Modellierung genregulatorischer Netzwerke mit stückweise linearen Differenzialgleichungen

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Jutta Gebert
aus Nürnberg

Hundt Druck GmbH, Köln
2007

Berichterstatter: Prof. Dr. U. Faigle
Prof. Dr. A. Burkovski

Tag der mündlichen Prüfung: 29.06.2007

Zusammenfassung

Das Verständnis der Prozesse innerhalb einer Zelle auf molekularbiologischer Ebene kann helfen, Krankheiten zu verstehen und Gene beziehungsweise Proteine zu bestimmen, die eine wichtige Rolle bei diesen Krankheiten spielen. Ebenso kann die Wirkungsweise von Medikamenten bei detaillierter Kenntnis der molekularen Prozesse vorausgesagt werden. Die vorliegende Arbeit behandelt die Modellierung biochemischer Netzwerke und stellt Differenzialgleichungen zur Beschreibung der dynamischen Prozesse auf, denen chemische Reaktionskinetiken zugrunde liegen. Eine Vereinfachung der Funktionen führt zu dem Vorteil, dass das System von Differenzialgleichungen analytisch lösbar wird.

Um bei genregulatorischen Netzwerken kleinere Subnetzwerke betrachten zu können, wird anhand eines graphentheoretischen sowie eines begriffsanalytischen Ansatzes, dem jeweils statistische Analysen folgen, eine Auswahl der Variablen des Differenzialgleichungssystems getroffen. Anschließend wird die Parameterschätzung für die Gleichungen mithilfe der Lösung eines Optimierungsproblems, das die Einbindung biologischen Wissens in lineare Nebenbedingungen erlaubt, durchgeführt. Ein weiterer Vorteil des Modells liegt darin, dass die aus Zeitreihen geschätzten Parameter direkte Rückschlüsse auf Abbauraten, Syntheseraten oder Regulierungsstärken zulassen.

Durch entsprechende Erweiterungen ist das Modell auch auf allgemeinere biochemische Netzwerke anwendbar. Zunächst wird die Stickstoffaufnahme im *Corynebacterium glutamicum* unter sich ändernden Stickstoffkonzentrationen modelliert. Dieses Modell konnte zum einen zufriedenstellende Ergebnisse bezüglich existierender Experimente liefern, zum anderen erlaubte es Schlussfolgerungen über Abbaumechanismen eines der beteiligten Proteine. Eine weitere Anwendung auf reale Daten beschäftigt sich mit dem DNA Reparatursystem im *Mycobacterium tuberculosis*. Hier konnte aufgrund der Komponentenauswahl ein Gen identifiziert werden, das möglicherweise eine wichtige Rolle in den Regulierungen übernimmt. Den Abschluss der Arbeit bilden Überlegungen zur Visualisierung, die teilweise wieder auf begriffsanalytische Methoden zurückgreifen.

Abstract

Understanding the processes in a cell on a molecular level can help to understand diseases and to determine genes and proteins that play an important role in these diseases. Moreover, the effects of drugs can be predicted given that the molecular processes are known in detail. This thesis focuses on model building of biochemical networks. Systems of differential equations, which are based on chemical reaction kinetics, are used to describe the dynamical processes. A simplification of the functions results in the advantage that the system becomes analytically solvable.

In order to select the variables for the system of differential equations two approaches for finding subnetworks of gene regulatory networks are proposed. These are a graph theoretical approach and an approach based on formal concept analysis. Both approaches are combined with a subsequent statistical analysis. Afterwards, the parameter estimation for the equations is carried out. An optimization problem with linear constraints that include mathematical formulations of biological knowledge is used. A further advantage of the model is the possibility to directly draw conclusions about regulation strengths, degradation and synthesis rates from the parameters of the equations.

Extending the model enables us to carry out applications on more general biochemical networks. One of these is the nitrogen uptake in *Corynebacterium glutamicum* under varying nitrogen concentrations. The model provides good results concerning existing experiments and additionally enables us to draw conclusions about degradation mechanisms of a protein involved in the system. A second application focuses on the DNA repair system in *Mycobacterium tuberculosis*. Using the proposed methods of finding variables which determine the sytem's behavior we are able to identify a gene that might play an important role in the repair mechanism as well. The thesis concludes with a presentation of visualization techniques, which include also formal concept analysis again.

Vorwort

In silico-Modelle einer Zelle sind eines der Hauptziele der Systembiologie, in der, wie auch in der klassischen Bioinformatik, mathematische Methoden von großer Relevanz sind. Im Bereich der klassischen Bioinformatik versucht man, die durch neue Technologien entstehende Informationsflut in der Biologie zu verarbeiten und Aussagen über einzelne Gene und Proteine zu erhalten. Mathematische Methoden, die in diesem Gebiet zum Einsatz kommen, sind zum Beispiel Hidden Markov Modelle zum multiplen Sequenz-Alignment, zum Clustern von Genexpressionsdaten oder zum Finden von CpG-Inseln, graphentheoretische Methoden zur Bestimmung phylogenetischer Bäume, geometrische Methoden zur Vorhersage von Bindestellen zwischen Proteinen und statistische Tests bei der Mikroarrayanalyse. Die Systembiologie dagegen verfolgt den Ansatz, ein gesamtes System wie eine Zelle oder ein Organ zu verstehen. Die Datenmengen, die über biologische Systeme gesammelt werden, sollen mithilfe mathematischer Modellierungen und Simulationen ein besseres Verständnis des Systems ermöglichen. Hauptgrund für die Versuche, eine Zelle zu simulieren, ist das Spektrum von Möglichkeiten, das sich für die Medizin eröffnet. Durch das Verständnis der Zusammenhänge auf molekularer Ebene können die Reaktionen auf äußere Einflüsse (wie Medikamente) vorhergesagt werden. Doch auch in Bereichen außerhalb der medizinischen Forschung ist solch ein tieferes Verständnis der molekularen Prozesse hilfreich, zum Beispiel können Aminosäuren und andere Produkte in Fermentationsprozessen effizienter hergestellt werden oder die biologische Wasserstofferzeugung durch Cyanobakterien könnte in Zukunft eine wichtige Rolle bei der Energieerzeugung spielen. Integriert werden sollen Daten aus den -omics-Bereichen wie Genomics, Transcriptomics, Proteomics und Metabolomics, also sowohl Informationen über genregulatorische Prozesse als auch über Signaltransduktionswege, metabolische Netzwerke etc.

Um dieses Ziel der Systembiologie zu erreichen, müssen zunächst die einzelnen Prozesse in der Zelle verstanden werden. Eine der Aufgaben besteht in der Rekonstruktion genregulatorischer Netzwerke aus Genexpressionsdaten. Diese Aufgabe wird oft auch als Reverse Engineering oder Inferenz von genregulatorischen Netzwerken bezeichnet. Jedoch ist nicht nur die Struktur, sondern insbesondere die Beschreibung der Dynamik des Netzes von Interesse. Oft werden die beiden Probleme separat betrachtet, so dass es zur Lösung des ersten Problems genügt herauszufinden, welche Komponenten sich gegenseitig beeinflussen. Bei der Modellierung der Dynamik dient die Struktur des Netzwerkes dagegen oft schon als Grundlage.

Es existieren jedoch einige Ansätze, die beide Probleme gleichzeitig lösen. Diese Ansätze unterscheiden sich meist im zugrunde liegenden Modell des dynamischen Verhaltens. In dieser Arbeit wird ein neues mathematisches Modell eingeführt, das einerseits komplex genug ist, um die genregulatorischen Prozesse widerspiegeln zu können, und das andererseits einfach genug strukturiert ist, um die Inferenzaufgabe lösen zu können.

Übersicht

Im ersten Kapitel wird eine Einführung in die Problemstellung gegeben. Die für das Problem wichtigen biologischen Prozesse in einer Zelle sowie Methoden zur Generierung von Genexpressions- und Proteindaten werden erläutert. Gängige Lösungsansätze werden diskutiert, wie beispielsweise die Modellierung genregulatorischer Netzwerke mit Booleschen oder Bayesschen Netzen. Ein Augenmerk liegt auf der Modellierung mit Differenzialgleichungen, da diese neben qualitativen auch quantitative Aussagen über das dynamische Verhalten eines Netzwerkes erlauben.
Im zweiten Kapitel wird der eigene Lösungsansatz erläutert. Dieser gründet auf der später von Yagil u. Yagil (1971) experimentell bestätigten Theorie von Jacob u. Monod (1961), dass Genregulationen durch sigmoidale Funktionen beschrieben werden können. Eine häufig verwendete Approximation solcher Funktionen ist durch Stufenfunktionen gegeben, da das hierdurch entstehende Modell einfache Analysemöglichkeiten aufweist, zum Beispiel ist es analytisch lösbar. Da gezeigt werden konnte, dass dieses Modell zumindest qualitativ die wesentlichen Eigenschaften des ursprünglichen Systems bewahren kann, werden diese sogenannten logischen Beschreibungen geschätzt. In dem hier vorgestellten neuen Ansatz wird die sigmoidale Funktion durch eine stückweise lineare Funktion approximiert, um eine genauere Beschreibung des Systems zu erhalten und dennoch auf Analysemethoden für lineare Differenzialgleichungen zurückgreifen zu können. Des Weiteren wird die Funktion so verschoben, dass sie konstant gleich Null ist, wenn die Regulierung nicht greift. Somit wird ein Term zu den Regulierungsfunktionen addiert, der als Grundexpression begriffen werden kann, also als diejenige Expressionsrate, die durch Fehlen jeglicher Regulierungen entsteht. Neben den Regulierungsfunktionen und der Grundexpression wird außerdem der Abbau der jeweiligen mRNA betrachtet, die als proportional zur Konzentration der mRNA angenommen wird. Dies führt zu folgendem Modell für die Beschreibung der mRNA-Konzentration x_i des Gens i, welches in der Menge $\Omega = \{1,...,n\}$ aller Gene liegt:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = c_i - \gamma_i \cdot x_i(t) + \sum_{j=1}^{n} rf_{i,j}(x_j(t)), \text{ mit } c_i, \gamma_i \in \mathbf{R}^+, rf_{i,j} : \mathbf{R} \to \mathbf{R}.$$

Hierbei beschreibt c_i die Grundexpression des Gens i, γ_i den Abbau der mRNA von Gen i und $rf_{i,j}(x_j(t))$ die Regulierungsfunktion, die den Einfluss von Gen j auf Gen i beschreibt und die sowohl positiv als auch negativ oder gleich Null sein kann. Ist die Konzentration x_j des regulierenden Gens j kleiner als ein erster Schwellwert, so findet keine Regulierung statt. Ist sie größer als ein zweiter Schwellwert, so hat die Einflussstärke ihren maximalen Wert erreicht, und zwischen den beiden Schwellwerten wird der entweder positive oder negative Einfluss jeweils linear abhängig von x_j beschrieben. Für ein System mit n Genen erhalten wir also ein System aus n gekoppelten Differenzialgleichungen erster Ordnung, die sowohl inhomogen als auch zeitinvariant sind. Graphentheoretisch lassen sich die Regulierungen darstellen durch den gerichteten Graphen (V,E) mit $V = \Omega$ und $E = \{(j,i) : \exists x_j \in \mathbf{R}^+, \text{ so dass } rf_{i,j}(x_j) \neq 0\}$. Jedem Knoten i werden die Parameter c_i und γ_i sowie jeder Kante (j,i) die Regulierungsfunktion $rf_{i,j}(x_j(t))$ zugeordnet. In ähnlicher Form werden diese Graphen oft zur Visualisierung der Abhängigkeiten genregulatorischer Netzwerke verwendet.
Die Annahme der Additivität der Regulierungen bringt den Vorteil, dass mit stückweise linearen Differenzialgleichungen gearbeitet werden kann. Für Systeme, in denen eine Komplexbildung

von verschiedenen Transkriptionsfaktoren erst zu einer Regulierung führt, muss eine Multiplikation von Regulierungsfunktionen ausgeführt werden. Der Grad der Regulierungsfunktion ist demnach gleich der Anzahl der verschiedenen Transkriptionsfaktoren, die am Komplex mitwirken. Der additive Modellansatz wird in diesem Kapitel hergeleitet und seine Reichweite sowie Vor- und Nachteile diskutiert. In Beispielen wird gezeigt, dass das Modell insbesondere häufig vorkommende biologische Dynamiken abbilden kann.
Zu Beginn der Modellierung stellt sich die Frage, welche Komponenten für ein betrachtetes System ausgewählt werden sollen bzw. ob die ausgewählten Komponenten ausreichen, um das Verhalten des Systems beschreiben zu können. Dieser bisher in der Literatur wenig beachteten Frage wird im dritten Kapitel der Arbeit nachgegangen. Hierbei wird einerseits ein graphentheoretischer Ansatz vorgestellt, dem statistische Analysen folgen, zum anderen die Möglichkeit, durch Anwendung der Begriffsanalyse zusammengehörende Komponenten zu finden. Beiden Ansätzen ist gemein, dass ein spezieller Korrelationskoeffizient verwendet wird, der unter anderem auch dem Rauschen auf den Daten Rechnung tragen kann, und dass eine Visualisierung der Ergebnisse möglich ist.
Bei der Modellierung stellt sich außerdem das Problem, dass viele Modellparameter nicht bekannt sind. Für die Parameterschätzung gibt es verschiedene Methoden, für die in vielen Anwendungsfällen die Datenlage jedoch nicht ausreicht. Oft sind die gemessenen Zeitreihen relativ kurz, so dass anstelle größerer Experimente andere biologische Informationen verwendet werden müssen, wie beispielsweise das Wissen über die Struktur des Netzes oder Abbauraten, um den Lösungsraum einzuschränken. Dieses Vorgehen wird im vierten Kapitel beschrieben.
Anschließend wird das ursprüngliche Modell auf allgemeinere biochemische Netzwerke erweitert, um posttranskriptionale Regulierungen oder spezielle Abbaumechanismen modellieren zu können. Es wird beschrieben, wie auch hier mit stückweise linearen Funktionen beziehungsweise Stufenfunktionen gearbeitet werden kann, so dass die Linearität in den einzelnen Bereichen des Zustandsraums nicht verloren geht. Im fünften und sechsten Kapitel wird das entwickelte Modell auf biologische Systeme und Daten angewendet. In Zusammenarbeit mit Nicole Radde und jeweils verschiedenen Kooperationspartnern - Prof. Andreas Burkovski vom Lehrstuhl für Mikrobiologie der Friedrich-Alexander-Universität Erlangen-Nürnberg, Dr. Christian Forst von der Bioscience Division der Los Alamos National Laboratories und Prof. Karin Schnetz vom Institut für Genetik der Universität Köln - sind bisher drei Modelle entstanden, von denen zwei in dieser Arbeit vorgestellt werden. In der ersten Anwendung wird die Stickstoffaufnahme des *Corynebacterium glutamicum* modelliert. Bei diesem Organismus gibt es verschiedene Transportwege von Ammonium in die Zelle, welches für den Aufbau von Proteinen benötigt wird. Einer dieser Wege existiert nur bei geringer Ammoniumkonzentration in der Umgebung des Bakteriums und führt über das Membranprotein AmtB, welches von AmtR inhibiert wird. Da AmtR wiederum von dem modifizierten GlnK-Protein inhibiert wird, bewirkt eine Modifikation von GlnK indirekt die Aktivierung von AmtB. Das nicht modifizierte oder wieder demodifizierte GlnK wird bei Vorhandensein von AmtB von Proteasen abgebaut. Hier konnte ein System von Differenzialgleichungen aufgestellt und dessen Parameter geschätzt werden, so dass das System nicht nur die bisherigen Experimente in der Simulation wiedergeben konnte, sondern auch mit veränderten Anfangszuständen oder Mutantensimulationen sinnvolle Ergebnisse sowie Hinweise für neue Experimente lieferte.
Im sechsten Kapitel ist die zweite Anwendung beschrieben. Hier wird ein Teil des DNA-Re-

paratursystems des *Mycobacterium tuberculosis* modelliert. Die Hauptkomponenten dieses Systems sind die beiden Gene *recA* und *lexA*, die ca. 35-40 weitere Gene regulieren. Das DNA-Reparatursystem wird angeschaltet, wenn die DNA so zerstört wird, dass einzelsträngige DNA vorliegt. An diese einzelnen Stränge bindet das Protein RecA, was zur Folge hat, dass das Protein LexA modifiziert wird, welches daraufhin nicht mehr an spezifische Bindestellen der DNA binden kann und dadurch die Fähigkeit verliert, die Reparaturgene zu inhibieren. Man sieht im Fall einer DNA-Zerstörung eine verstärkte Expression aller bisher inhibierten Reparaturgene, einschließlich der Gene *recA* und *lexA*. Zunächst werden die zugrunde liegenden Genexpressionsdaten beschrieben und mit den gängigen Methoden vorverarbeitet. Im Anschluss daran wird die Auswahl der Komponenten mithilfe der im dritten Kapitel vorgestellten Methoden beschrieben, um je nach Komponentenauswahl zwei verschiedene Modelle aufzustellen und deren Parameter zu schätzen. Der Vergleich der Simulationen dieser beiden Modelle legt nahe, dass das Gen *Rv2719c* möglicherweise eine große Rolle in dem Reparatursystem spielt. In weiteren Experimenten wird zur Zeit diese Behauptung überprüft.
Das weitere, in dieser Arbeit nicht beschriebene Modell behandelt die H-NS-Repression des *bgl*-Operons in *Escherichia coli*. Die spezielle Struktur des Modells mit zwei Feedbackschleifen erklärt die experimentelle Beobachtung einer hundertfach verstärkten Expression bei nur dreifach erhöhter Transkriptionsrate.
Im siebten Kapitel werden im Rahmen von Visualisierungsmöglichkeiten bei Differenzialgleichungen Alternativen zu Trajektorien, die üblicherweise die Bewegung im Zustandsraum zu gegebenen Anfangszuständen sichtbar machen, angegeben. Die chemische Organisationstheorie bildet Bewegungen im Zustandsraum auf Bewegungen in einem Verband ab. Diese alternative Möglichkeit der Visualisierung lässt es zu, dass mehrere Anfangszustände und deren Bewegung gleichzeitig innerhalb eines Verbandes veranschaulicht werden können. Eine Weiterentwicklung dieser Theorie auf Inhibierungen und Aktivierungen ist nötig und wird in diesem Kapitel erarbeitet. In der Begriffsanalyse gibt es einen ähnlichen Ansatz, der jedoch wie bei Trajektorien auch nur einen Anfangszustand betrachtet, für den der Begriffsverband gebildet wird. Beide Visualisierungsmethoden werden nach entsprechenden Veränderungen auf die Systeme des fünften und sechsten Kapitels angewendet.
Die Arbeit schließt im achten Kapitel mit einer Diskussion, die über die in dieser Arbeit erzielten Resultate geführt wird, und mit einem Ausblick über mögliche Erweiterungen. Es wurde ein neues Modell für genregulatorische Netzwerke eingeführt, für das Parameterschätzungen, Analyse- und Visualisierungsmöglichkeiten vorgestellt wurden, die dem Modell entsprechend modifiziert werden mussten. Des Weiteren wird die bisher weitestgehend unbeachtete Frage nach der Komponentenauswahl methodisch beantwortet. Ebenfalls diskutiert werden die Ergebnisse der beiden Anwendungen an *C. glutamicum* und *M. tuberculosis*.

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich während der Entstehung dieser Arbeit in irgendeiner Hinsicht unterstützt haben.
Zuvorderst danke ich für die fachliche Unterstützung meinem Betreuer Prof. Ulrich Faigle. Zusammen mit Prof. Rainer Schrader hat er es mir ermöglicht, in dem interdisziplinären und spannenden Gebiet der Bioinformatik zu forschen, wofür ich mich bei beiden bedanken möchte.

Für die intensiven Diskussionen während der Modellierung des *C. glutamicums* bedanke ich mich bei meinem Betreuer Prof. Andreas Burkovski. Für spannende Forschungsstunden und wertvolle Diskussionen danke ich Nicole Radde. Auch bedanke ich mich bei meinen weiteren Kooperationspartnern Susanne Motameny, Dr. Christian Forst, Dr. Röbbe Wünschiers, Prof. Karin Schnetz, Prof. Gerhard Wilhelm Weber und Prof. Stefan Pickl. Ebenfalls möchte ich mich bei Prof. Dietmar Schomburg bedanken sowie bei allen Korrekturlesern, die bisher noch nicht aufgeführt wurden, nämlich Volker Kudszus, Dr. Claus Gebert und Nicole Blumberg.
Insbesondere meinem Ehemann Volker Kudszus möchte ich besonderen Dank aussprechen, da er durch seine ständige seelische Unterstützung ebenfalls zum Gelingen dieser Arbeit beigetragen hat.
Zu guter Letzt danke ich meinen Eltern Christa und Alfred Gebert, die mich bereits mein Leben lang in jeder Hinsicht unterstützten.

Köln, Mai 2007 Jutta Gebert

Inhaltsverzeichnis

Kapitel 1

Einführung in die Problemstellung

1.1 Biologischer Hintergrund

Regulatorische Vorgänge in einer Zelle können dazu dienen, dass zum Beispiel der Organismus in vorteilhafter Weise auf eine veränderte Umgebung reagiert oder periodisch wiederkehrende Prozesse wie die Zellteilung gesteuert werden. Auf molekularbiologischer Ebene betrachtet man hierbei die Interaktionen von Makromolekülen in der Zelle, zu denen unter anderem DNA, RNA und Proteine gehören. Die verschiedenen Interaktionen vollziehen sich auf unterschiedlichen Zeitskalen. Eine Regulierung auf Genexpressionsebene findet beispielsweise im Minutenbereich statt, während sich Proteinkaskaden innerhalb von Sekunden abspielen. Spricht man abstrahiert von genregulatorischen Netzwerken, in denen ein Gen *a* ein Gen *b* aktiviert oder inhibiert, so hat man es mit einer Reihe von biologischen Prozessen zu tun, die in diesem ersten Teil des Kapitels beschrieben werden. Hier wird ebenfalls beschrieben, in welcher Weise im Folgenden abstrahiert wird und welche Probleme dies beinhalten kann. Im Anschluss folgt eine kurze Erklärung über die gängigsten Methoden zur Messung von mRNA und Proteinen sowie ein Überblick über allgemeine Analysemethoden von Mikroarray-Daten, mit denen die bisher größten Datenmengen produziert werden.

1.1.1 Molekularbiologische Prozesse

Die Produktion eines Proteins aus einem Gen, Genexpression genannt, vollzieht sich in zwei Schritten. Zunächst findet die Transkription, dann die Translation statt. Transkription heißt die Umschreibung der Nukleotidsequenz der DNA in die Nukleotidsequenz der Boten-RNA, welche wir im Folgenden mit mRNA[1] bezeichnen. Translation bezeichnet die Übersetzung der Nukleotidsequenz der mRNA in die Aminosäuresequenz des Proteins. Lange Zeit wurde behauptet, dass der Informationsfluss von der DNA über die RNA zum Protein ginge. Dies wurde lange Zeit als das zentrale Dogma der Molekularbiologie angesehen. Heutzutage wird dies nicht mehr so kategorisch formuliert, da nun bekannt ist, dass auch eine Umschreibung der Sequenzen von RNA zu DNA oder RNA zu RNA vollzogen werden kann (Janning u. Knust, 2004). Mit dem in der Genetik häufig verwendeten Begriff des Informationsflusses ist gemeint, dass die DNA-Sequenz bereits die RNA- und Proteinsequenz vorgibt, dass also die aufgrund der DNA produzierte RNA

[1] aus dem Englischen ‘messenger-RNA’

und die Proteine ‘keine Überraschung’ mehr enthalten[2].

DNA

Ein Gen ist definiert als ein proteinkodierender Bereich auf der DNA. Bereits 1953 wurde die Struktur der DNA von Watson und Crick beschrieben (Watson u. Crick, 1953). DNA ist die in dieser Arbeit bereits verwendete Abkürzung des englischen Wortes desoxyribonucleic acid. Eine weitere Abkürzung dieses Makromoleküls lautet DNS aufgrund des deutschen Wortes Desoxyribonukleinsäure. Die DNA hat vier verschiedene Nukleotide als Bausteine. Die Nukleotide bestehen jeweils aus einer Desoxyribose, einem Phosphat und einer organischen Base. Die vier Nukleotide unterscheiden sich nur in der Base, die durch A für Adenin, G für Guanin, C für Cytosin und T für Thymin abgekürzt werden.
Die Nukleotide können Nukleotidketten, sogenannte Polynukleotide, bilden, indem sie zwischen dem Phosphatrest der Desoxyribose des einen Nukleotids und der OH-Gruppe der Desoxyribose des anderen Nukleotids eine Bindung eingehen. Der Anfang der Kette besitzt also noch eine Phosphatgruppe, das Ende noch eine OH-Gruppe. Da das Phosphat am C5-Atom, die OH-Gruppe am C3-Atom der Desoxyribose sitzt, spricht man davon, dass die Nukleotidsequenz von 5’ nach 3’ gelesen wird. Die Desoxyribose und das Phosphat bilden das sogenannte Rückgrat der DNA, das in Form zweier Nukleotidketten vorliegt. Die Basen sind jeweils nach innen gerichtet und bilden Wasserstoffbrückenbindungen aus. Diese Wasserstoffbrückenbindungen kann es nur zwischen A und T beziehungsweise zwischen C und G geben. Die beiden Einzelstränge haben eine entgegengesetzte 5’ zu 3’-Orientierung (siehe auch Abbildung 1.1) und sind komplementär zueinander, so dass entsprechend der Anzahl der Basenpaare zwei Wasserstoffbrückenbindungen für jedes A-T-Basenpaar und drei für jedes C-G-Basenpaar ausgebildet werden, was die DNA zu einem sehr stabilen Molekül macht. Die Basenpaarung zwischen zwei einzelsträngigen DNA-Molekülen ist die Grundlage der im weiteren Verlauf des Kapitels beschriebenen Mikroarrays. Schematisch ist die Anordnung der Basen in der DNA in Abbildung 1.1 zu sehen. Aufgrund der spiralförmig gewundenen Struktur der DNA wird diese Struktur auch Doppelhelix genannt.

Transkription

Bei der Transkription wird ein Enzym benötigt, die sogenannte RNA-Polymerase, die am Promotor bindet, welcher aus einer Erkennungssequenz für die RNA-Polymerase vor der zu transkribierenden Region besteht. Im Folgenden werden wir uns auf die Beschreibung der prokaryontischen Prozesse beschränken, da die in den weiteren Kapiteln folgenden Modelle für Prokaryonten entwickelt wurden. Prokaryonten haben keinen Zellkern und es gehören hauptsächlich Bakterien zu dieser Klasse. Im Gegensatz zu den Prokaryonten haben Eukaryonten, also Organismen mit Zellkern, verschiedene RNA-Polymerasen und weniger stark konservierte Promotorregionen. Der wichtigste Grund, warum in dieser Arbeit Prokaryonten betrachtet werden, liegt jedoch darin, dass in Eukaryonten weitaus mehr Regulierungsmechanismen wirksam sind

[2]Der Begriff der Information in statistischer Sichtweise geht von einer Quelle aus, die Symbole generiert, welche statistisch unabhängig sind. Dies ist bei der DNA jedoch nicht gegeben, so dass von der Information nach Shannon unterschieden wird.

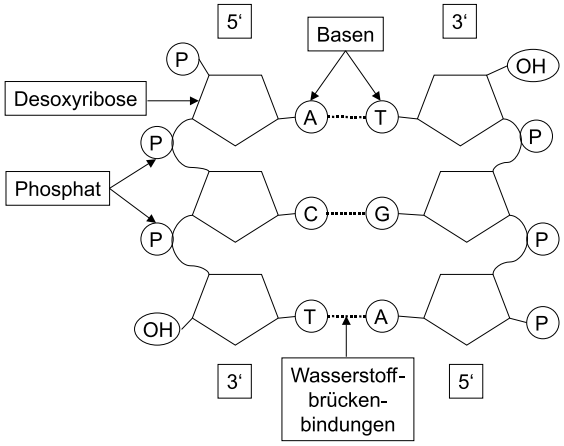


Abbildung 1.1: Die Abbildung zeigt drei Basenpaare der DNA und ihre Wasserstoffbrückenbindungen, welche durch gepunktete Linien dargestellt sind.

als in Prokaryonten, so dass das im zweiten Kapitel vorgestellte Grundmodell erweitert werden muss. Derartige Erweiterungen werden im fünften Kapitel diskutiert, sollen jedoch für das Grundmodell vernachlässigt werden. In Prokaryonten ist die DNA-Sequenz zu Beginn und zum Ende des Promotors sehr stark konserviert, das heißt, man findet in fast allen Prokaryonten diese oder eine sehr ähnliche DNA-Sequenz. Zehn Basenpaare vor dem Transkriptionsstart findet man eine spezielle Sequenz, TATA-Box genannt, 35 Basenpaare vor dem Transkriptionsstart die sogenannte -35-Region. Findet die RNA-Polymerase diese Stellen der DNA, kann sie binden und die DNA ungefähr 20 Basenpaare lang entwinden, so dass kurze Einzelstränge entstehen, die abgelesen werden können für die Bildung der mRNA. In dieser Arbeit wird wiederum die häufigere Bezeichnung RNA (Abkürzung für ribonucleic acid) statt RNS (Abkürzung für Ribonukleinsäure) verwendet. Die RNA liegt im Gegensatz zur DNA als Einzelstrang vor und ist im Vergleich zur DNA recht instabil. Die Lebensdauer einer mRNA beträgt meist nur wenige Dekaminuten (Wünschiers u. a., 2001). Sie unterscheidet sich von der DNA in zwei Aspekten: Die Nukleotide enthalten Ribose anstelle von Desoxyribose und Uracil anstelle von Thymin. Uracil wird wie üblich mit U abgekürzt. Wieder können Nukleotidketten entstehen, indem die 3’-OH-Gruppe des einen Ribonukleotids mit dem 5’-Phosphat eines weiteren Ribonukleotids eine Bindung ausbildet. Bei der Transkription findet die Synthese immer von 5’ nach 3’ statt. Zu einem Abbruch der Transkription kommt es, wenn entweder eine spezielle Struktur der RNA - die sogenannte Haarnadelschleife - erreicht ist oder ein Protein an die DNA gebunden hat, so dass die RNA-Polymerase sich an dieser Stelle von der DNA lösen muss. In Prokaryonten kommt es häufig vor, dass viele Gene direkt hintereinander auf der DNA liegen, dadurch denselben Promotor haben und häufig in eine einzige mRNA abgelesen werden, so dass diese mRNA für mehrere Proteine gleichzeitig kodiert. Der entsprechende DNA-Bereich aus Promotor, regulierenden Bindungsstellen und Genen wird Operon genannt. In dem Fall, dass ein Repressor an die DNA bindet, wird die regulierende Bindungsstelle Operator genannt. Im Fall eines Aktivators nennen wir sie Aktivatorbindungsstelle.

Translation

Bei der Translation findet nun die Übersetzung der Nukleotidsequenz der mRNA in die Aminosäuresequenz des Proteins statt. Hierfür sind neben der mRNA Ribosomen, tRNA, Initiations-, Elongations- und Terminationsfaktoren notwendig. Ribosomen sind Komplexe, die aus RNA und Proteinen bestehen und an denen die Translation ausgeführt wird, tRNA (transfer RNA) ist eine Klasse von RNA, die die Aminosäuren zum Ribosom transportiert und Initiationsfaktoren sind Proteine, die das Ribosom an die mRNA binden. Elongationsfaktoren sind dementsprechend Proteine, die zur Elongation der Kette aus Aminosäuren benötigt werden, und Terminationsfaktoren sind Proteine, die das Ribosom an den Nukleotidsequenzen UAA, UAG oder UGA, den sogenannten Stopcodons, zerfallen lassen und die Aminosäuresequenz loslösen. Die Bindungsstelle der Ribosomen ist bei Prokaryonten durch eine Nukleotidsequenz der Länge sechs festgelegt, die sich wiederum zwischen vier bis sieben Basen vor dem Tripel AUG befindet. Dieses Tripel ist bei Prokaryonten das Startcodon, es gibt nur wenige Ausnahmen. Die Ribosomenbindungsstellen können auch mehrfach innerhalb der mRNA-Sequenz vorkommen, so dass wie vorher beschrieben mehrere verschiedene Proteine aus derselben mRNA abgelesen werden können. Das Startcodon AUG kodiert die Aminosäure Methionin, abgekürzt mit MET oder M. Die meisten weiteren 3-Permutationen mit Wiederholung der vier Buchstaben A,C,G,U kodieren ebenfalls eine der zwanzig üblicherweise in Organismen zu findenden Aminosäuren. Der Code ist nicht injektiv, wie in Tabelle 1.1 zu erkennen ist.

Tabelle 1.1: Jede 3-Permutation aus den vier Buchstaben A, C, G und U bestimmt eine Aminosäure oder ein Start-/Stop-Codon.

Erste	Zweite Base				Dritte
Base	U	C	A	G	Base
U	Phenylalanin	Serin	Tyrosin	Cystein	U
U	Phenylalanin	Serin	Tyrosin	Cystein	C
U	Leucin	Serin	Stop-Codon	Stop-Codon	A
U	Leucin	Serin	Stop-Codon	Tryptophan	G
C	Leucin	Prolin	Histidin	Arginin	U
C	Leucin	Prolin	Histidin	Arginin	C
C	Leucin	Prolin	Glutamin	Arginin	A
C	Leucin	Prolin	Glutamin	Arginin	G
A	Isoleucin	Threonin	Asparagin	Serin	U
A	Isoleucin	Threonin	Asparagin	Serin	C
A	Isoleucin	Threonin	Lysin	Arginin	A
A	Methionin/Start-Codon	Threonin	Lysin	Arginin	G
G	Valin	Alanin	Asparaginsäure	Glycin	U
G	Valin	Alanin	Asparaginsäure	Glycin	C
G	Valin	Alanin	Glutaminsäure	Glycin	A
G	Valin	Alanin	Glutaminsäure	Glycin	G

Die Aminosäuresequenz beginnt sich noch während der Translation zu falten und bildet dann ein

Protein, eines der wichtigsten Bausteine der Zelle, da Proteine die verschiedensten Aufgaben übernehmen, wie beispielsweise Bildung der Struktur der Zelle, Katalyse von biochemischen Reaktionen oder Kontrolle der Genexpression. Weitere Studien zur Proteinfaltung finden sich beispielsweise in Wiebringhaus u. a. (2003, 2004).

Regulierung der Genexpression

Woher 'weiß' nun jedoch eine Zelle, welche Gene exprimiert werden sollen und welche nicht? Neben Ptashnes Arbeiten über Genregulierungen (Ptashne, 1986) war die Arbeit von Jacob u. Monod (1961) hier wegweisend. Jacob und Monod stellten sich genau diese Frage in Bezug auf das Bakterium *E. coli* und seine Reaktion auf verschiedene Zuckerarten in seiner Umgebung. Die Antwort auf die Frage, warum *E. coli* weiß, welche Proteine in den verschiedenen Zuckerumgebungen gerade exprimiert werden müssen, ist einfach: Das 'Wissen' ist in den Sequenzen der Operatoren bzw. Aktivatorbindungsstellen enthalten und wird über sogenannte Regulatorproteine genutzt. Die Regulatorproteine, die ausschließlich die Transkription regulieren, werden im Folgenden auch Transkriptionsfaktoren genannt. Ist die präferierte Umgebung vorhanden, bindet ein Repressor (auch Inhibitor genannt) an den Operator. Hier entsteht eine Überlappung mit dem Promotor, so dass die RNA-Polymerase schlechter an den Promotor binden kann als in der Situation ohne den Repressor und dass dadurch weniger mRNA und weniger Protein produziert wird. Der Repressor kann in einer anderen Situation als der präferierten jedoch auch an einen Induktor, in diesem speziellen Fall Laktose, binden, und somit den Promotor für eine verstärkte Expression freigeben. In diesem speziellen Fall wird Laktose in die Zelle hineintransportiert und geht eine Verbindung mit dem Repressor ein. Die Signalübertragung zum Repressor kann jedoch auch über Proteinkaskaden geschehen, so dass ein Signal übertragen wird, indem beispielsweise eine ganze Reihe von Strukturveränderungen von Proteinen hervorgerufen werden. In der Studie von Jacob und Monod existiert neben dem Repressor auch noch ein Aktivator, der in der präferierten Situation jedoch nicht an die DNA binden kann und erst durch Modifikation - im speziellen Fall durch Binden an den Induktor - zur Bindung an die DNA fähig wird (Knippers, 2001). Die entsprechenden Prozesse hierzu sowie weitere Möglichkeiten der Regulierung sind in Abbildung 1.2 verdeutlicht. Die Rolle des Induktors kann jedoch auch ein weiteres Regulatorprotein übernehmen. Beispielsweise sei ein Inhibitor durchgängig und unabhängig von der Umgebung des Bakteriums, also konstitutiv exprimiert. Dann kann das Vorhandensein eines Regulatorproteins zu einer Aktivierung des inhibierten Gens führen, indem der Regulator die Bindung des Inhibitors an die DNA verhindert. Ebenso kann im Fall eines konstitutiv exprimierten Aktivators eine Inhibierung durch ein Regulatorprotein erfolgen, indem es die zunächst vorhandene Aktivierung des Aktivators aufhebt. Wiederum muss jedoch eine Induzierung vorangegangen sein, das heißt, dass eine Signalübertragung von außerhalb der Zelle in die Zelle stattgefunden hat. Die im vorherigen Fall beschriebenen Regulierungen sind indirekt, wohingegen die zuvor als konstitutiv exprimiert angenommenen Regulatorproteine direkt wirken. Im Folgenden werden verschiedene Möglichkeiten der Regulierung auf der Ebene der Transkription definiert.

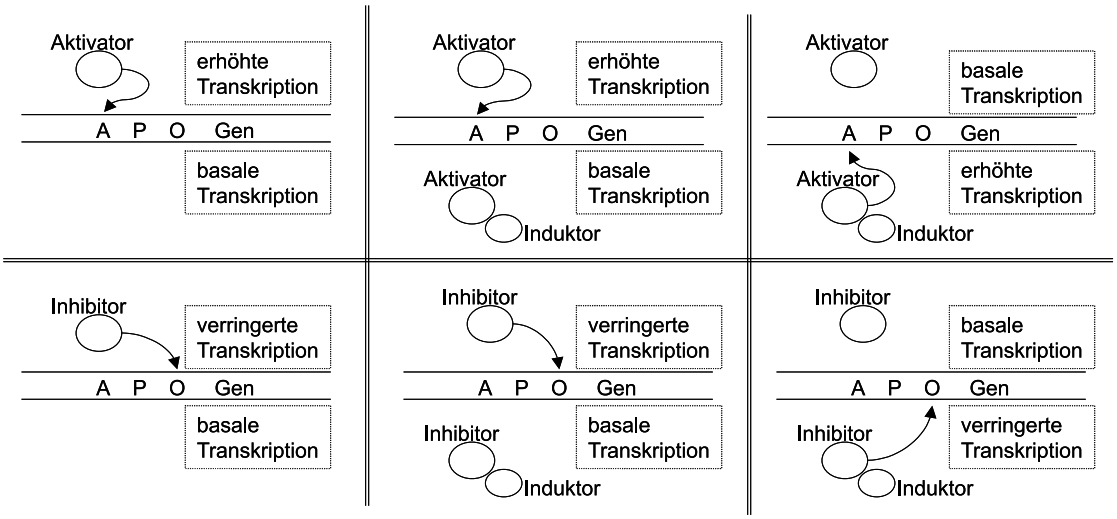


Abbildung 1.2: In den Skizzen sieht man eine Aktivatorbindungsstelle (A), einen Promotor (P), einen Operator (O) sowie ein Gen, die sich alle auf der DNA befinden, welche hier als länglicher Schlauch dargestellt ist. Aktivatoren binden an eine Aktivatorbindungsstelle und verstärken dadurch die Transkriptionsrate des Gens, Inhibitoren binden am Operator und wirken gegensätzlich, verringern also die Transkriptionsrate. Die Grundtranskription oder basale Transkription ist jene Transkription, die ohne jegliche Regulatorproteine stattfindet. Auf der linken Seite werden direkte Regulierungen dargestellt, in der Mitte indirekte Regulierungen. In den mittleren Skizzen wird gezeigt, wie ein Induktor das Regulatorprotein davon abhalten kann, an die DNA zu binden. Solche Induktoren können Moleküle aus der Umgebung des Bakteriums sein, die durch die Zellmembran ins Innere der Zelle transportiert werden oder wiederum Regulatorproteine, die unterschiedlicher Expression unterworfen sind. Alternativ kann eine indirekte Regulierung auch dadurch stattfinden, dass erst die Bindung an einen Induktor das Regulatorprotein zur Bindung an die DNA ermöglicht, wie auf der rechten Seite gezeigt.

Definition 1.1. *(Regulierungen auf Transkriptionsebene)*

- *Eine direkte Aktivierung (Inhibierung) eines Gens liegt vor, wenn durch Vorhandensein eines Aktivators (Inhibitors) in der Zelle die Expression des Gens erhöht (verringert) wird.*
- *Eine indirekte Aktivierung (Inhibierung) eines Gens liegt vor, wenn in der Zelle bereits ein Inhibitor (Aktivator) dieses Gens vorliegt und diese Inhibierung (Aktivierung) durch Erscheinen eines Aktivators (Inhibitors) aufgehoben wird.*

Die Ermöglichung der Bindung des Regulatorproteins durch ein weiteres Regulatorprotein bezeichnen wir ebenfalls als direkte Regulierung. Die bisher beschriebenen Regulierungen betreffen die Anziehung oder Blockierung der RNA-Polymerase an den Promotor. Regulierungen können jedoch auch stattfinden, wenn die RNA-Polymerase schon gebunden hat, jedoch das Ablesen der Sequenz nicht starten kann, weil die Doppelhelix noch nicht entwunden wurde. Auch hier können Aktivatoren oder Inhibitoren wirken, indem sie beispielsweise eine Änderung der Struktur der RNA-Polymerase oder der DNA herbeiführen. Ähnlich wirken Regulatorproteine, die die DNA beispielsweise durch Schleifenbildung derart verformen, dass ein weiter entfernt wirkendes Regulatorprotein eine Inhibierung oder Aktivierung hervorrufen kann. Dies ist ein Fall, wo erst kooperativ wirkende Proteine die Expression beeinflussen. Regulierungen während der Elongation der mRNA und bei Termination sind ebenso möglich. Ein Beispiel liefert das *trp*-Operon von *E. coli*, das neben einem Repressor auch einen vorzeitigen Abbruch der Transkription verwendet, wenn die Aminosäure Tryptophan vorhanden ist (Janning u. Knust, 2004). Eine Vielzahl von verschiedenen Wirkungsmechanismen ist demnach möglich, der Schwerpunkt bei Prokaryonten liegt jedoch bei der Regulierung der Transkriptionsinitiation.

1.1.2 DNA-Mikroarrays, Proteinarrays und weitere Messmethoden

Um Genexpressionen zu untersuchen, stehen zwei Möglichkeiten zur Verfügung. Entweder wird das Proteom untersucht, also alle vorliegenden Proteine, oder das Transkriptom, also alle vorliegenden mRNAs (Wünschiers, 2004). Da die Transkriptomanalyse sehr viel einfacher ist als die Proteomanalyse, liegt bisher der Schwerpunkt auf Messungen der mRNAs, die für die Messmethoden gut geeignete chemische Eigenschaften aufweisen. Hier ist insbesondere die Verwendung von Mikroarrays üblich, da mehrere Tausend mRNAs gleichzeitig untersucht werden können. Die Ergebnisse der Mikroarray-Experimente werden in Datenbanken gesammelt und sind teilweise frei zugänglich, zum Beispiel im Gene Expression Omnibus[3]. Im Folgenden werden DNA-Mikroarrays im Detail beschrieben, während einige weitere Messmethoden nur kurz skizziert werden.

DNA-Mikroarrays

DNA-Mikroarrays sind auf einem Träger angebrachte Oligonukleotide, ca. 20-100 Basen lange Nukleotidketten (Schena, 2003; Wünschiers u. a., 2001), die mit DNA-Molekülen hybridisieren,

[3] Gene Expression Omnibus at NCBI, http://www.ncbi.nlm.nih.gov/geo

also Bindungen aufgrund der in Kapitel 1.1.1 beschriebenen Basenpaarungen eingehen. Das Trägermaterial für die Oligonukleotide, die auch Sonden genannt werden, besteht meist aus Glas, Kunststoff oder Membranen. Mehrere derselben Sonden bilden einen 'Spot', so dass quantitative Aussagen gemacht werden können, da nicht nur eine, sondern mehrere Bindungen möglich sind. Diese Spots sind beispielsweise bei dem *geniom one*-Gerät quadratisch und $33 \times 33 \mu m$ groß (Wünschiers, 2004). Die Mikroarrays dieses Geräts bieten Platz für 48.000 Spots (Wünschiers, 2004), im Allgemeinen haben DNA-Mikroarrays 4.000-50.000 Spots (Butte, 2002). Die Sonden werden derart ausgewählt, dass die für die Studie interessierenden Gene daran binden. Da eine mRNA jedoch nicht so stabil ist wie DNA, wird aus dem mRNA-Ausgangsmaterial durch reverse Transkription cDNA (complementary DNA, also komplementäre DNS) hergestellt. Während dieses Prozesses werden die cDNA-Moleküle mit einem fluoreszierenden Farbstoff versehen. Werden Vergleichsexperimente durchgeführt, so werden die cDNA-Mengen aus zwei verschiedenen Ausgangsproben mit jeweils unterschiedlichem Farbstoff, meist Cyanin-3 und Cyanin-5 (Wünschiers u. a., 2001), gefärbt. Nach der Hybridisierung der cDNA-Mengen mit den Sonden werden die Mikroarrays mit einem Laserscanner abgetastet, um die Fluoreszenzintensität zu ermitteln. Insbesondere die Extraktion der mRNA aus der biologischen Probe, die Synthese der Sonden, das Aufbringen der Sonden auf das Trägermaterial, der Hybridisierungsprozess und das Ablesen und Zuordnen der Lichtsignale sind mit Fehlern behaftet. Die Synthese der Sonden kann beispielsweise durch ein photolithographisches Verfahren erfolgen. Hierbei werden mit einer Lochmaske jene Oligonukleotide ausgewählt und durch Belichtung aktiviert, die verlängert werden sollen. Die Übertragung der Sonden auf das Trägermaterial, auch Spotting genannt, kann durch verschiedenste Verfahren erfolgen. Im einfachsten Fall wird die Sonde durch eine Nadel aufgenommen und an die vorher bestimmte Position auf dem Trägermaterial abgegeben. Verbesserte Methoden sind beispielsweise die Split-Needle-Methode oder die Ring-and-Pin-Methode (Wünschiers u. a., 2001), die kleinere Distanzen zwischen den Spots ermöglichen als die einfache Methode. Die Probleme bleiben jedoch bestehen: Zum einen wird die vorherbestimmte Position eventuell nicht genau getroffen, zum anderen können Einflüsse wie zum Beispiel Temperaturschwankungen in der Umgebung zu veränderten Ergebnissen führen. Auch die Übertragung der Lichtsignale in eine Datenmatrix nach der Hybridisierung birgt Fehlerquellen. Wie eben erwähnt, liegen die Spots eventuell nicht auf einer Linie, so dass die Zuordnung der Lichtsignale zu der Sonde Schwierigkeiten bereiten kann. Auch ungebundene Sonden können ein Fluoreszenzsignal senden, so dass die gewonnenen Daten von dem ebenfalls gemessenen Hintergrundrauschen befreit werden müssen. Üblicherweise werden außerdem verschiedene Normalisierungen angewendet. Hat man mehrere Mikroarray-Experimente durchgeführt, kann beispielsweise die Annahme getroffen werden, dass bei verschiedenen Mikroarrays die cDNA-Mengen gleich sind und somit die Gesamtfluoreszenz jedes einzelnen Mikroarrays gleich sein muss. Eine andere Annahme ist, dass die meisten Gene sich in verschiedenen Experimenten nicht ändern oder dass es sogenannte Haushaltsgene gibt, deren Expressionsniveau unverändert bleibt (Butte, 2002). Einen nicht zu vernachlässigenden Einfluss hat auch das 'biologische Rauschen'. Nähme man beispielsweise aus demselben Organ morgens und abends eine Probe, so wären die Messergebnisse selbst dann unterschiedlich, wenn alle anderen Fehlerquellen beseitigt wären, da die Expression einiger Gene durch Hormone beeinflusst wird und somit den Tagesrhythmus widerspiegelt (Butte, 2002). Werden also Krebszellen mit gesunden Zellen verglichen, so ist ein in Krebszellen besonders hoch exprimiertes Gen eventuell nur Ergebnis der

biologischen Variabilität. An all den aufgezählten Fehlerquellen wird jedoch ständig weitergearbeitet, da trotz der vorhandenen Schwierigkeiten Mikroarrays eine Vielfalt an Möglichkeiten bieten. Der für diese Arbeit interessante Aspekt ist die Möglichkeit, Gen-Gen-Interaktionen aus den Mikroarrays herauszulesen und aufgrund der quantitativ verwertbaren Messungen Funktionen für diese Gen-Gen-Interaktionen zu bestimmen, die in Abhängigkeit der Lichtintensitäten aller mRNAs eine Aussage für die Lichtintensität jeder einzelnen mRNA zuläßt. Allgemeiner gefasst möchte man die Konzentration jeder einzelnen mRNA durch eine Funktion beschreiben, die abhängig ist von der Konzentration derjenigen mRNAs, die an einer Regulierung des entsprechenden Gens beteiligt sind. Ein weiterer möglicher Anwendungsbereich ist die Lokalisation von SNPs, einzelnen Nukleotidpolymorphismen[4] , also Positionen des Genoms, an denen in verschiedenen Organismen ein und derselben Spezies verschiedene Basen zu finden sind. Im Menschen sind diese Unterschiede im Genom verantwortlich für die unterschiedlichen physischen Eigenschaften wie die Augenfarbe, jedoch auch für die individuell verschiedenen Verträglichkeiten von Medikamenten und vermutlich auch für das Risiko des Ausbruchs verschiedener Krankheiten (Wünschiers u. a., 2001). Des Weiteren können Krankheiten klassifiziert werden, um unterschiedliche Behandlungsweisen zu ermöglichen. Beispielsweise kann für verschiedene Unterklassen von Tumoren die Überlebenszeit vorhergesagt werden (Kaderali u. a., 2006). In der Pharmakologie kommen Mikroarrays zum Einsatz, um die differenzielle Expression in Gewebe zu untersuchen, das unterschiedlichen Medikamenten ausgesetzt ist. Auch die Wirkungsweise eines Medikaments auf molekularer Ebene kann untersucht werden.

Proteinarrays und weitere Messmethoden

DNA-Mikroarrays messen nur den ersten Schritt der Genexpression, die Transkription. Proteinarrays sind eine Möglichkeit, um das Proteom, also die Gesamtheit der in der Zelle vorliegenden Proteine, zu messen. Der Grund, wieso die Transkriptionsanalyse verbreiteter ist als die Proteomanalyse, liegt darin, dass Letztere komplizierter durchzuführen ist. Statt Oligonukleotiden, die komplementär zur entsprechenden cDNA sind, werden bei Proteinarrays Antikörper verwendet, die das zu untersuchende Protein binden, welches ebenfalls vorher mit Fluoreszenzfarbstoffen oder radioaktivem Material markiert wird. Einige der Probleme bei Proteinarrays und ihre Nachteile im Vergleich zu DNA-Mikroarrays sind in (Wünschiers u. a., 2001) aufgelistet: Proteine sind instabiler als DNA-Moleküle, sie sind größer als DNA und ermöglichen daher nur schwächere Signalstärken und die Bindung der Antikörper an das Trägermaterial ist ebenfalls schwieriger als bei Oligonukleotiden. Des Weiteren müssen Antikörper für die zu untersuchenden Proteine gefunden werden, das heißt, es müssen Moleküle bzw. andere Proteine gefunden werden, die die zu untersuchenden Proteine binden.
Eine häufiger angewandte Methode als Proteinarrays ist zur Untersuchung weniger Proteine durch Westernblots gegeben. Im Jahr 1975 wurde von Edwin Southern ein Verfahren zur DNA-Fragmentanalyse entwickelt, das nach seinem Erfinder und dem englischen Verb ‘to blot’[5] den Namen ‘Southern Blotting’ erhielt. Die Weiterentwicklung für RNA wurde dementsprechend ‘Northern Blotting’ genannt, das Verfahren für Proteine ‘Western Blotting’. Das Grundprinzip des Western Blottings ist die Elektrophorese. Man verwendet ein elektrisches Feld, so dass

[4] SNP ist die Abkürzung von ‘Single Nucleotide Polymorphism’
[5] Übersetzung: mit Löschpapier aufsaugen

die geladenen Moleküle aufgrund ihrer unterschiedlichen Größe und ihrer Ladung verschieden schnell durch das Trägermaterial wandern und sich dadurch voneinander trennen. Anschließend werden die Proteine auf eine Membran übertragen, auf die eine Menge von ersten Antikörpern gebracht wird, welche eine Bindung zu den Proteinen eingehen können. Im dritten Schritt werden zweite Antikörper benötigt, die an den ersten Antikörper binden und sich zum Beispiel durch Fluoreszenz sichtbar machen lassen (Lodish u. a., 2001; Janning u. Knust, 2004).

1.1.3 Überblick über mathematische Methoden zur Analyse von Genexpressionsdaten

Der Fokus dieser Arbeit liegt auf der Konstruktion biochemischer Netzwerke, jedoch finden je nach Fragestellung weitere mathematische Methoden zur Analyse von Mikroarrays Verwendung (Slonim, 2002; Butte, 2002; Shamir u. Sharan, 2002; Kaminski u. Friedman, 2002). Zunächst wird eine Normalisierung der Daten durchgeführt mit dem Ziel, die Expression eines Gens nicht nur mit anderen Genen des Mikroarrays, sondern auch mit weiteren Mikroarrays vergleichen zu können. Unterschiedliche Expressionswerte ein und desselben Gens lassen sich auf entweder technische oder biologische Gründe zurückführen. Insbesondere technische Fehlerquellen wie das Hintergrundrauschen werden beseitigt. Die verschiedenen Fehlerquellen sowie deren Beseitigung sind in Sartor u. a. (2003) aufgeführt.
Nimmt man bereits normalisierte Daten an, liegt ein Schwerpunkt häufig auf dem Erkennen von differenziell exprimierten Genen, wofür üblicherweise verschiedene statistische Tests verwendet werden. Eine der einfachsten statistischen Methoden, um differenziell exprimierte Gene bei zwei unterschiedlichen Bedingungen zu finden, ist der Studentsche t-Test mit verschiedenen Modifikationen (Cui u. Churchill, 2003).
Ein weiterer Schwerpunkt ist die Gruppierung von Genen mit ähnlichen Transkriptionsmustern. Hier kommen verschiedene Clustermethoden zum Einsatz, die alle das Ziel haben, die Daten derart in Klassen, den sogenannten Clustern, zu gruppieren, dass die Objekte innerhalb eines Clusters große Ähnlichkeit zueinander haben, aber sehr große Unähnlichkeit zu Objekten aus anderen Clustern besteht. Hierbei sind die Objekte meistens die Gene, da Gene mit ähnlichen Genexpressionen oft ähnliche Funktionen in zellulären Prozessen haben. In diesem Fall können Clustermethoden dazu verwendet werden, Funktionen von neuen, bisher unbekannten Genen vorherzusagen. Ein Objekt entspricht dann einem Gen und besteht aus dem Vektor $(x_1, ..., x_n) \in \mathbf{R}^n$, wobei x_i die Expression dieses Gens im Experiment i ist. Ebenfalls möglich ist das Clustern von m Experimenten, so dass in diesem Fall ein Objekt $x \in \mathbf{R}^m$ ein Experiment ist und x_i die Expression des i-ten Gens in dem Experiment bezeichnet. In diesem Fall ist das Ziel beispielsweise, gesundes Gewebe von Tumorgewebe zu unterscheiden, wobei sich für solche Fragestellungen auch Klassifikationsverfahren wie Neuronale Netze oder SVMs[6] anbieten, siehe für weitere Mustererkennungsverfahren auch (Duda u. a., 2001).
Spricht man von Ähnlichkeit innerhalb der Gruppen und Unähnlichkeit zwischen den Gruppen, so müssen diese Begriffe definiert werden, üblicherweise über ein Ähnlichkeits- oder Distanzmaß. Bei Mikroarraydaten wird häufig der Pearsonsche Korrelationskoeffizient verwendet. Die

[6] Abkürzung des englischen Begriffs 'Support Vector Machines', frei übersetzt als 'Stützvektorverfahren'

Distanzen zwischen den Objekten seien in der Matrix $D = (d_{st})_{s,t=1,\ldots,n}$ beschrieben, so dass d_{st} den Abstand zwischen Objekt s und Objekt t beschreibt. Hierarchisches Clustern erzeugt einen Baum, dessen Wurzel ein Cluster ist, in dem sämtliche Objekte enthalten sind und dessen Blätter jeweils die Objekte selbst sind, also Cluster, die nur ein Objekt enthalten. Den Rest des Baumes erhält man im 'agglomerativen' Verfahren, indem man mit den Blättern beginnt und in jedem Schritt jeweils die zwei Cluster vereinigt, die die geringste Distanz zueinander aufweisen, bis sämtliche Objekte in einem Cluster sind, also die Wurzel erreicht ist. Im 'divisiven' Verfahren beginnt man mit der Wurzel und spaltet die Cluster, bis man die Blätter erreicht hat. Nach jeder Vereinigung beziehungsweise Spaltung von Clustern wird eine neue Distanzmatrix berechnet, die die Distanzen zwischen den Clustern enthält, die im neuen Schritt entstehen. Es gibt verschiedene Möglichkeiten, die neue Distanzmatrix zu berechnen. Die Definition der Distanz zwischen zwei Clustern ist unter anderem durch die kleinste Distanz (single linkage genannt), größte Distanz (complete linkage) oder durchschnittliche Distanz (average linkage) möglich.
Hierarchisches Clustern von Transkriptionsdaten ist zwar sehr weit verbreitet, hat aber den großen Nachteil, dass während des iterativen Vorgehens einmal begangene 'Fehler', also beispielsweise ein Zusammenfügen zweier Objekte, deren Distanz zwar gering ist, die jedoch von anderen Clustermethoden in zwei verschiedene Cluster eingeordnet werden würden, nicht mehr rückgängig gemacht werden.
Der Cluster-Algorithmus k-means macht diesen Fehler nicht, da die Objekte aus einmal zugeordneten Clustern wieder entfernt werden können, benötigt dafür jedoch die Anzahl und eine anfängliche Lage der Cluster als Eingabe. Zunächst werden die Objekte beliebig in die k Cluster aufgeteilt und danach die Zuordnungen derart verändert, dass die Distanzen zu den Clusterschwerpunkten minimiert werden. Anschließend werden die Clusterschwerpunkte aufgrund der Zuordnungen neu berechnet und wiederum die Objekte ihrem nächsten Clusterschwerpunkt zugeteilt. Ändert sich diese Zuordnung nicht mehr, so heißt es allerdings nicht, dass die Optimallösung gefunden wurde, die das globale Minimum der Summe aller Distanzen zwischen Objekten und den dazugehörigen Clusterschwerpunkten ergibt. Daher werden üblicherweise mehrere Durchgänge mit verschiedenen Anfangsclustern durchgeführt.
Bei Vorliegen von Zeitreihen ändern Permutationen der Zeitpunkte die Clusterergebnisse in den beiden zuvor genannten Algorithmen nicht. Um die Dynamik in Zeitreihen zu berücksichtigen, wurden spezielle Ansätze wie von Schliep u. a. (2003) gemacht, die ein modellbasiertes Clustern mithilfe von Hidden Markov Modellen durchführen. Ein weiterer Vorteil dieser Methode ist die mögliche Zuordnung eines Gens zu mehreren Clustern, die weder im hierarchischen Clustern noch bei k-means möglich ist.

In diesem Unterkapitel sind einige wichtige Analysemethoden für Expressionsdaten skizziert oder angesprochen worden, jedoch würde eine vollständige Auflistung den Rahmen einer Einbettung der Modellierung genregulatorischer Netzwerke in verwandte Methoden für Mikroarray-Analysen sprengen und es sei auf Duda u. a. (2001); Schena (1999); Slonim (2002); Butte (2002) verwiesen.

1.2 Problembeschreibung

Es seien n Gene gegeben. Weiterhin seien $\hat{x}_i(t)$, $i = 1,...,n$ für $t = t_0,...,t_l$ die $l+1$ Messungen der Konzentration der mRNA-Transkripte von Gen i zur Zeit t oder die entsprechenden Proteinkonzentrationen, falls Proteinmessungen statt mRNA-Messungen verwendet werden. Ein genregulatorisches Netzwerk definieren wir folgendermaßen:

Definition 1.2. *(genregulatorisches Netzwerk)*
Ein gerichteter Graph $G = (V,E)$ mit n Knoten heißt genregulatorisches Netzwerk, wenn die Kanten derart sind, dass für jedes Paar von Knoten $\{v_i, v_j\}$ mit $v_i, v_j \in V$ die gerichteten Kanten $(v_i,v_i),(v_j,v_j),(v_i,v_j)$ und $(v_j,v_i) \in E$ existieren. Außerdem ist jedem Knoten v_i die Variable x_i zugeordnet für $i = 1,...,n$ und jeder gerichteten Kante (v_j,v_i) eine Funktion $f_{i,j}(x_j) : \mathbf{R} \to \mathbf{R}$, die abhängig von x_j ist. Es ist erlaubt, dass diese Funktionen konstant gleich Null sind.

Die Variablen x_i entsprechen der mRNA-Konzentration von Gen i, die Funktion $f_{i,j}(x_j)$ für $i,j = 1,...,n$ beschreibt die Regulierung, der Gen i durch Gen j ausgesetzt ist. Genauer gesagt beschreiben sie die Veränderung der Konzentration des Produkts von Gen i, die die Konzentration des Produkts von Gen j hervorruft. Wenn sämtliche Kanten (v_j,v_i) entfernt werden, deren zugehörige Funktion $f_{i,j}(x_j)$ konstant gleich Null ist, so muss der Graph nicht notwendig zusammenhängend sein. Die Entfernung dieser Kanten liefert ein genregulatorisches Netzwerk wie es üblicherweise in der Literatur abgebildet wird. Hier wird meist ein Plus an die Kante (v_j,v_i) geschrieben, wenn $f_{i,j} \geq 0$, und ein Minus, wenn $f_{i,j} \leq 0$ gilt, siehe Abbildung 1.3.

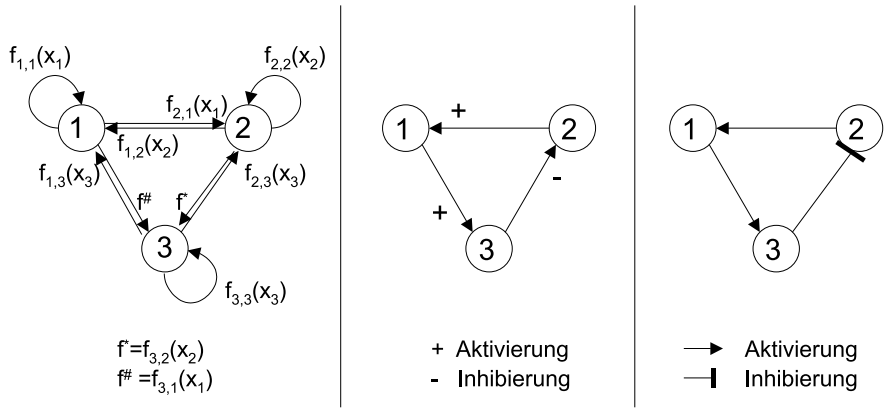


Abbildung 1.3: Links ist ein genregulatorisches Netzwerk dargestellt, in der jeder gerichteten Kante (j,i) eine Funktion $f_{i,j}(x_j)$ zugeordnet ist. In der Mitte und rechts ist dieses Netzwerk vereinfacht dargestellt, indem sämtliche Kanten, deren zugeordnete Funktionen konstant gleich Null sind, entfernt wurden, und die übrigen Kanten für positive Funktionen mit $+$ (Mitte) bzw. $\rightarrow$ (rechts) gekennzeichnet sind und für negative Funktionen mit $-$ (Mitte) bzw. $\dashv$ (rechts).

Genau genommen müssten nach der genannten Definition genregulatorische Netzwerke 'genregulatorische Graphen' heißen. Um bei den üblichen Bezeichnungen zu bleiben, wird jedoch

weiterhin der Begriff ‘genregulatorisches Netzwerk’ verwendet. Für genregulatorische Netzwerke gilt in den meisten Fällen, dass der Großteil der Funktionen $f_{i,j}$ in solchen Netzwerken konstant gleich Null ist.

Es ergeben sich zwei Problemstellungen:

Problemstellung 1.3. *(Inferenz von genregulatorischen Netzwerken)*
Zugrunde gelegt sei ein Modell für den zeitlichen Verlauf der mRNA-Konzentrationen $x(t) = (x_1(t),...,x_n(t))^t$. *Die Funktionen* $f_{i,j}$, $i,j = 1,...,n$ *seien in diesem Modell die Einflüsse von Gen* j *auf Gen* i *und seien in parametrisierter Form vorgegeben. Die Problemstellung lautet dann: Man sucht mithilfe der Daten* $\hat{x}(t_0),...,\hat{x}(t_l)$ *mit* $\hat{x}(t_k) = (\hat{x}_1(t_k),...,\hat{x}_n(t_k))^t$ *für* $k = 0,...,l$ *die Parameter der Funktionen* $f_{i,j}$ *für sämtliche Kanten* (v_j, v_i), $i,j = 1,...,n$, *so dass das zugrunde gelegte Modell für* $x(t)$ *ein vorgegebenes Ziel erfüllt. Dieses Ziel kann beispielsweise die Minimierung des Fehlers zwischen der Vorhersage des Modells und den Messungen* $\hat{x}(t)$ *sein, also zum Beispiel die Minimierung der von den Parametern* θ *abhängigen Funktion*

$$F(\theta) \quad = \quad \sum_{i=1}^{n} \sum_{j=0}^{l} ||x_i(t_j,\theta) - \hat{x}_i(t_j)||^2. \tag{1.1}$$

Hier liegt der Schwerpunkt auf der Suche nach der Struktur des Netzwerkes, es soll also die Frage beantwortet werden, welche Gene sich gegenseitig beeinflussen. Im besten Fall sind die Gene des gesamten Genoms eines Organismus gemessen worden und gesucht sind Regulierungen zwischen allen Genen. Dieses Unterfangen scheitert bisher daran, dass die Anzahl der Gene sehr viel größer ist als die Anzahl der Messungen. Eine bisher übliche Zeitreihenmessung umfasst ca. 5-50 Zeitpunkte, die Anzahl der Gene bei Prokaryonten liegt meist bei mehreren Tausend. Es existieren jedoch Teilmengen von Genen, die stark miteinander verbunden sind in dem genregulatorischen Netzwerk, die jedoch wenige Verbindungen zu Genen außerhalb dieser Teilmenge besitzen, siehe hierzu auch das dritte Kapitel. Daher läuft die Antwort der oben genannten Problemstellung oft in Einschränkung auf diese Teilmengen hinaus. Sind jedoch solche stark miteinander verbundenen Teilmengen bekannt, für die das Problem gelöst werden soll, so sind diese Gene oft gut untersucht, so dass auch schon die Frage beantwortet werden kann, welche Gene dieser Teilmenge welche anderen Gene aus dieser Teilmenge regulieren. Dadurch ergibt sich die zweite Problemstellung:

Problemstellung 1.4. *(Modellierung bei gegebener Struktur)*
Wieder sei ein zugrunde gelegtes Modell für den zeitlichen Verlauf der mRNA-Konzentrationen $x(t)$ *wie in der ersten Problemstellung gegeben. Es seien außerdem sämtliche* $i,j \in \{1,...,n\}$ *gegeben, für die* $f_{i,j}$ *konstant gleich Null ist, das heißt, es ist bekannt, welche Gene Regulierungen auf andere Gene ausüben. Die Art der Regulierung darf unbekannt sein. Man sucht nun die Parameter der verbleibenden Funktionen* $f_{i,j}$, *so dass das zugrunde gelegte Modell für* $x(t)$ *wiederum ein vorgegebenes Ziel erfüllt. Wie eben kann dieses Ziel beispielsweise die Minimierung*

von (1.1) sein.

Im zweiten Kapitel wird das eigene zugrunde gelegte Modell entwickelt, im vierten Kapitel wird eine allgemeine Lösung für die zweite Fragestellung vorgestellt und in den Kapiteln fünf und sechs wird die zweite Fragestellung für zwei geeignete Teilmengen von zwei verschiedenen Organismen beantwortet.

1.3 Lösungsansätze

Je nach Modellansatz gibt es verschiedene Lösungen für die genannten Probleme. Wie bei jeder mathematischen Modellierung von Netzwerken, gleich ob es sich um ein biochemisches, physikalisches, soziales oder wirtschaftswissenschaftliches Netzwerk handelt, muss ein Kompromiss zwischen Abstraktion und Genauigkeit gefunden werden, der zuvorderst abhängig ist von der zu beantwortenden Fragestellung, jedoch auch die vorhandene Datenlage berücksichtigen muss. Es muss also ein Abstraktionsniveau gefunden werden, der diejenigen Eigenschaften im Fokus behält, die für die Beantwortung der Frage wichtig sind, jedoch weniger wichtige Details außen vor lässt (Szallasi u. a., 2006). Neben der Detailliertheit des Modells müssen auch folgende Fragen geklärt werden:

- Es können bezüglich der Zeit diskrete oder stetige Modelle gewählt werden. Das Modell kann also in diskreten Zeitschritten eine Vorhersage von $x(t+1)$ liefern oder stetig bezüglich der Zeit $\frac{dx}{dt}$ beschreiben. Diskrete Modelle sind vorteilhafter, wenn die Variablen sich nur in Abhängigkeit von speziellen Ereignissen ändern, stetige Modelle sollten bevorzugt werden, wenn die Variablen sich durchgehend ändern (Kell u. Knowles, 2006).

- Es können deterministische oder stochastische Modelle gewählt werden. Die meisten Phänomene sind nicht stochastisch, jedoch werden oft nicht sämtliche Details in das Modell einbezogen, so dass es nützlich sein kann, ein stochastisches Modell zu wählen.

- Modelle können räumliche Aspekte mit einbeziehen oder räumliche Homogenität annehmen. Weist das biologische System starke Heterogenität bezüglich der räumlichen Ereignisse auf, so ist es sinnvoll, räumliche Aspekte in die Modellierung mit einzubeziehen (Kell u. Knowles, 2006). Beispielsweise ist in Eukaryonten ein Zellkern vorhanden, aus dem die mRNA heraustransportiert wird, bevor die Translation stattfindet. Hier kann es sinnvoll sein, zwischen Zellkern und Zytoplasma zu unterscheiden.

Die Modellierung unter Einbeziehung von räumlicher Heterogenität wird im fünften Kapitel näher beschrieben. Verschiedene Ansätze für diskrete, stetige, deterministische und stochastische Modelle werden in den folgenden Unterkapiteln mit ihren jeweiligen Vor- und Nachteilen erläutert.

1.3.1 Boolesche Netzwerke

Boolesche Netzwerke sind sehr einfache Modelle, die für die Beschreibung von Genregulierungen verwendet werden, siehe auch Smolen u. a. (2000); de Jong (2002); Somogyi u. Snie-

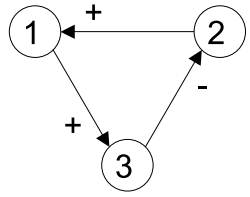
goski (1996). Sie sind in ihrer grundlegenden Form diskret und deterministisch. Trotz ihrer Einfachheit können sie selbst für große Netzwerke gute qualitative Aussagen machen, so dass sie weiterhin Verwendung finden. Für den Vektor der mRNA-Konzentrationen gilt die Annahme $x(t) = (x_1(t), ..., x_n(t))^t \in \{0,1\}^n$. Die möglichen Zustände des genregulatorischen Netzwerkes belaufen sich dadurch auf 2^n Zustände. Da die mRNA-Konzentrationen nur die Werte 0 und 1 zugeordnet bekommen können, spricht man hier auch von den zugehörigen Genen, welche aktiv (entspricht 1) und inaktiv (entspricht 0) sein können.

Definition 1.5. *(Boolesches Netzwerk, (Thierolf, 2003))*
Das Paar (X,B) mit Genen $X = \{x_1, ..., x_n\}$ und Booleschen Funktionen $B = \{b_1, ..., b_n\}$ heißt Boolesches Netzwerk, wenn der Vektor $x(t+1) = (x_1(t+1), ..., x_n(t+1))^t$ über die Gleichungen

$$x_i(t+1) = b_i(x(t)) \tag{1.2}$$

für jedes $i = 1, ..., n$ berechnet wird.

Zwei mögliche Visualisierungen ergeben sich durch Input-Output-Tabellen und Übergangsgraphen, siehe auch Abbildung 1.4. Die Tabellen bestehen aus zwei Spalten, in denen die erste



$x_1(t+1)=x_2(t)$

$x_2(t+1)=\neg x_3(t)$

$x_3(t+1)=x_1(t)$

Eingabe x(t)			Ausgabe x(t+1)		
x_1	x_2	x_3	x_1	x_2	x_3
0	0	0	0	1	0
0	0	1	0	0	0
0	1	0	1	1	0
1	0	0	0	1	1
0	1	1	1	0	0
1	0	1	0	0	1
1	1	0	1	1	1
1	1	1	1	0	1

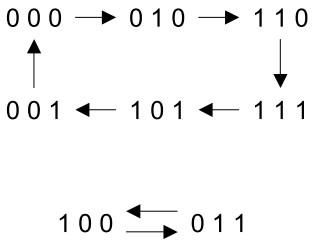


Abbildung 1.4: Links ist ein genregulatorisches Netzwerk mit zugehörigen Booleschen Funktionen dargestellt. In der Mitte befindet sich die zugehörige Input-Output-Tabelle, rechts der Übergangsgraph.

Spalte alle möglichen Ausgangswerte für $x(t)$, also die Vektoren $(x_1, ..., x_n)^t \in \{0,1\}^n$, und die zweite Spalte die jeweiligen neuen Zustände der Gene auflistet. Da die Anzahl der möglichen Zustände endlich ist, muss jeder Anfangsvektor $x(t_0)$ in einem Attraktor enden. Dies kann ein Punktattraktor sein, das heißt, der Zustand verändert sich ab einem Zeitpunkt $\tilde{t}$ nicht mehr und es gilt $x(t+1) = x(t)$ für $t \geq \tilde{t}$, oder es kann ein periodischer Attraktor sein, das heißt, die Abfolge von Zuständen wird ab einem Zeitpunkt $\tilde{t}$ periodisch und es gilt $x(t+k) = x(t)$ für ein festes $k \geq 2$ und $t \geq \tilde{t}$. Um die Attraktoren zu veranschaulichen, ist die zweite Visualisierungsmöglichkeit mit Übergangsgraphen gut geeignet. Die n Knoten des Graphen entsprechen den n möglichen Zuständen. Der Übergang von einem Zustand zum nächsten Zustand ist deterministisch

und wird durch eine gerichtete Kante dargestellt. Kreise in den Übergangsgraphen entsprechen dann den Attraktoren. Im Beispiel 1.4 lassen sich zwei zyklische Attraktoren erkennen. Da ein deterministisches Modell vorliegt, ist durch den Anfangszustand bereits festgelegt, in welchem der beiden Attraktoren das System bleibt.

Definition 1.6. *(Konsistenz, (Thierolf, 2003))*
Es sei eine Menge $M = \{(I_0, O_0), ..., (I_{l-1}, O_{l-1})\}$ von l Input-Output-Paaren gegeben. Ein Paar (I_j, O_j) für ein $j \in \{0, ..., l-1\}$ besteht aus Messungen der mRNA-Konzentrationen sämtlicher Gene zu den zwei aufeinanderfolgenden Zeitpunkten t_j, t_{j+1}. Die Messungen zum ersten Zeitpunkt t_j sind in I_j enthalten, die Messungen zum darauf folgenden Zeitpunkt t_{j+1} in O_j, das heißt, $I_j = (I_j(x_1), ..., I_j(x_n))$ und $O_j = (O_j(x_1), ..., O_j(x_n))$. Ein Boolesches Netzwerk (G, B) mit Genen $G = \{x_1, ..., x_n\}$ und Booleschen Funktionen $B = \{b_1, ..., b_n\}$ heißt konsistent mit der Menge M, wenn $O_j(x_i) = b_i(I_j(x_1), ..., I_j(x_n))$ gilt.

Mit dieser Definition von Konsistenz lassen sich nun folgende Probleme definieren:

Problemstellung 1.7. *(Thierolf, 2003)*
Das Konsistenzproblem stellt die Frage, ob zu gegebenem $M = \{(I_0, O_0), ..., (I_{l-1}, O_{l-1})\}$ und gegebenem $n \in \mathbb{N}^+$ ein Boolesches Netzwerk (G, B) existieren kann, welches konsistent mit M ist.
Das Zählproblem fragt nach der Anzahl dieser konsistenten Booleschen Netzwerke und das Enummerierungssproblem bezieht sich darauf, sämtliche dieser konsistenten Netzwerke aufzulisten.
Das Identifikationsproblem wiederum fragt danach, ob es zu gegebenem M und n ein eindeutiges Netzwerk gibt und wie dieses im Existenzfall aussieht.

Insgesamt gibt es für n Gene 2^n Expressionsmuster. Um aus einer Menge M auf ein konsistentes Boolesches Netzwerk zu schließen, müsste man alle diese Expressionsmuster untersuchen. Daher würde die Laufzeit exponentiell mit der Anzahl von Genen anwachsen. Bekannte Algorithmen zur Lösung der oben genannten Problemstellungen führen daher einen 'inneren Grad K' ein, der eine obere Schranke der Anzahl von beeinflussenden Genen definiert. In dem gesuchten Modell, das auf einem Booleschen Netzwerk basiert, können also nicht mehr sämtliche der n Gene einen Einfluss auf ein Gen haben, sondern maximal K Gene. Ein sehr einfacher Algorithmus, der alle vier Fragen nach der Konsistenz, der Anzahl, der Enummerierung und der Existenz lösen kann, ist in Akutsu u. a. (1999) vorgestellt worden:

Bemerkung 1.8. *(BOOL-1, (Akutsu u. a., 1999))*
Nacheinander werden x_i mit $i = 1, ..., n$ untersucht. Für sämtliche K-Kombinationen von Genen $\{i_1, ..., i_K\}$ mit $i_1, ..., i_K \in \{1, ..., n\}$ und für alle Booleschen Funktionen $b_1, ..., b_s$ mit $b_r : \{0,1\}^K \to \{0,1\}$ für $r = 1, ..., s$ wird untersucht, ob für x_i

$$O_j(x_i) \quad = \quad b_r(I_j(x_{i_1}), ..., I_j(x_{i_K})) \tag{1.3}$$

für sämtliche Paare $(I_j, O_j) \in M$ gilt. Sämtliche Booleschen Funktionen b_r mit der zugehörigen K-Kombination von Genen $\{i_1, ..., i_K\}$, für die (1.3) gilt, können für die Definition eines

konsistenten Booleschen Netzwerkes verwendet werden. Existiert ein x_i mit $i \in \{1,...,n\}$, für das keine Boolesche Funktion für eine K-Kombination von Genen gefunden werden kann, so ist die Existenzfrage negativ zu beantworten. Andernfalls können durch Auflistung der Funktionen sämtliche Booleschen Netzwerke, die konsistent mit M sind, definiert werden. Für einen festen inneren Grad K löst der Algorithmus BOOL-1 das Konsistenzproblem polynomial in n, siehe Akutsu u. a. (1999); Thierolf (2003).

Einer der bekanntesten Algorithmen, die einen inneren Grad K vorgeben, ist REVEAL[7] (Liang u. a., 1998). Dieser betrachtet nicht sämtliche Paare (I_j, O_j), sondern nur eine Teilmenge, da nur eine Teilmenge benötigt wird, um - im Existenzfall - ein eindeutiges Netzwerk zu bestimmen. Ein Boolesches Netzwerk ist bereits eindeutig bestimmt, wenn sämtliche $2K$-elementigen Teilmengen von G mit allen 2^{2K} möglichen Werten als I_j in M erscheinen (Thierolf, 2003).

Der große Vorteil des Modellierungsansatzes mit Booleschen Netzwerken liegt darin, dass durch die einfache Modellierung mit ein- oder ausgeschalteten Genen Algorithmen existieren, die große Datenmengen verarbeiten können und daher dieser Ansatz nicht nur auf kleine Subnetzwerke beschränkt ist. Dieser Ansatz ist also vorrangig geeignet, um die Problemstellung 1.3 zu lösen. Kleinere Probleme, wie beispielsweise die Berücksichtigung von Messfehlern, können dadurch behoben werden, dass die Algorithmen entsprechend angepasst werden. In dem Fall der Berücksichtigung von Messfehlern führt beispielsweise eine Inkonsistenz mit einem Paar (I_j, O_j) nicht automatisch zum Verwerfen eines Netzwerkes. Die sehr vereinfachte Modellierung durch nur zwei Zustände für jedes Gen lässt jedoch vermuten, dass selbst das qualitative Verhalten eines Systems durch solch ein Modell nur in seltenen Fällen zufriedenstellend beschrieben werden kann. Vergleicht man die komplexen biochemischen Prozesse mit der binären Beschreibung, erscheinen erfolgreiche Anwendungen dieses Modells kaum möglich. Mehrere Untersuchungen haben jedoch ergeben, dass in vielen Fällen diese Modellierung bereits ausreicht, um das qualitative Verhalten des Systems beschreiben zu können (Thomas u. Kaufman, 2001; Glass u. Kauffman, 1973). In Thomas (1998) wurde beispielsweise gezeigt, dass sich das qualitative Verhalten eines Modells, das auf speziellen nichtlinearen Differenzialgleichungen basierte[8] und durch ein Boolesches Netzwerk approximiert wurde, nicht änderte. Des Weiteren sind viele dynamische Verhaltensweisen, die in der Biologie häufig vorkommen, wie Multistationarität, Hysterese oder Oszillationen, durch Boolesche Netzwerke darstellbar (Shmulevich u. a., 2002a,b). Diese positiven Ergebnisse nehme ich unter anderem zum Anlass, bei dem in Kapitel 2 vorgestellten Modell, das eine detailliertere und quantitative Beschreibung der Dynamik verfolgt, dennoch weiterhin starke Vereinfachungen zuzulassen, wenn diese die Analyse- oder Berechnungsmöglichkeiten des Modells verbessern.

1.3.2 Probabilistische Netzwerke

Kennt man jegliche Regulierungsschritte wie unter anderem sämtliche Details des Bindens der Transkriptionsfaktoren an die DNA auf molekularer Ebene, so könnte man zu dem Schluss kommen, dass detaillierte Modelle deterministisch sein müssen. Allerdings können trotz gleicher

[7] **Rev**erse **E**ngineering **Al**gorithm

[8] Hier wurden sogenannte Hillfunktionen verwendet.

Anfangszustände unterschiedliche Verhaltensweisen beobachtet werden, was darauf zurückgeführt wird, dass individuelle Synthese und individueller Abbau zweier gleicher Moleküle leicht unterschiedlich sein können.
Selbst wenn deterministische Prozesse zugrunde gelegt werden könnten, so würden sie in den Messdaten stochastisch erscheinen, da nicht sämtliche Variablen beachtet werden können (Murphy u. Mian, 1999). Ein stochastisches Modell, nämlich die Modellierung mit Bayesschen Netzwerken, wird im Folgenden beschrieben.

Bayessche Netzwerke

Bayessche Netzwerke - diskrete und stochastische Modelle - wurden in Pearl (1988) eingeführt und von Friedman u. a. (2000) auf Genexpressionsdaten sowie von Markowetz (2006) auf allgemeinere biochemische Prozesse angewendet.

Definition 1.9. *(Bayessche Netzwerke, (Friedman u. a., 2000))*
Ein Bayessches Netzwerk ist ein Paar (G,Θ), das aus einem gerichteten azyklischen Graphen $G=(V,E)$ und bedingten Verteilungen Θ besteht. Die Knoten $V=\{v_1,...v_n\}$ von G entsprechen den n Genen. Ihnen zugeordnet sind n Zufallsvariablen $X_1,...,X_n$ über dem Ereignisraum Ω, der die diskretisierten Expressionen der n Gene enthält. Die Kanten E entsprechen den bedingten Abhängigkeiten zwischen den einzelnen Variablen. Der Vektor $\Theta=(p_1,...,p_n)^t$ enthält die bedingten Verteilungen $p_i=p(X_i|Eltern(X_i))$ für $i=i_1,...,i_l$, $\{i_1,...,i_l\}\subset\{1,...,n\}$, und die unbedingten Verteilungen $p_i=p(X_i)$ für $i=i_{l+1},...,i_n$, $\{i_{l+1},...,i_n\}=\{1,...,n\}\setminus\{i_1,...,i_l\}$. Hierbei gilt, dass die Variablen $X_{i_1},...,X_{i_l}$ jeweils die Eltern $Eltern(X_{i_1}),...,Eltern(X_{i_l})$ im Graphen G besitzen, wohingegen die Variablen $X_{i_{l+1}},...,X_{i_n}$ keine Eltern in G besitzen. Aufgrund der azyklischen Eigenschaft von G ist $\{i_{l+1},...,i_n\}\neq\emptyset$.

Nun lässt sich die gemeinsame Verteilung $p(X_1,...,X_n)=p(X_1)p(X_2|X_1)\cdot...\cdot p(X_n|X_{n-1},...,X_1)$ durch

$$p(X_1,...,X_n) \quad = \quad \prod_{i=1}^{n} p_i \tag{1.4}$$

beschreiben.

Ein Bayessches Netzwerk ist also eine graphische Darstellung einer gemeinsamen Wahrscheinlichkeitsverteilung.

Bei Problemstellung 1.4 ist der Graph G bereits vorgegeben, so dass nur noch die Parameter der bedingten Verteilungen Θ geschätzt werden müssen. Dazu wird zumeist ein Maximum-Likelihood-Ansatz gewählt, wenn für sämtliche Variablen Messungen vorliegen (Murphy u. Mian, 1999). Ein Vorteil dieses Modells liegt darin, dass auch nicht messbare Variablen berücksichtigt werden können bzw. der Graph als nur unvollständig bekannt angenommen werden kann. Ein Überblick über die Parameterschätzung in diesen verschiedenen Spezialfällen ist in Murphy u. Mian (1999) aufgelistet.

Der Vorteil der Bayesschen Netzwerke liegt darin, dass die stochastische Modellierung zum einen die stochastischen Prozesse in der Zelle widerspiegelt sowie zum anderen den stark verrauschten experimentellen Daten entspricht. Auch kann das Modell dahin gehend erweitert werden, nicht beobachtbare Variablen und bereits vorhandenes biologisches Wissen über Abhängigkeiten zwischen Genen zu berücksichtigen. Dem Nachteil, die Dynamik der Genexpression nicht einzufangen, kann durch sogenannte dynamische Bayessche Netzwerke (Perrin u. a., 2003) begegnet werden. Allerdings ist die Kreisfreiheit für genregulatorische Netzwerke eine sehr einschränkende Annahme, welche viele Subnetzwerke ausschließt, da Kreise bei genregulatorischen Netzwerken sehr häufig beobachtet werden.

Mastergleichung

Eine weitere Klasse von Modellen sind die sogenannten Mastergleichungen, die wie auch die Bayesschen Netzwerke nichtdeterministische Gestalt haben. Es sei $x = (x_1(t),...,x_n(t))^t \in \mathbb{N}^n$ der Vektor, der die Anzahl $x_1,..,x_n$ der Moleküle $1,...,n$ zum Zeitpunkt t beinhaltet. Die Verteilung $p(x,t)$ gibt die Wahrscheinlichkeit zum Zeitpunkt t an, dass x_i Moleküle der Art i in der Zelle enthalten sind. Es gebe m mögliche Reaktionen und es sei $\alpha_j \Delta t$, $j = 1,...,n$, die Wahrscheinlichkeit, dass Reaktion j im Zeitintervall $[t,\Delta t]$ stattfindet, wenn das System im Zustand x zum Zeitpunkt t ist. Des Weiteren sei $\beta_j \Delta t$, $j = 1,...,n$, die Wahrscheinlichkeit, dass Reaktion j das System in den Zustand x bringt im Zeitintervall $[t,\Delta t]$. Die zeitliche Veränderung von $p(x,t)$ lässt sich dann darstellen als

$$p(x,t+\Delta t) \quad = \quad p(x,t)\left(1-\sum_{j=1}^{m}\alpha_j \Delta t\right)+\sum_{j=1}^{m}\beta_j \Delta t, \tag{1.5}$$

siehe auch Gillespie (1977). Betrachtet man den Grenzwert $\Delta t \to 0$, so erhält man die sogenannte Mastergleichung (Gillespie, 1977; de Jong, 2002):

$$\frac{\mathrm{d}p(x,t)}{\mathrm{d}t} \quad = \quad \sum_{j=1}^{m}(\beta_j - \alpha_j p(x,t)). \tag{1.6}$$

Im Gegensatz zu gewöhnlichen Differenzialgleichungen sind numerische Simulationen der Mastergleichung sehr aufwendig. Zur einfacheren Simulation wurden verschiedene Approximationen betrachtet, zum Beispiel in Gillespie (1977) der stochastische Simulationsalgorithmus. Dieser Algorithmus geht von einem Zustand aus und bestimmt aufgrund von Wahrscheinlichkeitsverteilungen, wann die nächste Reaktion stattfindet und welche der m möglichen Reaktionen diese ist. Daraufhin wird der nächste Zustand berechnet, indem zu dem zuvor bestimmten Zeitpunkt die zuvor bestimmte Reaktion ausgeführt wird.
Laut Smolen u. a. (2000) werden jedoch deterministische Modelle wie Differenzialgleichungen oder Boolesche Netzwerke auf lange Sicht weiterhin wichtig bleiben, da die Datenlage nicht ausreicht, um befriedigende stochastische Modelle zu konstruieren.

1.3.3 Differenzialgleichungen

Liegt der Fokus bei der Modellierung genregulatorischer Netzwerke nicht auf der Fragestellung, welches Gen welches andere Gen reguliert, sondern eher auf der quantitativen zeitlichen

Entwicklung der Genexpression, so sind Differenzialgleichungen und Differenzengleichungen prädestiniert, da sie beide die zeitliche Veränderung einer Variable beschreiben.
Eine gewöhnliche Differenzialgleichung ist eine Gleichung, die eine unbekannte Funktion $y(x)$ sowie Ableitungen dieser Funktion enthält, sie wird also durch

$$f(x,y,y',...,y^{(k)}) = 0 \tag{1.7}$$

beschrieben. Die Funktion $f : G \to \mathbf{R}$ sei auf dem Gebiet $G \subseteq \mathbf{R}^{k+2}$ definiert und stetig. Eine Lösung y von (1.7) ist eine reellwertige Funktion, die auf einem Intervall $I \subseteq \mathbf{R}$ erklärt und dort k-mal stetig differenzierbar ist und für die gilt:

$$(x,y(x),...,y^{(k)}(x)) \in G \text{ für } x \in I \text{ und} \tag{1.8}$$

$$f(x,y(x),...,y^{(k)}(x)) = 0 \text{ für } x \in I, \tag{1.9}$$

siehe auch Behnke (2003). Gleichungen, die nach der höchsten vorkommenden Ableitung aufgelöst sind, heißen explizit, andernfalls heißen sie implizit. Explizite Gleichungen haben also die Form

$$y^{(k)} = g(x,y,y',...,y^{(k-1)}) \tag{1.10}$$

mit $g : G \to \mathbf{R}$ auf dem Gebiet $G \subseteq \mathbf{R}^{k+1}$ definiert und stetig.
Differenzialgleichungen k-ter Ordnung sind dadurch definiert, dass die höchste existierende Ableitung $y^{(k)}$ ist. Systeme von Differenzialgleichungen erster Ordnung bestehen aus n Gleichungen

$$\begin{aligned} \frac{\mathrm{d}y_1(x)}{\mathrm{d}x} &= f_1(y_1(x),...,y_n(x)) && (1.11)\\ &\vdots \quad \vdots \\ \frac{\mathrm{d}y_n(x)}{\mathrm{d}x} &= f_n(y_1(x),...,y_n(x)), && (1.12) \end{aligned}$$

kurz $y' = f(y)$ mit $y = (y_1,...,y_n)^t \in \mathbf{R}^n$ und $f = (f_1,...,f_n)^t : \mathbf{R}^n \to \mathbf{R}$. Wir möchten sich zeitlich verändernde Systeme beschreiben, so dass wir im Folgenden die Variable x durch die Zeitvariable t ersetzen sowie die Funktion $y(x)$ durch $x(t)$. Systeme erster Ordnung, deren unabhängige Variable die Zeit t ist, heißen auch dynamische Systeme. Der Vektor $x(t)$ heißt auch Zustand des Systems zum Zeitpunkt t und die Gesamtheit aller möglichen Zustände eines Systems nennen wir Zustandsraum. Jede Lösung beschreibt eine Kurve im $\mathbf{R}^{n+1}$ durch $\{(t,x(t)) \in \mathbf{R}^{n+1} : t \in I\}$, die auch Lösungskurve genannt wird. Eine Trajektorie ist die Bildmenge $\{x(t) \in \mathbf{R}^n : t \in I\}$, also eine Kurve im $\mathbf{R}^n$. Hat man einen Zustand $\bar{x} \in \mathbf{R}^n$, der sich für $t \to \infty$ nicht mehr verändert, so nennt man $\bar{x}$ einen Gleichgewichtspunkt oder Fixpunkt. Die Definition für einen Fixpunkt $\bar{x}$ einer Differenzialgleichung $\mathrm{d}x/\mathrm{d}t = f(x)$ lautet, dass $f(\bar{x}) = 0$ gilt.
Wir betrachten im Folgenden reellwertige lineare Systeme von Differenzialgleichungen mit konstanten Koeffizienten. Dies sind Gleichungen der Form

$$\begin{aligned} \frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= a_{11}x_1(t) + ... + a_{1n}x_n(t) + b_1 && (1.13)\\ &\vdots \quad \vdots \\ \frac{\mathrm{d}x_n(t)}{\mathrm{d}t} &= a_{n1}x_1(t) + ... + a_{nn}x_n(t) + b_n && (1.14) \end{aligned}$$

mit konstanten Koeffizienten $a_{ij}, b_i \in \mathbf{R}$, $i, j = 1, ..., n$, kurz $\frac{dx(t)}{dt} =: \dot{x}(t) = Ax(t) + b$ mit $A \in \mathbf{R}^{n \times n}$ und $b \in \mathbf{R}^n$. Für $b = 0$ heißt solch ein System von Differenzialgleichungen homogen, für $b \neq 0$ inhomogen.
Geben wir eine Anfangswertbedingung an, das heißt eine Bedingung $x(\tau) = x_0$ mit $x_0 \in \mathbf{R}^n$, $\tau \in \mathbf{R}$, die beschreibt, in welchem Zustand das System zu dem Zeitpunkt τ ist, so gibt es für die von uns betrachteten Systeme von Differenzialgleichungen eine eindeutige Lösung. Dies gilt nach dem Satz von Picard-Lindelöf sogar für allgemeinere Systeme von Differenzialgleichungen:

Satz 1.10. *(Satz von Picard-Lindelöf für Systeme von Differenzialgleichungen, siehe Natterer (1998))*
Es sei

$$\dot{x}(t) \quad = \quad f(t, x(t)), \;\; x(\tau) = \eta \tag{1.15}$$

mit f stetig in einer Umgebung J von $(\tau, \eta) \in \mathbf{R}^{n+1}$. Des Weiteren erfülle f in dieser Umgebung J eine Lipschitz-Bedingung, das heißt, es existiert eine Konstante L, so dass für $(t,x), (t,y) \in J$

$$||f(t,x) - f(t,y)|| \quad \leq \quad L||x - y|| \tag{1.16}$$

gilt, wobei $||\cdot||$ eine beliebige Norm im $\mathbf{R}^n$ sei. Dann gibt es ein Intervall I mit $x_0 \in I$ und eine Funktion $x(t) : I \to \mathbf{R}$, die in I stetig differenzierbar ist und die Lösung von 1.15 in I ist.

Im Folgenden gehen wir auf verschiedene Modellierungsansätze, die auf Differenzialgleichungen basieren, ein.

Lineare Systeme von Differenzialgleichungen

Chen u. a. (1999) unternahmen bereits 1999 den Versuch, Systeme von Differenzialgleichungen zur Modellierung genregulatorischer Netzwerke heranzuziehen. Dazu haben sie homogene lineare Systeme von Differenzialgleichungen mit konstanten Koeffizienten verwendet. In diesem Modell beinhaltet der Vektor $x(t) \in \mathbf{R}^n$ sämtliche mRNA- und Proteinkonzentrationen des Systems zum Zeitpunkt t. Es wird angenommen, dass die Dynamik durch

$$\dot{x}(t) \quad = \quad Mx(t) \tag{1.17}$$

beschreibbar ist, wobei M eine konstante, reellwertige $(n \times n)$-Matrix ist, die den Einfluss der Variablen aufeinander beschreibt und $x : \mathbf{R} \to \mathbf{R}^n$ eine vektorwertige Funktion ist, die von $t \in \mathbf{R}$ abhängt. d'Haeseleer u. a. (1999) publizierten einen entsprechenden diskreten Ansatz, indem sie die Dynamik des genregulatorischen Netzwerkes durch

$$x(t + h) \quad = \quad Wx(t) \tag{1.18}$$

mit einer konstanten, reellwertigen $(n \times n)$-Matrix W und $h > 0$ beschrieben, wobei $x : \mathcal{N}(t_0, h) \to \mathbf{R}^n$ gilt. Hierbei sei $\mathcal{N}(t_0, h)$ für ein reellwertiges t_0 und $h \neq 0$ als $\mathcal{N}(t_0, h) = \{t_0 + kh : k =$

$0,1,...\}$ definiert (siehe Faigle (1983)). Man kann Systeme von Differenzialgleichungen als stetiges Analogon des entsprechenden Systems von Differenzengleichungen auffassen (Faigle, 1983). Führt man einen linearen Operator Δ_h ein, der durch

$$\Delta_h x(t) \quad = \quad \frac{1}{h}(x(t+h)-x(t)) = Ax(t) \tag{1.19}$$

mit $A \in \mathbf{R}^{n\times n}$ definiert ist, so gilt $W = hA + E_n$. Für $h \to 0$ erhält man den Differenzialoperator $D = \lim_{h\to 0} \Delta_h$, welcher auf die vektorwertige Funktion $x : \mathbf{R} \to \mathbf{R}^n$ angewendet werden kann. Somit erhält man ein stetiges System von Differenzialgleichungen über $Dx(t) = x'(t) = Ax(t)$ für die differenzierbare Funktion $x : \mathbf{R} \to \mathbf{R}^n$.

Hoon u. a. (2002) wählten das Modell (1.17) als Grundlage für ihr Modell, in dem sie sich aufgrund der Datenverfügbarkeit auf mRNA-Messungen beschränkten, das heißt, der Vektor $x(t) \in \mathbf{R}^n$ enthält jeweils die mRNA-Konzentrationen von n Genen. Um die Matrix dünn besetzt zu halten, setzen sie jedoch im Gegensatz zu Chen u. a. (1999) die maximale Anzahl an beeinflussenden Genen pro Gen nicht im Vorfeld bereits fest, sondern verwenden Akaikes Informationskriterium[9], um diesen Parameter, der den Zusammenhang des genregulatorischen Netzwerkes bestimmt, zu wählen.
Beide Modelle widersprechen jedoch der biologischen Realität in der Hinsicht, dass in genregulatorischen Netzwerken wenige einzelne Gene sehr viele andere Gene beeinflussen, jedoch die meisten Gene nur mit wenigen anderen verbunden sind. Diese wenigen einzelnen Gene mit vielen Einflüssen prägen ein genregulatorisches Netzwerk, so dass sie nicht vernachlässigt werden sollten.

Beide Ansätze haben außerdem den Nachteil, dass ein konstanter Einfluss, beispielsweise eine konstante Syntheserate, nicht modelliert werden kann, da nur Abhängigkeiten von den n Variablen möglich sind. Dieses Modell gibt auch nur einen Teil der Verhaltensweisen wieder, die genregulatorische Netzwerke zeigen können. Multistationaritäten, also die Existenz mehrerer verschiedener Gleichgewichtspunkte, wie sie in Kapitel 2 beschrieben werden, können durch solche Modelle beispielsweise nicht abgebildet werden.

Der Vorteil dieser Modelle, die auf linearen Differenzialgleichungen basieren, liegt jedoch auf der im Vergleich zu den meisten nichtlinearen Differenzialgleichungen geringen Anzahl an zu schätzenden Parametern. Daher sind auch mit linearen Modellen bereits gute Ergebnisse erzielt worden, wie beispielsweise bei der Vorhersage von Transkriptionsfaktoren in *Bacillus subtilis* (Hoon u. a., 2003). Hier wurde eine (5×5)-Matrix mit nur 8 Datenpunkten geschätzt. Biologisch sinnvolle Ergebnisse können also durchaus schon mit solch einem einfachen Ansatz erzielt werden. Dies liegt vermutlich daran, dass die einzelnen Regulierungsmechanismen selbst monoton sind, also bei einer Erhöhung der Transkriptionsfaktorkonzentration eine gleichstarke oder stärkere Regulierung zu beobachten ist, wenn alle weiteren Einflüsse vernachlässigt werden. Allerdings werden wir in Kapitel 2 sehen, dass eine Sättigung der Regulierungsstärke erfolgt, dass also ab einer gewissen Konzentration des Transkriptionsfaktors nur noch geringfügige Änderungen der Regulierungsstärke zu beobachten sind. Daher ist solch ein lineares Modell durchaus

[9] $AIC = -2L + 2K$ mit L logarithmiertem Maximum-Likelihood des geschätzten Modells und K Anzahl der geschätzten Parameter

dazu geeignet, beeinflussende Faktoren hervorzusagen, jedoch nicht, die Dynamik vorherzusagen.

Als Beispiel betrachte man den Fall eines einzigen Gens, das sich selbst reguliert. Die Gleichung lautet nach dem linearen Modell für die mRNA-Konzentration x_1:

$$\dot{x}_1(t) = m_{11}x_1(t) \tag{1.20}$$

mit $m_{11} \in \mathbf{R}$. Für $x_1(t_0) = 0$ verändert sich der Zustand trivialerweise nicht mehr. Wir betrachten nun $x_1(t_0) > 0$. Ist $m_{11} < 0$, so strebt $x_1(t)$ für $t \to \infty$ gegen 0. Ist $m_{11} > 0$, so strebt $x_1(t)$ für $t \to \infty$ gegen unendlich. Für $m_{11} = 0$ verändert sich der Anfangszustand nicht. Schätzt man den Parameter m_{11} aus Messwerten zu Zeitpunkten zwischen t_0 und t_l, so wird das Modell für eine negative Selbstregulierung gegen Null laufen, diesen Wert selbst jedoch nicht erreichen. Um hier einen Fixpunkt größer Null zu erreichen, müsste das Modell inhomogene Systeme von Differenzialgleichungen verwenden, also affine Gleichungen $m_{11}x_1(t) + b_1$ anstelle von $m_{11}x_1(t)$. Hat man eine positive Regulierung, kann das System (1.20) in vielen Fällen das Verhalten von Zeitpunkt t_0 bis Zeitpunkt t_l abbilden. Je größer jedoch der Abstand zwischen t und t_l wird, desto größer wird auch die Vorhersage $x_1(t)$ für den Zustand des Systems zum Zeitpunkt t. Quantitative Vorhersagen sind mit diesem Modell also nur lokal, nicht jedoch für Zeitpunkte $t > t_l$ möglich.

Man erkennt an diesem kleinen Beispiel, dass ein Modell basierend auf (1.17), in der eine Matrix M aus einer Zeitreihe mit Messungen von den Zeitpunkten t_0 bis t_l geschätzt wird, nur eingeschränkte Ergebnisse liefern kann. Es ist jedoch möglich, Transkriptionsfaktoren vorherzusagen und auch die Art des Einflusses - Aktivierung oder Inhibierung - zu bestimmen. Es ist jedoch nicht möglich, das Verhalten des Systems über die Messungen hinaus vorherzusagen, also für $t > t_l$.

Da ein lineares Modell in vielen Fällen bereits die Hauptregulierungen erfassen kann, scheint es sinnvoll, das Modell derart zu erweitern, dass das Modell auch über den Zeitraum von t_0 bis t_l hinweg zufriedenstellende Ergebnisse liefern kann. Dazu ist es nötig, nichtlineare Ansätze zu verfolgen.

Verschiedene nichtlineare Ansätze werden im folgenden Unterkapitel vorgestellt. Ein nichtlinearer Ansatz, der auf den zugrunde liegenden Prozessen basiert und außerdem in den verschiedenen Bereichen des Zustandsraums weiterhin lineare Differenzialgleichungen verwendet, die allerdings inhomogen sind, wird im nächsten Kapitel 2 beschrieben.

Nichtlineare Differenzialgleichungen

Allgemeine nichtlineare Regulierungsfunktionen Sakamoto u. Iba (2001) haben das Modell (1.17) verallgemeinert, indem die mRNA-Konzentration von Gen i beschrieben wird durch

$$\dot{x}_i = f_i(x_1, ..., x_n), \tag{1.21}$$

wobei $f_i : \mathbf{R}^n \to \mathbf{R}$ eine nichtlineare Funktion ist. Diese wird in Sakamoto u. Iba (2001) als frei wählbar erklärt und in ihrem Beispiel als Polynom vom Grad 2 angenommen. Zur Schätzung

der Koeffizienten wird ein genetischer Algorithmus verwendet. Im Beispiel wird das genregulatorische Netzwerk

$$\begin{aligned} \dot{x}_1(t) &= -0.9x_2(t) + x_3(t), \\ \dot{x}_2(t) &= 0.2x_1(t)x_2(t), \\ \dot{x}_3(t) &= x_3 - 0.5x_1(t) \end{aligned} \tag{1.22}$$

zugrunde gelegt und 100 Messwerte hieraus generiert. Ausgehend von einer Population von Individuen, die jeweils aus drei Gleichungen der Form $ax_1^2 + bx_2^2 + cx_1x_2 + dx_1 + ex_2 + f$ bestehen, wird die Fitness jedes Individuums aus einer Summe der Fehlerquadrate und einem Strafterm für den Grad der Gleichung bestimmt. Individuen mit der besten Fitness werden in der Population beibehalten und neue Individuen der Population hinzugefügt, die über Mutationen und Rekombinationen gewonnen wurden. Die geschätzten Parameter des Systems von Differenzialgleichungen sind im Beispiel

$$\begin{aligned} \dot{x}_1(t) &= -0.9020x_2(t) + 1.000x_3(t), \\ \dot{x}_2(t) &= 0.0369x_2(t)^2 + 0.1920x_1(t)x_2(t), \\ \dot{x}_3(t) &= 1.0208x_3(t) - 0.5200x_1(t), \end{aligned} \tag{1.23}$$

wobei hier Werte unter 0.1 auf Null gesetzt wurden. Für den Grad 2 und 100 Messwerte ist es also möglich, mit einem genetischen Algorithmus zufriedenstellende Ergebnisse für ein kleines Netzwerk mit drei Genen zu erhalten. Lokal können also wiederum gute Ergebnisse erreicht werden. Möchte man jedoch nicht nur einen kurzen Zeitabschnitt modellieren, ist es fraglich, ob eine Funktion der Form $\dot{x}_i(t) = -ax_k^2(t) + b$ mit $a, b > 0$ zur Beschreibung von mRNA-Konzentrationen biologisch sinnvoll ist. Dies würde bedeuten, dass bei geringer Konzentration des Transkriptionsfaktors $\dot{x}_i$ zunächst positiv ist, jedoch für eine Konzentration größer als $\sqrt{b/a}$ die zeitliche Änderung von x_i negativ wird. Der Transkriptionsfaktor x_k würde also sowohl x_i aktivieren, als auch inhibieren, und zwar allein bedingt durch die Konzentration von x_k. Auch wird solch eine negative Änderung in der Methode nicht durch beispielsweise eine positive Syntheserate aufgefangen, so dass für $t \to \infty$ der Zustand $x_i(t) < 0$ möglich ist, was dem biologischen System jedoch widerspricht. Eine lokale Beschreibung des Systems innerhalb der Zeit, in denen die Messdaten gewonnen wurden, ist also auch hier zufriedenstellend möglich, jedoch produziert das berechnete System von Differenzialgleichungen wiederum Lösungen, die außerhalb dieses Zeitraumes biologisch nicht sinnvolle Ergebnisse liefern. Auch die Regulierungsfunktionen können keinen biologischen Reaktionen oder Prozessen zugeordnet werden. Sowohl eine lokal lineare Approximation als auch eine lokal quadratische Approximation kann also für die Erkennung der Hauptregulatoren verwendet werden, liefert jedoch für ein Modell, das Prognosen ermöglichen soll, keine geeigneten Regulierungsfunktionen.

Eine weitere Schwierigkeit, die bei nichtlinearen Differenzialgleichungen zu bewältigen ist, wird in dem Beispiel von Sakamoto u. Iba (2001) ebenfalls deutlich. Lässt man sämtliche Formen von Regulierungsfunktionen zu, wächst die Anzahl der Parameter so stark, dass die Parameter durch Zeitreihenmessungen, die üblicherweise bei unter 50 Zeitpunkten liegen, nicht mehr schätzbar sind, weil die Anzahl der Parameter bei bereits wenigen Genen schon unter der

Anzahl der Messungen liegt. Um jedoch die Parameter eines Polynoms von Grad n zu schätzen, müssen mindestens $n+1$ Messwerte vorhanden sein.
Es muss also biologisches Wissen über das genregulatorische Netzwerk mit einfliessen. Dies ist darüber möglich, dass parametrisierte Regulierungsfunktionen wie im folgenden Ansatz vorgegeben werden.

Prozessbasierter Ansatz Ein Modellansatz, der die Form der Differenzialgleichungen aufgrund der zugrunde liegenden biochemischen Prozesse wählt und dennoch einfache Methoden zur Analyse der Differenzialgleichungen bereitstellt, ist von de Jong u. a. (2004) entwickelt und vertieft worden und beruht auf grundlegenden Arbeiten von Glass u. Kauffman (1973). Das Modell wurde aufgrund der Beobachtung aufgestellt, dass die Aktivität eines Gens als Funktion der Transkriptionsfaktorkonzentration einer sehr steilen sigmoidalen Funktion folgt. Nach diesem Ansatz wird die Dynamik zellulärer Proteinkonzentrationen $x = (x_1, ..., x_n)^t$ für die Proteine $i = 1, ..., n$ durch

$$\dot{x}_i(t) = f_i(x(t)) - \gamma_i x_i(t) \tag{1.24}$$

beschrieben, wobei $f_i : \mathbf{R}_+^n \to \mathbf{R}_+$ die Syntheserate und $\gamma_i x_i(t)$ die Abbaurate des Proteins i beschreibt. Die Funktion f_i wird parametrisiert durch

$$f_i(x) = \sum_{l \in L} \kappa_{il} b_{il}(x), \tag{1.25}$$

mit $\kappa_{il} > 0$ als Regulierungsstärken, $b_{il} : \mathbf{R}_+^n \to \{0,1\}$ als Regulierungsfunktionen, die die Bedingungen angeben, unter denen die Regulierungsstärke wirksam ist, und $L \subseteq \{1, ..., n\}$ als Indexmenge, die die Indizes der regulierenden Proteine beinhaltet. Analog kann die Funktion $\gamma_i(x)$ definiert werden, wenn auch der Abbau von anderen Proteinkonzentrationen als von Protein i selbst abhängig gemacht werden soll. Die Regulierungsfunktionen b_{il} werden wiederum als Multiplikation verschiedener Stufenfunktionen $s^+, s^- : K \times \mathbf{R}_+ \to \{0,1\}$, $K \subseteq \mathbf{R}_+$,

$$s^+(x_j, \theta_j^i) = \begin{cases} 1 & \text{für } x_j > \theta_j^i \\ 0 & \text{für } x_j < \theta_j^i \end{cases} \tag{1.26}$$

und

$$s^-(x_j, \theta_j^i) = 1 - s^+(x_j, \theta_j^i) \tag{1.27}$$

definiert. Laut dieser Definition sind die Stufenfunktionen s^+ und s^- für die Schwellwerte θ_j^i nicht definiert, was dazu führt, dass auch die Regulierungsfunktionen b_{il} Definitionslücken aufweisen. Dies verursacht wiederum Komplikationen bei der Analyse des dynamischen Verhaltens. Andererseits ist das Verhalten des Systems abseits der Definitionslücken sehr gut zu analysieren, da die Funktion auf der rechten Seite von Gleichung (1.24) stückweise linear ist. Wird der Zustandsraum nun aufgrund der Schwellwerte θ_j^i der Regulierungsfunktionen in Hyperrechtecke zerteilt, indem man Hyperebenen $x_i = \theta_j^i$ definiert, so unterscheiden de Jong u. a. (2004) in ihrem Modell zwischen regulatorischen und umschaltenden Bereichen. Regulatorische Bereiche sind die durch die Hyperebenen beschränkten n-dimensionalen Teilmengen des $\mathbf{R}^n$, für umschaltende

Bereiche gilt, dass mindestens ein Paar $i, j \in \{1,...,n\}$ existiert, für das $x_i = \theta_j^i$ gilt. Umschaltende Bereiche sind also beispielsweise die Hyperebenen selbst. In jedem regulatorischen Bereich ist ein System von Differenzialgleichungen der Form

$$\dot{x}_i \quad = \quad \kappa - \gamma x_i \tag{1.28}$$

gegeben, wobei κ eine Konstante ist, die sämtliche regulatorischen Einflüsse beinhaltet. Dieses System strebt gegen den Fixpunkt $\bar{x}_i = \kappa/\gamma$. Liegt die Anfangsbedingung $x_i(\tau) = \eta_i$ ebenso wie der Fixpunkt in demselben regulatorischen Bereich D, so gilt $x_i(t) \in D$ und $x_i(t) \to \bar{x}_i$ für $t \to \infty$. Liegt der Fixpunkt nicht in D, so kann das System entweder augenblicklich den umschaltenden Bereich D' durchqueren oder in D' verbleiben, siehe Casey u. a. (2006); de Jong u. a. (2004). Um bestimmen zu können, ob D' augenblicklich durchquert wird oder ein Gleichgewichtspunkt in D' existiert, müssen Lösungen der Differenzialgleichungen für die Definitionslücken der Regulierungsfunktionen definiert werden. Hierzu greifen de Jong u. a. (2004) auf Filippov-Lösungen zurück. Zunächst wird für regulatorische Bereiche D und umschaltende Bereiche D'

$$\begin{aligned} H(x) \quad &= \quad \{\kappa_D - \gamma x\} \text{ für } x \in D \\ \text{und} & \\ H(x) \quad &= \quad co(\{\kappa_D - \gamma x | D \in R(D')\}) \text{ für } x \in D' \end{aligned} \tag{1.29}$$

definiert, wobei $co(X)$ die konvexe Hülle von X ist, κ_D die Konstante, die den regulatorischen Einfluss in D beschreibt und $R(D')$ sämtliche regulatorische Bereiche beinhaltet, die an den umschaltenden Bereich D' angrenzen.

Definition 1.11. *(Bronstein u. Semendjajew, 1991) Eine Funktion $f : D \to \mathbf{R}$, $D \subseteq \mathbf{R}$ heißt absolut stetig in $[a,b] \subseteq D$, falls für alle $\varepsilon > 0$ ein $\delta > 0$ existiert, so dass für jede Folge paarweise disjunkter, in $[a,b]$ enthaltener Intervalle $I_k = [x_{k-1}, x_k]$, $k = 1,...,n$, mit $\sum_{k=1}^{n}(x_k - x_{k-1}) < \delta$ auch $|\sum_{k=1}^{n} f(x_k) - f(x_{k-1})| < \varepsilon$ gilt.*

Lösungen des Systems von Differenzialgleichungen werden dann definiert durch

Definition 1.12. *(Casey u. a., 2006)*
Eine Lösung von $\dot{x} \in H(x)$ auf $[0,T]$ im Sinne von Filippov ist eine (bzgl. t) absolut stetige Funktion $\xi_t(x_0)$, so dass $\xi_t(x_0) = x_0$ und $\dot{\xi}_t \in H(\xi_t)$ für fast alle $t \in [0,T]$ gilt (also für alle bis auf endlich viele $t \in [0,T]$).

Ein Gleichgewichtspunkt $\bar{x}$ in einem umschaltenden Bereich D' wird dann dadurch definiert, dass $0 \in H(\bar{x})$ liegt.

Nach der Berechnung der Übergänge des Systems zwischen den Bereichen wird ein qualitativer Zustandsübergangsgraph entworfen, in der die Dynamik des Systems nachvollzogen werden kann. Jeder Knoten des Zustandsübergangsgraphen entspricht einem regulatorischen oder umschaltenden Bereich. Die Übergänge von einem Bereich zum nächsten werden durch gerichtete

Kanten angezeigt. Für das Beispiel

$$\begin{aligned} \dot{x}_1(t) &= \kappa_1 s^-(x_2, \theta_2), \\ \dot{x}_2(t) &= \kappa_2 s^-(x_1, \theta_1) \end{aligned} \tag{1.30}$$

erhält man den in Abbildung 1.5 gezeigten Zustandsübergangsgraphen und den entsprechend in regulatorische und umschaltende Bereiche zerlegten Zustandsraum.

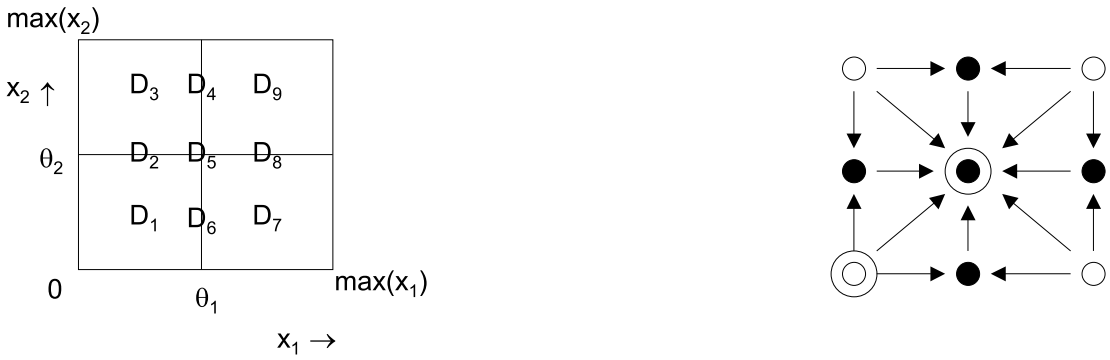


Abbildung 1.5: Dargestellt sind der durch Hyperebenen unterteilte Zustandsraum (links) und der Zustandsübergangsgraph (rechts). Im Zustandsübergangsgraphen werden regulatorische Bereiche durch leere Kreise und umschaltende Bereiche durch ausgefüllte Kreise dargestellt. Bereiche, in denen sich Gleichgewichtspunkte befinden, werden durch einen zusätzlichen Kreis markiert. Positionen im Zustandsübergangsgraph entsprechen den Positionen im Zustandsraum.

Die Stärke des Modells liegt darin, Regulierungsfunktionen zuzulassen, die über rein additive Modelle hinausgehen, da die Stufenfunktionen miteinander multipliziert werden und hierdurch biochemische Prozesse abgebildet werden können, in denen die Konzentration von Proteinkomplexen eine Rolle spielt. Andererseits muss das Modell durch die Stufenfunktionen als rein qualitatives Modell angesehen werden und verhindert durch die unstetig definierten Regulierungsfunktionen mit Definitionslücken eine einfache Analyse des Systems. Regulierungsfunktionen, die an den Schwellwerten stetig sind, würden diesen Nachteil beheben und könnten neben einer einfacheren Analyse auch quantitative Aussagen zulassen. Die Herleitung und Analyse eines solchen Modells wird im folgenden Kapitel erklärt.

Kapitel 2

Ein neues Modell für Genregulierungen, das auf stückweise linearen Differenzialgleichungen basiert

Das durch Gleichung (1.24) beschriebene Modell verwendet sehr einfache Differenzialgleichungen, da stückweise konstante Regulierungsfunktionen verwendet werden, die je nach Lage des Systems im Zustandsraum von einem konstanten Wert zu einem anderen konstanten Wert springen. Die Regulierungsfunktion wird als Stufenfunktion angenommen. Für geringe Konzentrationen eines Regulators ändert sich die Expression des regulierten Gens nicht, für hohe Konzentrationen wird ein maximaler Regulierungseinfluss angenommen. Die Einteilung in geringe und hohe Konzentrationen des Regulators geschieht nur aufgrund eines einzigen Schwellwertes.
Aufgrund der sehr vereinfachten Annahme von Stufenfunktionen für Regulierungsvorgänge zählt das Modell zu den qualitativen Modellen. In dieser Arbeit soll nun ein Modell vorgestellt werden, das auch quantitative Aussagen ermöglicht und aus diesem Grund auf einer verfeinerten Regulierungsfunktion basiert, das jedoch weiterhin analytisch lösbar bleibt. Das Modell wurde in Zusammenarbeit mit Nicole Radde entwickelt, die sich in ihrer Arbeit auf sigmoidale Funktionen stützt, wohingegen in der vorliegenden Arbeit stückweise lineare Funktionen verwendet werden. Beide Modelle basieren auf der Theorie der Genregulierung von Jacob u. Monod (1961). Im folgenden Unterkapitel wird zunächst der quantitative Zusammenhang zwischen Regulator und reguliertem Protein erläutert, der zu einer Regulierungsfunktion führt, die durch eine stetige, stückweise lineare Funktion approximiert wird. Die Ergebnisse liefern am Ende des ersten Abschnitts ein Modell für genregulatorische Netzwerke, das auf stückweise linearen Differenzialgleichungen beruht. Im zweiten Unterkapitel wird das entstandene Modell diskutiert. Es werden Analysemöglichkeiten aufgezeigt und anschliessend wird das Modell im letzten Unterkapitel durch ein Beispiel veranschaulicht.
Zunächst muss bei Modellen für genregulatorische Netzwerke geklärt werden, was die Komponenten im Netzwerk darstellen. In manchen Fällen sind dies die Gene selbst, in anderen die mRNA oder das Protein und in wieder anderen wird unterschieden zwischen mRNA und Protein. Hierbei ist es wichtig, zwischen Prokaryonten und Eukaryonten zu unterscheiden. Da diese Arbeit auf Prokaryonten ausgerichtet ist, können posttranskriptionale Regulierungen als vernachlässigbar angenommen und Regulierungen auf Transkriptionsebene als hauptsächlicher Regulierungsmechanismus angenommen werden. Ein wichtiger Grund hierfür ist, dass starke posttranskriptionale Regulierungsmechanismen, wie z. B. das alternative Spleißen, in Proka-

ryonten nicht existieren. Die Proteinkonzentration wird als proportional zur zugehörigen mRNA-Konzentration angenommen. Somit werden im Folgenden direkt die Einflüsse zwischen verschiedenen mRNA-Konzentrationen modelliert, ohne die Translation miteinzubeziehen. Mit Expression ist demnach in diesem Kapitel vorrangig der Schritt der Transkription gemeint, die Komponenten im Netzwerk beziehen sich auf mRNAs.

2.1 Herleitung der stückweise linearen Regulierungsfunktionen

Im Folgenden werden indirekte und direkte Regulierungen betrachtet. Bei einer direkten Regulierung bindet der Transkriptionsfaktor selbst an die DNA und löst dadurch eine Aktivierung oder Inhibierung aus. Bei einer indirekten Regulierung existiert ein Repressor, der im Fall einer Aktivierung durch Transkriptionsfaktoren von der DNA gelöst wird oder im Fall einer Inhibierung erst mithilfe der Transkriptionsfaktoren an die DNA binden kann, siehe auch Kapitel 1.

2.1.1 Indirekte Regulierungen

Die Expression eines Gens kann nicht nur durch das Produkt eines anderen Gens reguliert werden, sondern es lassen sich auch quantitative Zusammenhänge zwischen der Expression des Gens und der mRNA-Konzentration des anderen Gens herleiten. Im Folgenden werde ich mich an Veröffentlichungen von Yagil u. Yagil (1971); Yagil (1975) orientieren, die folgendes Ergebnis für diesen quantitativen Zusammenhang präsentierten:

$$\log\left(\frac{\alpha}{1-\alpha}-\frac{[K_2]}{[R_t]}\right)=\pm n\log[E]+\log\frac{K_2}{[R_t]}\mp\log K_1, \tag{2.1}$$

wobei α für den Anteil freier Operatoren steht, [E] ist die Konzentration des Enzyms (hier Transkriptionsfaktor), n die Anzahl der Enzyme, die mit dem Repressor eine Verbindung eingehen, K_1 die Dissoziationskonstante dieser Dissoziationsreaktion und $\frac{K_2}{[R_t]}$ das Verhältnis von Repressor-Operator-Dissoziationskonstante zur gesamten Repressorkonzentration. Das obere Zeichen gilt für eine Aktivierung, das untere für eine Inhibierung.
Im Folgenden soll analog zu Yagil und Yagil eine Funktion $f([T])$ hergeleitet werden, die die Expressionsrate des regulierten Gens in Abhängigkeit von der Konzentration des Transkriptionsfaktors T beschreibt. Hierbei werden einige mit den biologischen Gegebenheiten im Einklang stehende Annahmen gemacht. Zunächst gehen wir davon aus, dass in jeder Zelle mehr Repressormoleküle existieren als Operatoren. Des Weiteren seien in der chemischen Reaktion $RE_n \rightleftharpoons R+nE$ Zwischenprodukte wie $RE_{\tilde{n}}$ mit $1 \leq \tilde{n} < n$ vernachlässigbar.
Eine Dissoziationskonstante ist die Gleichgewichtskonstante einer Dissoziationsreaktion. Die Gleichgewichtskonstante beschreibt das Mengenverhältnis der Produkte zu den Edukten, für das ein chemisches Gleichgewicht herrscht, das heißt, dass die Geschwindigkeiten der Hin- und Rückreaktion gleich groß sind und sich die Konzentrationen von Produkten und Edukten somit nicht mehr ändern (Darnell u. a., 1994). Das Verhältnis ist unter den üblichen Standardbedingungen mit einer Temperatur von 25°C und einem Druck von 1 bar für jede chemische Reaktion

eine unveränderliche Größe und wird mit K_{eq} bezeichnet. Hat man eine allgemeine reversible chemische Reaktion

$$a_1A_1 + a_2A_2 + ... + a_nA_n \quad \rightleftharpoons \quad b_1B_1 + b_2B_2 + ... + b_mB_m \tag{2.2}$$

mit Edukten $A_1,...,A_n$ und Produkten $B_1,...,B_m$ sowie $a_i, b_j \in \mathbb{N}$ für $i = 1,...,n,\ j = 1,...,m$, so ist die Gleichgewichtskonstante K_{eq} gegeben durch

$$K_{eq} \quad = \quad \frac{[B_1]^{b_1} \cdot [B_2]^{b_2} \cdot ... \cdot [B_m]^{b_m}}{[A_1]^{a_1} \cdot [A_2]^{a_2} \cdot ... \cdot [A_n]^{a_n}}. \tag{2.3}$$

Mit der Gleichgewichtskonstanten kann man also auch die Richtung bestimmen, in der die Reaktion abläuft. Ist das Mengenverhältnis der Produkte zu den Edukten kleiner als die Gleichgewichtskonstante, so findet die Hinrichtung schneller als die Rückrichtung statt, bis das Gleichgewicht wieder hergestellt ist. Ist das Verhältnis größer als die Gleichgewichtskonstante, so findet entsprechend die Rückrichtung schneller statt als die Hinrichtung, bis ebenfalls wieder das Gleichgewicht erreicht ist. Erhöht man also beispielsweise die Konzentration von A_1, so erhält man mehr Produkte als vorher. Dasselbe Prinzip lässt sich auch in der Regulierung von Genexpressionen erkennen.
Als Dissoziationsreaktion bezeichnet man chemische Reaktionen, in denen ein Molekül in seine Bestandteile zerfällt, also z. B.

$$AB \quad \rightleftharpoons \quad A + B. \tag{2.4}$$

Analog zum Zerfall eines Moleküls lassen sich Aktivierungen und Inhibierungen bei Genregulierungen beschreiben. Wir betrachten zunächst den Fall, dass eine indirekte Aktivierung vorliegt, also eine vorliegende Inhibierung aufgehoben wird. Dies bedeutet, dass ein Repressor an einen Operator binden kann und eine Aktivierung der Expression der zum Operator gehörenden Gene dadurch stattfindet, dass Enzyme an den Repressor binden, so dass die Inhibierung durch den Repressor verhindert wird. Im Folgenden interessieren wir uns für den Anteil der freien bzw. gebundenen Operatoren in der Gesamtpopulation. Mit α wird der Anteil der freien Operatoren in der Gesamtpopulation bezeichnet, entsprechend erfasst man mit $1-\alpha$ den Anteil der gebundenen Operatoren. Analog sei β der Anteil der freien, $1-\beta$ der Anteil der gebundenen Aktivatorbindungsstellen. Unser Interesse an diesen Verhältnissen leitet sich daher ab, dass wir annehmen, die Syntheserate der mRNA sei proportional zu α bzw. $1-\beta$: Bezeichnet also c_i die Syntheserate der mRNA eines Gens i, so nehmen wir $c_i = k\alpha$ mit $k > 0$ an, wenn indirekte Aktivierungen, indirekte Inhibierungen oder direkte Inhibierungen beteiligt sind. Im Fall, dass die Aktivierung direkt durch die Bindung des Transkriptionsfaktors an die entsprechende Aktivatorbindungsstelle hervorgerufen wird, gilt die Annahme $c_i = l(1-\beta)$ mit $l > 0$, dass also die Syntheserate proportional zum Anteil der gebundenen Aktivatorbindungsstellen ist. Für alle vier Fälle der direkten bzw. indirekten Aktivierung und Inhibierung wird nun α bzw. im Fall der direkten Aktivierung $1-\beta$ berechnet.

Satz 2.1. *(indirekte Aktivierung, (Yagil u. Yagil, 1971))*
Bei einer indirekten Aktivierung, in der der Repressor R durch n Moleküle des Enzyms E von

einem Operator O gelöst wird, ist das Verhältnis der Konzentration von freien Operatoren $[O]$ zu der Gesamtkonzentration der Operatoren $[O_t]$ abhängig von der Enzymkonzentration $[E]$, der Gesamtkonzentration von Repressoren $[R_t]$ und den Gleichgewichtskonstanten K_1 und K_2 der Dissoziationsreaktion, in der n Enzyme mit dem Repressor binden bzw. der Operator mit dem Repressor bindet. Das Verhältnis berechnet sich zu

$$\frac{[O]}{[O_t]} = \frac{\frac{\alpha_b}{K_1}[E]^n + \alpha_b}{\frac{\alpha_b}{K_1}[E]^n + \alpha_b + 1} =: f_1([E]) \tag{2.5}$$

mit $\alpha_b = \frac{K_2}{[R_t]}$, $[R_t], K_1, K_2 \in \mathbf{R}^+$, $n \in \mathbf{N}^+$ *und* $f_1 : \mathbf{R}_0^+ \to \mathbf{R}$.

BEWEIS:
In dem Fall einer indirekten Aktivierung finden die zwei Dissoziationsreaktionen

$$RE_n \rightleftharpoons nE + R \tag{2.6}$$

und

$$OR \rightleftharpoons O + R \tag{2.7}$$

statt, in denen Gleichung (2.6) die Bindung von n Enzymen E an den ungebundenen Repressor R und Gleichung (2.7) die Bindung des ungebundenen Repressors R an den Operator O beschreibt. Die Rolle des von uns betrachteten Transkriptionsfaktors übernimmt hier das Enzym E. RE_n beschreibt den Repressor, an den n Enzyme gebunden haben, und OR den Operator, an dem ein Repressor gebunden hat. Die Gesamtkonzentration von Operatoren in unserer Population - bezeichnet mit $[O_t]$ - setzt sich zusammen aus den ungebundenen Operatoren und den gebundenen Operatoren, also

$$[O_t] = [O] + [OR]. \tag{2.8}$$

Die Gesamtkonzentration des Repressors in unserer Population - bezeichnet mit $[R_t]$ - besteht ebenfalls aus gebundenen und ungebundenen Repressoren, also

$$[R_t] = [R] + [RE_n]. \tag{2.9}$$

Die Gleichgewichtskonstante der Dissoziationsgleichungen (2.6) bzw. (2.7) lauten

$$K_1 = \frac{[R][E]^n}{[RE_n]} \text{ bzw.} \tag{2.10}$$

$$K_2 = \frac{[O][R]}{[OR]}. \tag{2.11}$$

Setze Gleichung (2.9) in Gleichung (2.10) ein:

$$\begin{aligned} & K_1 = \frac{[R][E]^n}{[RE_n]} \\ & \Leftrightarrow K_1[RE_n] = [R][E]^n \\ & \Leftrightarrow_{(2.9)} K_1[R_t] - K_1[R] = [R][E]^n \\ & \Leftrightarrow K_1[R_t] = [R]K_1 + [R][E]^n \\ & \Leftrightarrow [R] = \frac{K_1[R_t]}{K_1 + [E]^n}. \end{aligned} \tag{2.12}$$

Nun erhält man die in Yagil u. Yagil (1971) theoretisch hergeleitete und experimentell bestätigte Gleichung, die zu Anfang dieses Unterkapitels zitiert wurde:

$$\begin{aligned}\frac{\alpha}{1-\alpha} &= \frac{[O]}{[O_t]}\frac{[O_t]}{[OR]} = \frac{[O]}{[OR]} =_{(2.11)} \frac{K_2}{[R]} \\ &=_{(2.12)} \frac{K_2}{K_1[R_t]}(K_1+[E]^n) = \frac{K_2}{K_1[R_t]}[E]^n + \frac{K_2}{[R_t]} \end{aligned} \tag{2.13}$$

bzw. logarithmiert:

$$\log\left(\frac{\alpha}{1-\alpha} - \frac{K_2}{[R_t]}\right) = n\log([E]) + \log\frac{K_2}{[R_t]} - \log K_1. \tag{2.14}$$

Durch Auflösen dieser Gleichung und Setzen von $\frac{K_2}{[R_t]} = \alpha_b$ erhält man

$$\begin{aligned}&\frac{\alpha}{1-\alpha} = \frac{\alpha_b}{K_1}[E]^n + \alpha_b \\ &\Leftrightarrow \alpha + \frac{\alpha_b}{K_1}[E]^n\alpha + \alpha_b\alpha = \frac{\alpha_b}{K_1}[E]^n + \alpha_b \\ &\Leftrightarrow \alpha = \frac{\frac{\alpha_b}{K_1}[E]^n + \alpha_b}{\frac{\alpha_b}{K_1}[E]^n + \alpha_b + 1} =: f_1([E]). \end{aligned} \tag{2.15}$$

□

In Abbildung 2.1 wird diese Funktion mit verschiedenen Werten von n, K_1 und α_b gezeigt. Je größer n ist, desto steiler steigt die Funktion an. Für K_1 gilt der umgekehrte Zusammenhang. Je größer K_1 ist, desto langsamer steigt die Funktion an. Für $[E] = 0$ erhält man $\alpha = \frac{\alpha_b}{\alpha_b+1}$, das heißt, dass im Fall von nicht vorhandenen Regulatorproteinen die relative Häufigkeit von ungebundenen Operatoren im Vergleich zur Gesamtoperatorenzahl als $\frac{\alpha_b}{\alpha_b+1}$ berechnet werden kann. Je größer α_b ist, desto höher ist daher die Grundexpression, die für $[E] = 0$ vorhanden ist.

Bemerkung 2.2. *Für die Funktion* $f_1 : \mathbf{R}_0^+ \to \mathbf{R}$ *gilt für* $\alpha_b > 0$ *das asymptotische Verhalten* $\lim_{[E]\to\infty} f_1([E]) = 1$ *und des Weiteren* $f_1([E]) = \frac{\alpha_b}{\alpha_b+1}$ *für* $[E] = 0$.

Als nächstes sollen indirekte Inhibierungen betrachtet werden. Bei einer indirekten Inhibierung binden zunächst n Moleküle des Enzyms an einen Repressor, bevor dieser an die DNA binden kann.

Satz 2.3. *(indirekte Inhibierung, (Yagil u. Yagil, 1971))*
Bei einer indirekten Inhibierung, in der der Repressor R erst mit n Enzymen E binden muss, bevor er an einen Operator O binden kann, ist das Verhältnis der Konzentration der freien Operatoren $[O]$ *zu der Gesamtkonzentration der Operatoren* $[O_t]$ *abhängig von der Enzymkonzentration* $[E]$*, der Gesamtkonzentration von Repressoren* $[R_t]$ *und den Gleichgewichtskonstanten*

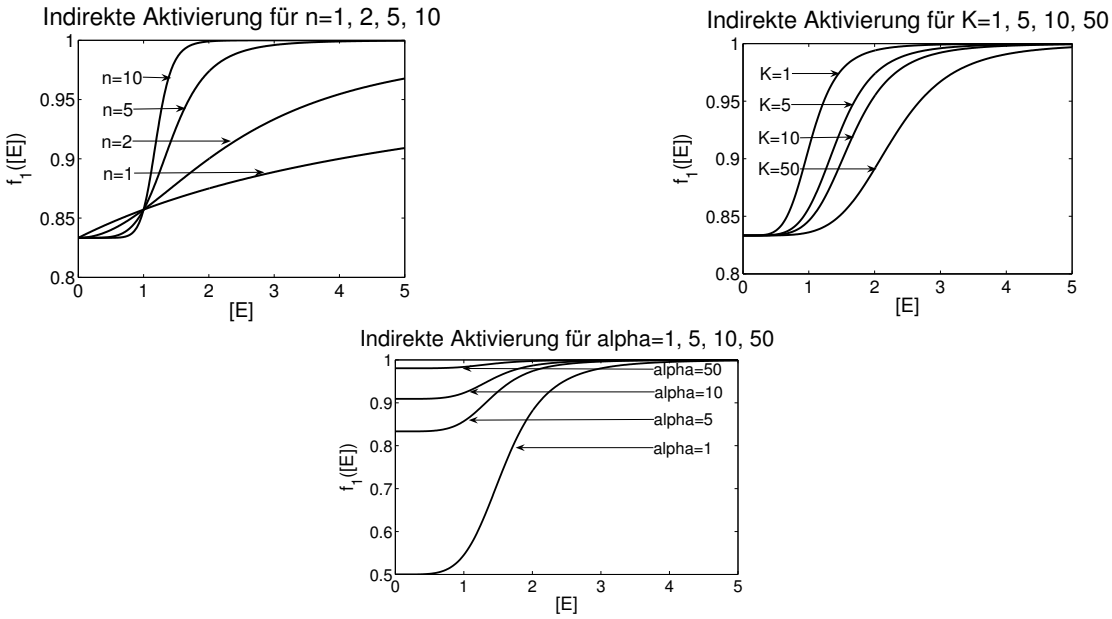


Abbildung 2.1: In dieser Abbildung wird die Funktion $f_1([E]) = \frac{\frac{\alpha_b}{K_1}[E]^n + \alpha_b}{\frac{\alpha_b}{K_1}[E]^n + \alpha_b + 1}$ gezeigt. Links sind verschiedene Werte für den Parameter n gewählt, in der Mitte verschiedene Werte für den Parameter K, rechts für den Parameter α_b. Die nicht variierten Parameter sind als $n = 5$, $K_1 = 5$ bzw. $\alpha_b = 5$ gewählt.

K_1 und K_2 der Dissoziationsreaktionen, in denen n Enzyme E mit dem Repressor binden bzw. der Operator mit dem Enzym-Repressor-Komplex bindet. Das Verhältnis berechnet sich zu

$$\frac{[O]}{[O_t]} = \frac{K_1\alpha_b + \alpha_b[E]^n}{K_1\alpha_b + (\alpha_b + 1)[E]^n} =: f_2([E]) \tag{2.16}$$

mit $\alpha_b = \frac{K_2}{[R_t]}$, $[R_t], K_1, K_2 \in \mathbf{R}^+$, $n \in \mathbb{N}^+$ und $f_2 : \mathbf{R}_0^+ \to \mathbf{R}$.

BEWEIS:
Gleichung (2.6) bleibt bestehen, es gilt also $RE_n \rightleftharpoons R + nE$, wohingegen Gleichung (2.7) nun durch

$$ORE_n \rightleftharpoons O + RE_n \tag{2.17}$$

ersetzt werden muss. Dadurch ändern sich K_1 und $[R_t]$ nicht, jedoch berechnet sich die Dissoziationskonstante K_2 nun zu

$$K_2 = \frac{[O][RE_n]}{[ORE_n]} \tag{2.18}$$

und $[O_t]$ muss ersetzt werden durch

$$[O_t] \quad = \quad [O] + [ORE_n]. \tag{2.19}$$

Analog zum vorherigen Fall der indirekten Aktivierung kann hier $[RE_n]$ in Abhängigkeit von $[R_t]$ beschrieben werden:

$$\begin{aligned} & K_1 = \frac{[R][E]^n}{[RE_n]} \\ & \Leftrightarrow K_1[RE_n] = [R][E]^n \\ & \Leftrightarrow_{(2.9)} K_1[RE_n] = [R_t][E]^n - [RE_n][E]^n \\ & \Leftrightarrow [RE_n](K_1 + [E]^n) = [R_t][E]^n \\ & \Leftrightarrow [RE_n] = \frac{[R_t][E]^n}{K_1 + [E]^n}. \end{aligned} \tag{2.20}$$

Der Wert α gibt in diesem Fall wieder das Verhältnis $[O]/[O_t]$ an, wohingegen $1-\alpha$ das Verhältnis $[ORE_n]/[O_t]$ beschreibt. Die Bestimmung von $\alpha/(1-\alpha)$ ergibt sich somit folgendermaßen:

$$\begin{aligned} \frac{\alpha}{1-\alpha} \quad = \quad & \frac{[O]}{[ORE_n]} =_{(2.18)} \frac{K_2}{[RE_n]} \\ & =_{(2.20)} \frac{K_2(K_1 + [E]^n)}{[R_t][E]^n} = \frac{K_2}{[R_t]} + \frac{K_1 K_2}{[R_t][E]^n}. \end{aligned} \tag{2.21}$$

Durch Logarithmieren erhält man das zu Anfang vorgestellte Ergebnis für indirekte Inhibierungen:

$$\log\left(\frac{\alpha}{1-\alpha} - \frac{K_2}{[R_t]}\right) \quad = \quad -n\log[E] + \log\frac{K_2}{[R_t]} + \log K_1. \tag{2.22}$$

Diese Gleichung wird nach α aufgelöst, um auf die Syntheserate schliessen zu können:

$$\alpha = \frac{K_1\alpha_b + \alpha_b}{K_1\alpha_b + (\alpha_b + 1)[E]^n} =: f_2([E]) \tag{2.23}$$

mit $\alpha_b = \frac{K_2}{[R_t]}$ wie auch im vorherigen Fall.

□

In Abbildung 2.2 wird die Funktion $f_2([E])$ in Abhängigkeit von verschiedenen Werten für n, K_1 und α_b gezeigt.

Bemerkung 2.4. *Für die Funktion* $f_2 : \mathbf{R}_0^+ \to \mathbf{R}$ *gilt für* $\alpha_b > 0$ *das asymptotische Verhalten* $\lim_{[E]\to\infty} f_2([E]) = \frac{\alpha_b}{\alpha_b+1}$ *und des Weiteren* $f_2([E]) = 1$ *für* $[E] = 0$.

Neben diesen bisher betrachteten enzymatischen Regulierungen können Transkriptionsfaktoren auch direkt binden. Dies soll im nächsten Unterkapitel untersucht werden.

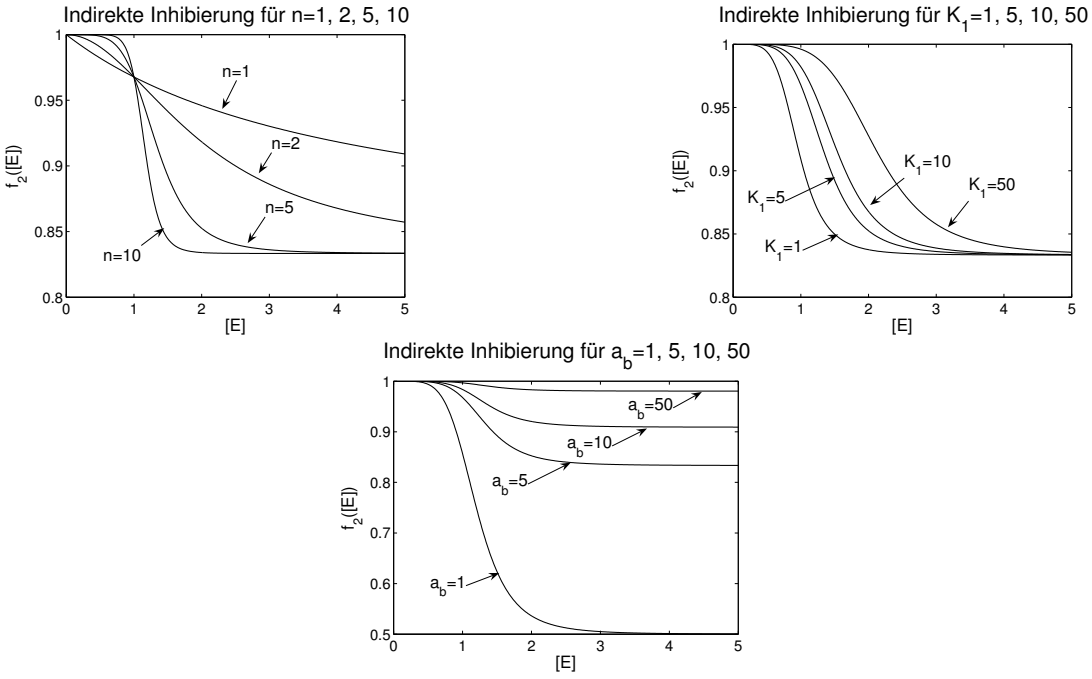


Abbildung 2.2: In dieser Abbildung wird die Funktion $f_2([E]) = \frac{K_1\alpha_b+\alpha_b[E]^n}{K_1\alpha_b+(\alpha_b+1)[E]^n}$ gezeigt. Jeweils einer der Parameter n, K_1 bzw. α_b wird variiert, die übrigen werden durch $n = 5$, $K_1 = 5$ bzw. $\alpha_b = 5$ festgesetzt.

2.1.2 Direkte Regulierungen

Bisher wurden indirekte Aktivierungen und indirekte Inhibierungen beschrieben. Die relativen Häufigkeiten von ungebundenen Operatoren bzw. gebundenen Aktivatorbindungsstellen sollen nun für direkte Regulierungen berechnet werden.
Im ersten Fall sei die Regulierung eine Aktivierung: Der Transkriptionsfaktor T bindet mit n seiner Moleküle an eine Aktivatorbindungsstelle A und erhöht dadurch die Transkriptionsrate. Im Gegensatz zu den vorher beschriebenen Fällen sind wir daher nun an der relativen Häufigkeit interessiert, mit der die Aktivatorbindungsstellen gebunden sind, um auf die Transkriptionsrate schliessen zu können.

Satz 2.5. *(direkte Aktivierung)*
Bei einer direkten Aktivierung durch den Transkriptionsfaktor T ist das Verhältnis der Konzentration von gebundenen Aktivatorbindungsstellen $[AT_n]$ zu der Gesamtkonzentration der Aktivatorbindungsstellen $[A_t]$ abhängig von der Transkriptionsfaktorkonzentration $[T]$ und der Gleichgewichtskonstanten K der Dissoziationsreaktion, in der der Operator mit n Transkriptionsfaktoren bindet. Das Verhältnis berechnet sich zu

$$\frac{[AT_n]}{[A_t]} \quad = \quad \frac{[T]^n}{K+[T]^n} =: f_3([T]) \tag{2.24}$$

mit $n \in \mathbf{N}^+, K \in \mathbf{R}^+, f_3 : \mathbf{R}_0^+ \to \mathbf{R}$.

BEWEIS:

Die Dissoziationsreaktion hat folgendes Aussehen:

$$AT_n \quad \rightleftharpoons \quad nT + A. \tag{2.25}$$

Es gilt

$$[A_t] \quad = \quad [A] + [AT_n]. \tag{2.26}$$

Wieder lässt sich die Gleichgewichtskonstante

$$K \quad = \quad \frac{[A][T]^n}{[AT_n]} \tag{2.27}$$

und daraus ein Zusammenhang zwischen dem Anteil gebundener Aktivatorbindungsstellen und der Transkriptionsfaktorkonzentration über die Gleichung (2.26) berechnen:

$$\begin{aligned}
&K = \frac{[A][T]^n}{[AT_n]} \\
&\Leftrightarrow K[AT_n] = [A][T]^n \\
&\Leftrightarrow_{(2.26)} K[AT_n] = [A_t][T]^n - [AT_n][T]^n \\
&\Leftrightarrow [AT_n](K + [T]^n) = [A_t][T]^n \\
&\Leftrightarrow \frac{[AT_n]}{[A_t]} = \frac{[T]^n}{K + [T]^n} = 1 - \frac{K}{K + [T]^n} = 1 - \beta
\end{aligned} \tag{2.28}$$

mit $\beta = [A]/[A_t]$ als relativer Häufigkeit der Konzentration von ungebundenen Aktivatorbindungsstellen zur Gesamtkonzentration von Aktivatorbindungsstellen.

□

Bemerkung 2.6. *Für die Funktion* $f_3 : \mathbf{R}_0^+ \to \mathbf{R}$ *gilt das asymptotische Verhalten* $\lim_{[T]\to\infty} f_3([T]) = 1$, *und des Weiteren* $f_3([T]) = 0$ *für* $[T] = 0$. *Für* $[T] = \sqrt[n]{K}$ *gilt* $f_3([T]) = 0.5$.

Im zweiten Fall sei die Regulierung eine Inhibierung: Es sei T ein Transkriptionsfaktor, der mit n seiner Moleküle an einen Operator O binden kann und dadurch die Transkriptionsrate der zu dem Operator gehörenden Gene verringert. Die Gleichungen (2.25) und (2.27) gelten wiederum. Nun möchte man jedoch Aufschluss über die relative Häufigkeit haben, mit der die Operatoren frei sind, also keine Inhibierung stattfindet.

Satz 2.7. *(direkte Inhibierung)*
Bei einer direkten Inhibierung durch den Transkriptionsfaktor T ist das Verhältnis der Konzentration von freien Operatoren $[O]$ *zu der Gesamtkonzentration der Operatoren* $[O_t]$ *abhängig von der Transkriptionsfaktorkonzentration* $[T]$ *und der Gleichgewichtskonstanten K der Dissoziationsreaktion, in der der Operator mit n Transkriptionsfaktormolekülen bindet. Das Verhältnis berechnet sich zu*

$$\frac{[O]}{[O_t]} \quad = \quad \frac{K}{K + [T]^n} =: f_4([T]) \tag{2.29}$$

mit $n \in \mathbf{N}^+, K \in \mathbf{R}^+, f_4 : \mathbf{R}_0^+ \to \mathbf{R}$.

BEWEIS:

Die relative Häufigkeit α berechnet sich wie folgt:

$$\begin{aligned} & K = \frac{[O][T]^n}{[OT_n]} \\ & \Leftrightarrow K[OT_n] = [O][T]^n \\ & \Leftrightarrow_{(2.8)} K[O_t] - K[O] = [O][T]^n \\ & \Leftrightarrow K[O_t] = [O]K + [O][T]^n \\ & \Leftrightarrow \alpha = \frac{[O]}{[O_t]} = \frac{K}{K+[T]^n} =: f_4([T]) \end{aligned} \tag{2.30}$$

□

In Abbildung 2.3 wird im Fall von direkter Aktivierung und Inhibierung die Funktion $1 - \alpha([T])$ bzw. $\alpha([T])$ mit verschiedenen Werten für n und K gezeigt. Wieder sieht man, dass die Funktion um so steiler fällt, je größer n ist. Je größer K ist, umso weiter nach rechts verschiebt sich der Wendepunkt der Funktion. Für $[T] = \sqrt[n]{K}$ ist $\alpha([T]) = 0.5$.

Bemerkung 2.8. *Für die Funktion* $f_4 : \mathbf{R}_0^+ \to \mathbf{R}$ *gilt das asymptotische Verhalten* $\lim_{[T]\to\infty} f_4([T]) = 0$ *und des Weiteren* $f_4([T]) = 1$ *für* $[T] = 0$. *Für* $[T] = \sqrt[n]{K}$ *gilt* $f_4([T]) = 0.5$.

Bemerkung 2.9. *Es gilt*

$$f_3([T]) \quad = \quad \frac{[T]^n}{K+[T]^n} = 1 - \frac{K+[T]^n}{K+[T]^n} + \frac{[T]^n}{K+[T]^n} = 1 - \frac{K}{K+[T]^n} = 1 - f_4([T]). \tag{2.31}$$

2.1.3 Das Modell

Fassen wir also die bisherigen Ergebnisse noch einmal zusammen. Wir betrachten vier Fälle von Genregulierungen und für jeden Fall berechnen wir die relative Häufigkeit von gebundenen bzw. ungebundenen Operatoren in der Gesamtpopulation. Der Grund hierfür ist, dass die Syntheserate als proportional zu dieser relativen Häufigkeit angenommen werden kann (Yagil u. Yagil, 1971).

- Indirekte Aktivierung: $f_1([E]) = \frac{\frac{\alpha_b}{K_1}[E]^n + \alpha_b}{\frac{\alpha_b}{K_1}[E]^n + \alpha_b + 1}$.
- Indirekte Inhibierung: $f_2([E]) = \frac{K_1\alpha_b + \alpha_b}{K_1\alpha_b + (\alpha_b+1)[E]^n}$.
- Direkte Aktivierung: $f_3([T]) = \frac{[T]^n}{K+[T]^n}$.
- Direkte Inhibierung: $f_4([T]) = \frac{K}{K+[T]^n}$.

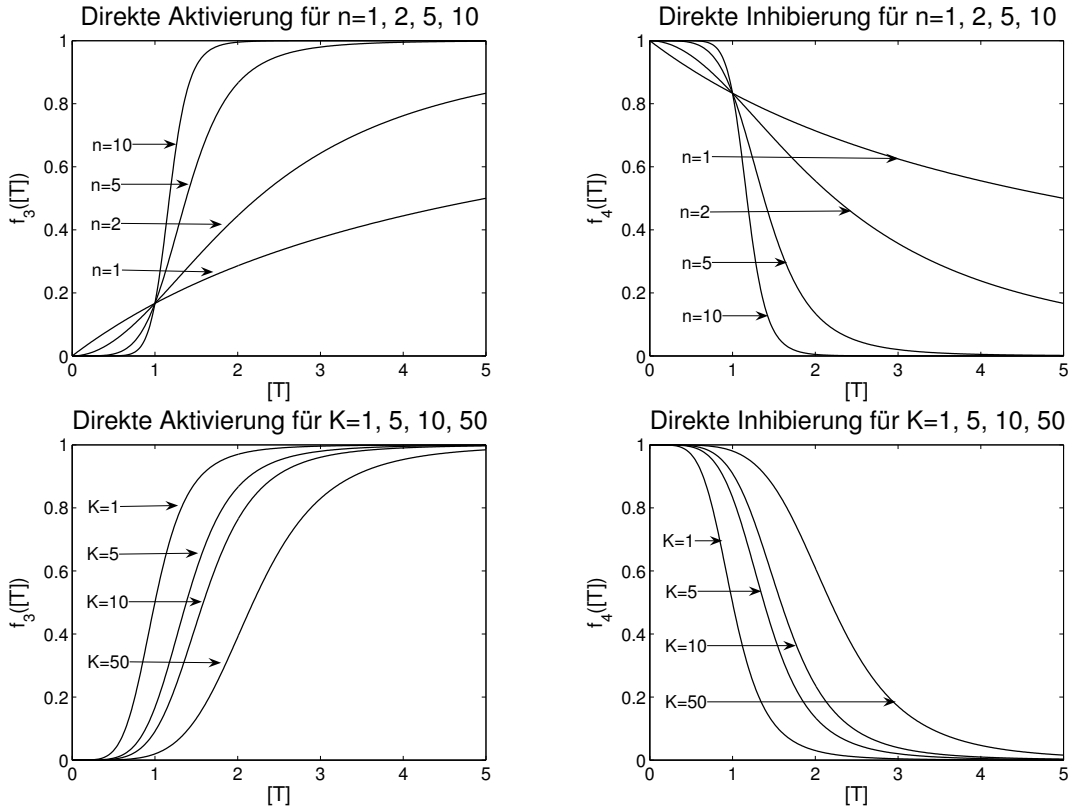


Abbildung 2.3: In der linken Abbildung wird die direkte Aktivierung $f_3([T]) = \frac{[T]^n}{K+[T]^n}$ mit verschiedenen Werten für n und K gezeigt, in der rechten Abbildung ist die direkte Inhibierung $f_4([T]) = \frac{K}{K+[T]^n}$ mit denselben variierenden Werten für n und K zu sehen.

Für $n > 1$ erhalten wir eine sigmoidale Funktion in sämtlichen der betrachteten Fälle. Je größer n wird, desto mehr nähern sich die Funktionen $f_1, ..., f_4$ einer Stufenfunktion an. Dies erklärt die zumindest qualitativ guten Ergebnisse, die Boolesche Netzwerke oder Differenzialgleichungen mit Stufenfunktionen liefern können. Solche Vereinfachungen sind nötig, da meist nur wenig Kenntnisse über die genauen Regulierungsprozesse vorliegen und dennoch eine Modellbildung ermöglicht werden soll. Das Ziel ist es nun, ähnlich wie bei der Approximierung der Regulierungsfunktionen durch Stufenfunktionen (vgl. Gleichung (1.26)) eine parametrisierte Regulierungsfunktion aufzustellen, die allgemein genug ist, um jeden der Prozesse zu beschreiben, die jedoch auch quantitative Aussagen zulässt.
Wie an den Funktionen f_1 und f_2 zu sehen ist, gibt es eine Grundexpression des regulierten Gens, das heißt, das Gen wird selbst ohne die Regulierungskomponenten mit einer gewissen Rate transkribiert. Im Folgenden sei mit der Grundexpression c_i von Gen i immer die Rate gemeint, mit der das Gen transkribiert wird, wenn sämtliche Transkriptionsfaktoren nicht vorhanden sind. Wir unterscheiden nun nur noch zwischen Inhibierungen und Aktivierungen. Bei Inhibierungen wird die Grundexpression entsprechend einer Inhibierungsfunktion $rf^- : \mathbf{R}_0^+ \to \mathbf{R}$ verringert, bei Aktivierungen wird sie entsprechend einer Aktivierungsfunktion $rf^+ : \mathbf{R}_0^+ \to \mathbf{R}$ erhöht. Die stückweise linearen und stetigen Funktionen rf^- und rf^+ sind abhängig von der Konzentration des jeweiligen Transkriptionsfaktors und werden wie folgt definiert:

$$rf^+([T]) = \begin{cases} 0 & \text{für } [T] < \theta_1 \\ \frac{k}{\theta_2 - \theta_1}([T] - \theta_1) & \text{für } \theta_1 \leq [T] \leq \theta_2 \\ k & \text{für } \theta_2 < [T] \end{cases}, \tag{2.32}$$

$$rf^-([T]) = k - rf^+([T]). \tag{2.33}$$

Der Parameter k bezeichnet bei der Aktivierungsfunktion die maximal mögliche Erhöhung der Syntheserate, bei der Inhibierungsfunktion die maximal mögliche Herabsetzung derselben. Somit steht $k > 0$ für Aktivierungen und $k < 0$ für Inhibierungen. Die Parameter θ_1 und θ_2 sind Schwellwerte. Liegt die Konzentration des Regulators unterhalb des ersten Schwellwertes θ_1, so beeinflusst der Regulator die Syntheserate noch nicht. Auch bei sehr hohen Konzentrationen des Regulators, also in dem Fall, dass die Konzentration oberhalb des zweiten Schwellwertes θ_2 liegt, ist eine Abhängigkeit von $[T]$ für den Einfluss auf die Syntheserate nicht mehr vorhanden, die Regulierung liegt im maximalen Bereich. Für $[T] \geq \theta_2$ bleibt der Einfluss also bei seinem Maximalwert k. Bewegt sich die Konzentration des Regulators innerhalb der Schwellwerte, so steigt der Einfluss auf die Syntheserate linear mit der Konzentration an. Für den Bereich zwischen den beiden Schwellwerten gilt $rf^+(\theta_1) = 0$ und $rf^+(\theta_2) = k$, so dass rf^+ und damit auch rf^- auf ganz $\mathbf{R}^+$ stetige Funktionen sind. Die Approximierung findet derart statt, dass die approximierte Funktion $l \cdot f$ und die approximierende Funktion rf sich an dem Punkt schneiden, an dem $l \cdot f([T^*]) = rf([T^*]) = k/2$ gilt. Der Vorteil an dieser Approximierung ist, dass die Funktionen in dem interessanten Bereich, der nahe an der Hälfte des Maximaleinflusses liegt, sehr gut übereinstimmen. Der Nachteil ist, dass bessere Approximierungen in dem Sinne existieren, dass sie einen geringeren Fehler im Vergleich zur ursprünglichen Funtkion aufweisen. Dies liegt daran, dass die Regulierungsfunktion f sich für $[T] \to 0$ schneller dem Grenzwert nähert als für $[T] \to \infty$ und so die favorisierte Approximierung im zweiten Bereich für $[T] > \sqrt[n]{K}$ schlechter ist als im ersten Bereich für $[T] < \sqrt[n]{K}$. Es lassen sich die Parameter k, θ_1 und θ_2 für eine

stückweise lineare Approximation rf beispielsweise für direkte Aktivierungen und Inhibierungen wie folgt berechnen: Es sei eine direkte Aktivierung über $l \cdot f_3([T]) = l \cdot ([T]^n/(K+[T]^n))$ gegeben. Dann lässt sich zunächst der Parameter k der Funktion rf^+ über $k = l$ bestimmen, da die Funktion $[T]^n/(K+[T]^n)$ für T gegen unendlich gegen 1 strebt. Die beiden Parameter θ_1 und θ_2 erhält man über eine Linearisierung der Funktion $l \cdot f_3([T])$ an derjenigen Stelle T^*, an der $l \cdot f_3([T^*]) = l/2$ gilt. Aufgrund von

$$l \cdot ([T^*]^n/(K+[T^*]^n)) = l/2 \quad \Leftrightarrow \quad [T^*] = \sqrt[n]{K}$$

wird die Linearisierung durch die Funktion

$$\tilde{f}_3([T]) \quad = \quad f_3(\sqrt[n]{K}) + \frac{f_3'(\sqrt[n]{K})}{1!}([T] - \sqrt[n]{K}) = l/2 + \frac{nl}{4\sqrt[n]{K}}([T] - \sqrt[n]{K})$$

nach der Taylor-Approximierung (Forster, 1992) vorgenommen. Setzt man die Funktion gleich Null beziehungsweise gleich l, so ergeben sich die Werte für θ_1 und θ_2 der Funktion rf^+ zu

- $\theta_1 = \sqrt[n]{K} - (2\sqrt[n]{K}/n)$,
- $\theta_2 = \sqrt[n]{K} + (2\sqrt[n]{K}/n)$.

In Abbildung 2.4 ist für die direkte Aktivierung $f([T]) = 1 \cdot \frac{[T]^5}{50+[T]^5}$ die entsprechende Approximierung $rf^+([T])$ mit $k = 1$, $\theta_1 = \sqrt[5]{50} - (2\sqrt[5]{50}/5)$ und $\theta_2 = \sqrt[5]{50} + (2\sqrt[5]{50}/5)$ zu sehen. Über

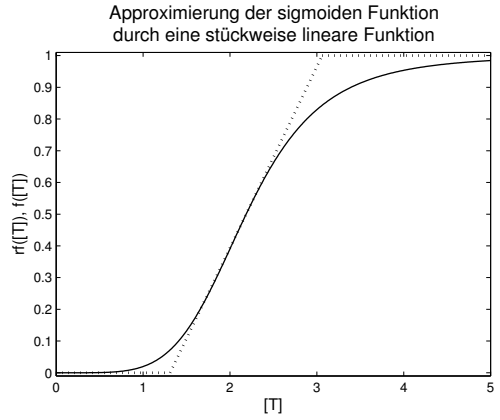


Abbildung 2.4: Regulierungsfunktion mit Approximierung durch eine stückweise lineare Funktion wie in Gleichung (2.32) beschrieben

$f_3([T]) = k - f_4([T])$ erhält man analog eine Approximierung für direkte Inhibierungen.

Sind multiplikative Effekte zwischen verschiedenen Transkriptionsfaktoren bekannt, so können diese ebenfalls parametrisiert in das Modell mit aufgenommen werden. Ist kein Wissen über solche multiplikativen Effekte vorhanden, so folgen wir der Annahme der Additivität, siehe auch

Chen u. a. (1999); Hoon u. a. (2002), um die Einfachheit des Modells zu erhalten.

Neben der Syntheserate muss ein Modell, das die Dynamik eines genregulatorischen Netzes abbilden will, auch die Abbaurate der jeweiligen Komponente berücksichtigen. Hier folgen wir der gängigen Annahme (siehe auch Chen u. a. (1999); Hoon u. a. (2002)), dass der Abbau proportional zur Konzentration der jeweiligen Komponente angenommen werden kann. Auch diese Annahme kann bei detailliertem Wissen über bestimmte Abbauvorgänge zurückgenommen werden, siehe dazu auch Kapitel 5.1.
An dieser Stelle ist die Herleitung des Modells abgeschlossen, und wir fassen zusammen:

Definition 2.10. *(Modell für genregulatorische Netzwerke)*
Gegeben seien n Gene, deren mRNA bzw. Proteine als Transkriptionsfaktoren wirken können. Die Konzentration der mRNA von Gen i definieren wir als x_i. Dann lässt sich die zeitliche Änderung von x_i für $i = 1,...,n$ beschreiben durch

$$\frac{dx_i(t)}{dt} = c_i - \gamma_i \cdot x_i(t) + \sum_{(i,j)\in\Omega_+} rf^+_{i,j}(x_j(t)) + \sum_{(i,j)\in\Omega_-} rf^-_{i,j}(x_j(t)), \tag{2.34}$$

wobei c_i die Grundsyntheserate bezeichnet, $\gamma_i x_i(t)$ die Abbaurate, $rf^+_{(i,j)}(x_j(t))$ bzw. $rf^-_{(i,j)}(x_j(t))$ die Aktivierungs- bzw. Inhibierungsfunktionen, die den Einfluss von Gen j auf Gen i beschreiben und die Kantenmengen Ω_+ bzw. Ω_- enthalten geordnete Paare (i, j) mit $i, j \in \{1,...,n\}$, wobei Gen j einen aktivierenden bzw. inhibierenden Einfluss auf Gen i hat.

Veröffentlicht wurde dieses Modell bzw. Vorläufer dieses Modells bisher in Gebert u. a. (2002, 2003a,b, 2004a, 2006, 2004b, 2007a,b); Gebert u. Radde (2006); Radde u. a. (2006); Akhmet u. a. (2005).
Die Schwellwerte der Regulierungsfunktionen rf^+ und rf^- zerlegen den Zustandsraum in Hyperrechtecke, so dass in jedem dieser Bereiche die Differenzialgleichungen aus (2.34) linear sind. Fasst man nun alle Konzentrationen in einem Vektor zusammen, also $x(t) = (x_1(t),...,x_n(t))^t$, so erhält man in einem Hyperrechteck q des Zustandsraums die Beschreibung der Dynamik des Systems durch

$$\dot{x}(t) := \frac{\mathrm{d}x(t)}{\mathrm{d}t} = A_q x(t) + b_q, \tag{2.35}$$

mit einer Matrix $A_q = (a_{q_{ij}})_{i,j=1,...,n} \in \mathbf{R}^{n\times n}$ und einem Vektor $b_q = (b_{q_1},...,b_{q_n})^t \in \mathbf{R}^n$. In der Matrix sind die Abbauparameter γ_i für $i = 1,...,n$ sowie die Regulierungen in den linearen Bereichen der Regulierungsfunktionen zusammengefasst. Aus einem Matrixeintrag $a_{ij} > 0$ mit $i \neq j$ kann gefolgert werden, dass Gen j das Gen i positiv reguliert, also aktiviert. Aus $a_{ij} < 0$ mit $i \neq j$ kann gefolgert werden, dass Gen j das Gen i negativ reguliert, also inhibiert.
Dieser Modellansatz enthält stetige Funktionen zur Beschreibung der Dynamik, jedoch auch einen diskreten Aspekt aufgrund der Zuordnung des Zustandsraumes zu bestimmten Matrix-Vektor-Paaren, so dass dieses Modell in die Klasse der hybriden Systeme fällt, siehe dazu auch Akhmet u. a. (2005); Öktem (2005). In der folgenden Definition wird das Modell mit stetigen sowie diskreten Funktionen beschrieben.

Definition 2.11. *(Modell als hybrides System)*
Das Modell (2.34) kann folgendermaßen beschrieben werden:

$$\begin{aligned} \dot{x}(t) &= A_{q(t)}x(t) + b_{q(t)}, \\ q(t) &= f(z(x(t))), \\ z(x(t)) &= (z_1(x(t)), ..., z_n(x(t))), \\ z_i(x(t)) &= \begin{cases} 0 & \text{für } x_i(t) \leq \theta_{i,1} \\ 1 & \text{für } \theta_{i,1} < x_i(t) \leq \theta_{i,2} \\ \vdots & \\ d_i & \text{für } \theta_{i,d_i} < x_i(t). \end{cases} \end{aligned} \tag{2.36}$$

Hierbei seien die Schwellwerte für Gen i geordnet durch $\theta_{i,1} < \theta_{i,2} < ... < \theta_{i,d_i}$, und es seien $A_{q(t)} \in \mathbf{R}^{n\times n}$, $b_{q(t)} \in \mathbf{R}^n$. Des Weiteren sei $z : \mathbf{R}^n \to N^n$ eine Funktion, deren Funktionswert angibt, wo sich der Vektor $x(t)$ bezüglich sämtlicher Schwellwerte zum Zeitpunkt t im Zustandsraum befindet. Die Funktion $f : \mathbf{N}^n \to \mathbf{N}$ wertet $z(x(t))$ aus und gibt den Index des Matrix-Vektor-Paares an, das in diesem Bereich des Zustandsraums zur Beschreibung der Dynamik gewählt werden muss.

2.2 Vorteile und Eigenschaften des Modells

In den letzten zwei bis drei Jahren ließ sich erkennen, dass sich Modelle für genregulatorische Netzwerke durchsetzen, die auf Differenzialgleichungen basieren. Dies liegt daran, dass nicht mehr nur Vergleichsexperimente, sondern auch Zeitreihenexperimente verfügbar sind und in Zukunft vermutlich in größerem Maße verfügbar sein werden. Genauer heißt dies, dass vor einigen Jahren überwiegend Experimente gemacht wurden, so dass eine Messung für Bedingung A und eine Messung für Bedingung B erfolgte und die Resultate in ihrer Höhe miteinander verglichen wurden. Dies wird bei Mikroarrays differenzielle Expressionsanalyse genannt und zielt darauf ab, einzelne Gene zu bestimmen, die unter bestimmten Bedingungen höher exprimiert sind als zu den Grundbedingungen. Aufgrund der gesunkenen Kosten für Mikroarrayexperimente und dem gestiegenen Interesse an Fragestellungen der Systembiologie ist jedoch zu erwarten, dass in Zukunft mehr Experimente verfügbar sein werden, in denen Messungen nach Eintreten einer Bedingung A zu mehreren aufeinander folgenden Zeitpunkten durchgeführt werden, bis beispielsweise keine Veränderung in der Expression mehr eintritt. Solche Experimente sind die Voraussetzung, um Parameter von Differenzialgleichungsmodellen zu schätzen.
Dennoch gibt es zwei Gründe, wieso möglichst einfache Differenzialgleichungssysteme zu bevorzugen sind, wie z. B. in Chen u. a. (1999); Hoon u. a. (2002); de Jong u. a. (2004). Zum einen nimmt die Anzahl der Parameter sehr stark zu, wenn man die erlaubten Funktionenklassen vergrößert (Hoon u. a., 2002), zum anderen sind Messungen innerhalb kürzester Zeit oft technisch nicht möglich, so dass beispielsweise Parameter einer Funktion, die zwischen zwei Messungen mehrere Extrema besitzt, nicht geschätzt werden können. Mit dem hier entwickelten Modell wird die Gratwanderung zwischen der benötigten Einfachheit der Funktionen und der Genauigkeit der Beschreibung der zugrunde liegenden biochemischen Prozesse zustande gebracht. In dem Modell wird eine Verallgemeinerung der verschiedenen Regulierungsfunktionen

vorgenommen, so dass nur noch zwischen Aktivierung und Inhibierung unterschieden wird. Dies wird durch Funktionen erreicht, die annähernd die Prozesse beschreiben und dennoch Einfachheit bewahren, indem affine Funktionen in den durch die Schwellwerte eingeteilten Bereichen verwendet werden. Hierdurch erhält man den großen Vorteil dieses Modells gegenüber Differenzialgleichungsmodellen mit höhergradigen Polynomen auf der rechten Seite, dass es in den einzelnen Bereichen des Zustandsraums analytisch lösbar ist. In einem Bereich Q werde das System beschrieben durch das konstante Matrix-Vektor-Paar (A,b), also

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = Ax(t)+b \text{ mit } A \in \mathbf{R}^{n\times n} \text{ und } b \in \mathbf{R}^n. \tag{2.37}$$

Dann lässt sich dieses System mit Methoden aus der Theorie der gewöhnlichen Differenzialgleichungen leicht lösen (Walter, 1993).

Neben dem Vorteil, die Lösungen explizit aufschreiben zu können, hat das Modell jedoch noch weitere Vorteile. Insbesondere ist hervorzuheben, dass die Regulierungsfunktionen des Modells auf den biologischen Prozessen der Bindungen von Transkriptionsfaktoren an die DNA basieren. Daraus folgt, dass sämtliche Verhaltensweisen, die auf diesen Prozessen basieren, auch modelliert werden können. Hierzu muss jedoch die Approximierung der sigmoidalen Funktion durch stückweise lineare Funktionen, die aus zwei konstanten und einer affinen Funktion bestehen, genau genug sein. Modellierbare Verhaltensweisen, die in biologischen Systemen zu beobachten sind, sind zum Beispiel Multistationarität, Oszillationen und Hysterese, siehe auch Thomas (1998); de Jong (2002).
Multistationarität tritt beispielsweise in Lambda Phage auf (Ptashne, 1986). Hierbei lassen sich verschiedene Anfangsbedingungen finden, die auf unterschiedliche Gleichgewichtspunkte hinauslaufen, also Zustände, in denen sich $x(t)$ für $t \to \infty$ nicht mehr verändert. Solch ein Verhalten lässt sich bereits mit einer Variablen modellieren:

$$\dot{x}_1(t) = c_1 - \gamma_1 x_1(t) + r f^+_{1,1}(x_1(t)). \tag{2.38}$$

Hierbei unterliegt x_1 einer positiven Selbstregulierung

$$r f^+_{1,1}(x_1(t)) = \begin{cases} 0 & \text{für } x_1(t) \le \theta_{1,1,1} \\ \frac{k_{1,1}}{\theta_{1,1,2}-\theta_{1,1,1}}(x_1(t)-\theta_{1,1,1}) & \text{für } \theta_{1,1,1} < x_1(t) < \theta_{1,1,2} \\ k_{1,1} & \text{für } \theta_{1,1,2} \le x_1(t) \end{cases}. \tag{2.39}$$

Es gibt drei Gleichgewichtspunkte, für die $\dot{x}_1 = 0$ gilt, und zwar

$$\bar{x}_1 = c_1/\gamma_1, \qquad \bar{x}_1 = (c_1+k_{1,1})/\gamma_1 \text{ und } \bar{x}_1 = \frac{-c_1+\theta_{1,1,1}}{k_{1,1}/(\theta_{1,1,2}-\theta_{1,1,1})-\gamma_1}. \tag{2.40}$$

Liegen die Anfangsbedingungen $x_1(t_0) < \frac{-c_1+\theta_{1,1,1}}{k_{1,1}/(\theta_{1,1,2}-\theta_{1,1,1})-\gamma_1}$, läuft $x_1(t)$ für $t \ge t_0$ auf den ersten Gleichgewichtspunkt c_1/γ_1 zu, für $x_1(t_0) > \frac{-c_1+\theta_{1,1,1}}{k_{1,1}/(\theta_{1,1,2}-\theta_{1,1,1})-\gamma_1}$ läuft $x_1(t)$ für $t \ge t_0$ auf den zweiten Gleichgewichtspunkt $(c_1+k_{1,1})/\gamma_1$ zu und nur für $x_1(t_0) = \frac{-c_1+\theta_{1,1,1}}{k_{1,1}/(\theta_{1,1,2}-\theta_{1,1,1})-\gamma_1}$ bleibt $x_1(t)$ für

$t \geq t_0$ bei diesem Wert.
Oszillationen benötigen mindestens zwei Variablen. In Kapitel 2.3 wird ein Beispiel zur Erläuterung des Modells gegeben, das auch Oszillationen enthält.

Das Modell hat noch einen weiteren Vorteil, da es in sich stabil in dem Sinne ist, dass es Grenzen $K_{\max} \in \mathbf{R}^n$ und $K_{\min} \in \mathbf{R}^n$ gibt, in denen sich sämtliche Trajektorien letztendlich befinden. Das bedeutet, dass es für jede Trajektorie $x(t)$, die der Anfangsbedingung $x(t_0)$ gehorcht, einen Zeitpunkt t^* gibt, so dass $K_{\min} \leq x(t) \leq K_{\max}$ für jedes $t \geq t^*$ gilt. Es lassen sich Schranken $K_{\min}$ und $K_{\max}$ finden, indem man die maximalen Einflussstärken k_{ij} der Regulierungsfunktionen für jedes $i = 1, ..., n$ addiert:

$$K_{i,\max} = (c_i + \sum_{j:\exists rf^+_{i,j}}^{n} k_{ij})/\gamma_i \tag{2.41}$$

und

$$K_{i,\min} = (c_i - \sum_{j:\exists rf^-_{i,j}}^{n} k_{ij})/\gamma_i. \tag{2.42}$$

Hierbei sind die Parameter k_{ij} entsprechend der Regulierungsfunktionen (2.32) definiert. Allerdings sind dies nicht die kleinsten oberen und größten unteren Schranken, die zu finden sind. Solch ein Supremum und Infimum muss je nach System von Differenzialgleichungen spezifisch bestimmt werden. Gibt es beispielsweise nur einen Attraktor, so können Infimum und Supremum trivialerweise über die Grenzen dieses Attraktors definiert werden.

Gründe für die Wahl dieses Modells

Die oben genannten Vorteile sind ein Grund dafür, wieso das hier eingeführte Modell für die späteren Anwendungen ausgewählt wurde. Hauptgrund ist jedoch die Zielsetzung, die mit diesem Modellansatz verfolgt wurde. Im Gegensatz zu den Booleschen Netzwerken (siehe Kapitel 1.3.1) sollen nicht nur zwei mögliche Zustände für die Konzentrationen erlaubt sein, sondern es sollen stetige Übergänge möglich sein. Um dynamisches Verhalten in genregulatorischen Netzwerken zu modellieren, existiert bereits eine Vielzahl von Modellen, die auf Differenzialgleichungen basieren. Unser Modell soll jedoch nicht nur lokal gute Ergebnisse erzeugen, sondern auch nur solche globalen Verhaltensmuster ermöglichen, die biologisch plausibel erscheinen. Dies steht im Gegensatz zu den einfachen linearen Modellen, wie sie in Kapitel 1.3.3 zu finden sind. Stattdessen wurden die biologischen Prozesse selbst zugrunde gelegt, so dass die Funktionen parametrisiert vorgegeben werden können. Die Parameter haben außerdem ihre biologischen Entsprechungen (z. B. Abbau- und Syntheseraten) und können daher einfach interpretiert werden. Das Modell wurde demnach entworfen, um den folgenden Zielsetzungen gerecht zu werden:

- Die Funktionen sollen direkt von den biologischen Prozessen abgeleitet werden.
- Die Funktionen sollen so weit vereinfacht werden, dass das System analytisch lösbar ist.

- Bei den Vereinfachungen sollen die möglichen Verhaltensmuster, die das System zeigen kann, erhalten bleiben.
- Die Parameter sollen einfach interpretierbar sein.
- Das Grundmodell soll derart erweiterbar sein, dass zusätzliche Variablen in späteren Modellierungsschritten einfügbar sind.

2.3 Beispiel

Im Folgenden wird anhand eines Beispiels das Modell erläutert. Es gebe zwei Gene, die sich jeweils gegenseitig und sich selbst beeinflussen, jedoch keinen weiteren Einwirkungen unterworfen sind. Gen 1 beeinflusse sowohl sich selbst als auch Gen 2 negativ. Gen 2 beeinflusse sich selbst und auch Gen 1 positiv, siehe auch Abbildung 2.5 links. Nach dem hier vorgestellten Modell sieht das System von Differenzialgleichungen wie folgt aus:

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = c_1 - \gamma_1 x_1(t) + rf^-_{1,1}(x_1(t)) + rf^+_{1,2}(x_2(t)) \tag{2.43}$$

$$\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} = c_2 - \gamma_2 x_2(t) + rf^-_{2,1}(x_1(t)) + rf^+_{2,2}(x_2(t)) \tag{2.44}$$

Wieder bezeichnet x_i für $i = 1,2$ die mRNA-Konzentration von Gen i. In diesem Beispiel wählen wir die Parameter wie in Tabelle 2.1 aufgelistet. Wie der Tabelle zu entnehmen ist, gehen

Tabelle 2.1: Parameter für das Beispielmodell.

Parameter	Wert	Parameter	Wert	Parameter	Wert	Parameter	Wert
c_1	3.5	$k_{1,1}$	-0.3	$\theta_{1,1,1}$	4.0	$\theta_{2,1,1}$	4.0
c_2	1.9	$k_{1,2}$	1.5	$\theta_{1,1,2}$	4.3	$\theta_{2,1,2}$	4.3
γ_1	1	$k_{2,1}$	-1.5	$\theta_{1,2,1}$	1.3	$\theta_{2,2,1}$	1.3
γ_2	1	$k_{2,2}$	0.9	$\theta_{1,2,2}$	1.6	$\theta_{2,2,2}$	1.6

wir davon aus, dass die beiden Regulierungen durch Gen i ($i = 1$ oder 2) jeweils dieselben zwei Schwellwerte besitzen. Dadurch erhält man eine Einteilung des Zustandsraums in neun Bereiche, siehe auch Abbildung 2.5 rechts. Die vier Regulierungsfunktionen teilen sich auf in zwei Inhibierungsfunktionen und zwei Aktivierungsfunktionen:

$$rf^-_{1,1}(x_1) = \begin{cases} 0 & \text{für } x_1 < 4.0 \\ -x_1 + 4 & \text{für } 4.0 \le x_1 \le 4.3, \\ -0.3 & \text{für } 4.3 < x_1 \end{cases} \quad rf^+_{1,2}(x_2) = \begin{cases} 0 & \text{für } x_2 < 1.3 \\ 5x_2 - 6.5 & \text{für } 1.3 \le x_2 \le 1.6 \\ 1.5 & \text{für } 1.6 < x_2 \end{cases}$$

$$rf^-_{2,1}(x_1) = \begin{cases} 0 & \text{für } x_1 < 4.0 \\ -5x_1 + 20 & \text{für } 4.0 \le x_1 \le 4.3, \\ -1.5 & \text{für } 4.3 < x_1 \end{cases} \quad rf^+_{2,2}(x_2) = \begin{cases} 0 & \text{für } x_2 < 1.3 \\ 3x_2 - 3.9 & \text{für } 1.3 \le x_2 \le 1.6 \\ 0.9 & \text{für } 1.6 < x_2 \end{cases}$$

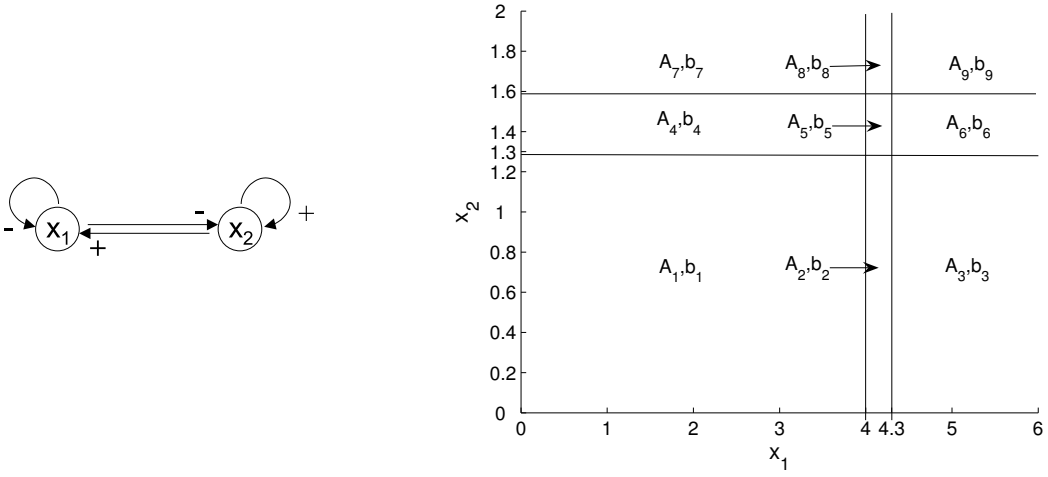


Abbildung 2.5: In der linken Abbildung ist das Beispielnetzwerk schematisch aufgezeichnet. In der rechten Abbildung ist die Einteilung des Zustandsraums bezüglich der Schwellwerte dargestellt. Für jedes Rechteck lässt sich die zeitliche Änderung des genregulatorischen Netzwerkes durch eine affine Funktion $A_i x + b_i$ für ein $i \in \{1,...,9\}$ beschreiben.

Insgesamt sind im Beispiel also neun Matrix-Vektor-Paare zu bestimmen, um das Verhalten des Beispielnetzwerkes berechnen zu können. Im Fall von verschiedenen Schwellwerten erhielte man 25 Bereiche. Allgemein hat man für n_1 verschiedene Schwellwerte für Gen 1 und n_2 verschiedene Schwellwerte für Gen 2 insgesamt $(n_1+1)\cdot(n_2+1)$ Bereiche im Zustandsraum, für die die Matrix-Vektor-Paare bestimmt werden müssen. Noch allgemeiner lässt sich diese Beobachtung wie folgt formulieren:

Bemerkung 2.12. *Für ein genregulatorisches Netzwerk mit n Genen, in denen n_i verschiedene Schwellwerte für Gen i existieren, erhält man $\prod_{i=1}^{n}(n_i+1)$ verschiedene Bereiche im Zustandsraum, in denen das Verhalten des genregulatorischen Netzwerkes durch affine Funktionen beschrieben werden kann.*

Gilt $x_1 > 4.3$ und $x_2 > 1.6$, so liegen beide Variablen über den Schwellwerten, das heißt, der Einfluss ist nicht mehr proportional zur regulierenden Komponente, sondern konstant. Es gilt in diesem Fall das folgende System von Differenzialgleichungen:

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = c_1 - \gamma_1 x_1(t) + k_{1,2} + k_{1,1} = 3.5 - x_1(t) + 1.5 - 0.3 = 4.7 - x_1(t) \quad (2.45)$$

$$\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} = c_2 - \gamma_2 x_2(t) + k_{2,2} + k_{2,1} = 1.9 - x_2(t) + 0.9 - 1.5 = 1.3 - x_2(t). \quad (2.46)$$

Das Paar (A_9, b_9) setzt sich also zusammen aus

$$A_9 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{und} \quad b_9 = \begin{pmatrix} 4.7 \\ 1.3 \end{pmatrix}.$$

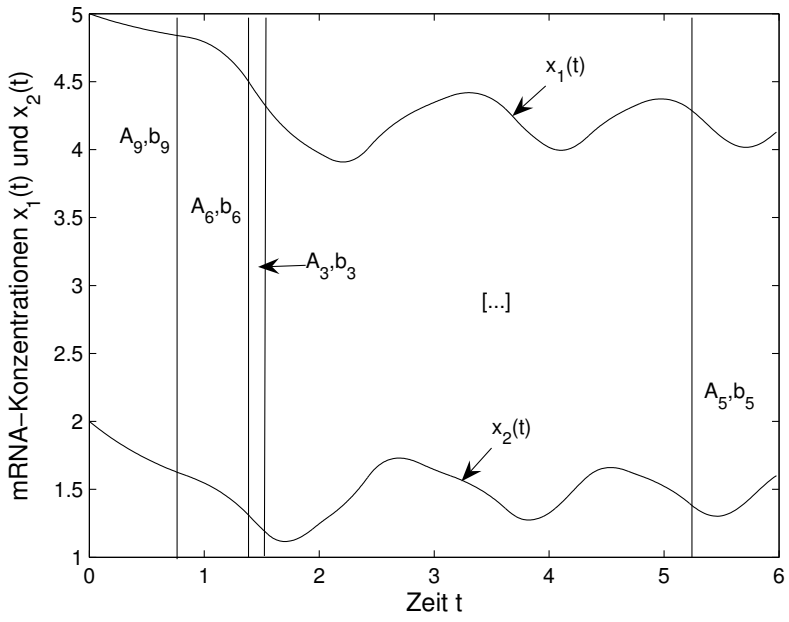


Abbildung 2.6: Die Lösungen $(t, x_1(t))$ und $(t, x_2(t))$ für das System von Differenzialgleichungen (2.43) sind für die Anfangsbedingung $x(0) = (5,2)^t$ dargestellt. Die vertikalen Linien zeigen die Schwellwertüberschreitungen an. Das System läuft also zunächst aus dem Bereich, in dem es durch A_9, b_9 beschrieben wird in den Bereich, in dem es durch A_6, b_6 charakterisiert wird und von dort in den Bereich mit der Beschreibung des Systems durch A_3, b_3 usw.

Für die Anfangswerte $x(0) = (5,2)^t$ wird somit $\dot{x}(t)$ beschrieben durch $\dot{x}(t) = A_9 x(t) + b_9$. Die Lösung dieser Differenzialgleichung lautet für unsere Anfangsbedingung $x(0)$:

$$x(t) \quad = \quad \begin{pmatrix} 0.3e^{-t} + 4.7 \\ 0.7e^{-t} + 1.3 \end{pmatrix}.$$

Die Trajektorie läuft zunächst in den Bereich, in dem das System durch A_6, b_6 beschrieben wird, da x_2 zuerst den Schwellwert 1.6 unterschreitet (x_1 kommt an seinen Schwellwert in diesem Bereich nicht heran, da der Fixpunkt von x_1 bei $\bar{x}_1 = 4.7$ liegt). Die Schwellwertunterschreitungen sind in Abbildung 2.6 skizziert. Der genaue Zeitpunkt des Erreichens des Schwellwertes für x_2 ist durch die Gleichung

$$0.7e^{-t} + 1.3 \quad = \quad 1.6$$

berechenbar. Durch Lösen der Gleichung nach t berechnet man den Zeitpunkt $t = -ln(\frac{3}{7}) = 0.8473$, zu dem x_2 den Schwellwert 1.6 erreicht. Das System von Differenzialgleichungen in

diesem neuen Bereich lautet

$$\frac{dx_1(t)}{dt} = 3.5 - x_1(t) - 0.3 + 5x_2(t) - 6.5 = -x_1(t) + 5x_2(t) - 3.3 \quad (2.47)$$

$$\frac{dx_2(t)}{dt} = 1.9 - x_2(t) - 1.5 + 3x_2(t) - 3.9 = 2x_2(t) - 3.5, \quad (2.48)$$

somit ist (A_6, b_6) also gegeben durch

$$A_6 = \begin{pmatrix} -1 & 5 \\ 0 & 2 \end{pmatrix} \quad \text{und} \quad b_6 = \begin{pmatrix} -3.3 \\ -3.5 \end{pmatrix}.$$

Zur Lösung dieses Systems von Differenzialgleichungen benötigen wir neben der neuen Anfangsbedingung $x_2(-ln(\frac{3}{7})) = 1.6$ auch den Wert für x_1 zum Zeitpunkt $t = -ln\frac{3}{7}$. Dieser Wert berechnet sich zu $0.3e^{-(-ln\frac{3}{7})} + 4.7 = 4\frac{29}{35}$. Die Lösung ergibt sich damit zu

$$x(t) = \begin{pmatrix} -\frac{13}{15}e^{-t} - \frac{9}{196}e^{2t} + \frac{109}{20} \\ -\frac{27}{980}e^{2t} + \frac{7}{4} \end{pmatrix}.$$

Wieder erreicht x_2 zuerst einen Schwellwert, nämlich den unteren Schwellwert 1.3. Durch Lösen der Gleichung $-\frac{27}{980}e^{2t} + \frac{7}{4} = 1.3$ nach t erhält man $t = ln(\frac{49}{3})/2 \approx 1.3966$ als Zeitpunkt, an dem $x(t)$ den Bereich mit A_6, b_6 verlässt und in den Bereich übergeht, in dem das System durch $\dot{x}(t) = A_3 x(t) + b_3$ mit

$$A_3 = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{und} \quad b_3 = \begin{pmatrix} 3.2 \\ 0.4 \end{pmatrix}$$

beschrieben wird. Die Variable x_1 hat zum Zeitpunkt $t = ln(\frac{49}{3})/2$ den Wert $x_1(ln(\frac{49}{3})/2) = 4.7 - \frac{13\sqrt{3}}{105} \approx 4.4856$. Anschließend unterschreitet x_1 den Schwellwert 4.3, so dass das System nun durch $\dot{x}(t) = A_2 x(t) + b_2$ beschrieben wird mit

$$A_2 = \begin{pmatrix} -2 & 0 \\ -5 & -1 \end{pmatrix} \quad \text{und} \quad b_2 = \begin{pmatrix} 7.5 \\ 21.9 \end{pmatrix}.$$

Nun überschreitet die Variable x_2 wieder ihren Schwellwert 1.3, so dass das System in dem mittleren Bereich landet, wo

$$\dot{x}(t) = \begin{pmatrix} -2 & 5 \\ -5 & 2 \end{pmatrix} x(t) + \begin{pmatrix} 1 \\ 18 \end{pmatrix}$$

gilt. Das System durchläuft also nacheinander Bereiche des Zustandsraums, in denen das System jeweils durch lineare Differenzialgleichungen charakterisiert ist. Letztendlich nähert sich die Lösung einem periodischen Attraktor.
Solch ein periodisches Verhalten ist auch bereits in einem einzigen Bereich möglich. In diesem Beispiel enthält der durch A_5, b_5 beschriebene Bereich solch einen periodischen Attraktor, siehe auch Abbildung 2.7. Mit der Anfangsbedingung $x(0) = (4.2, 1.5)^t$ befindet sich $x(0)$ in dem Bereich des Zustandsraums, der durch A_5 und b_5 beschrieben wird. Dieses System von linearen

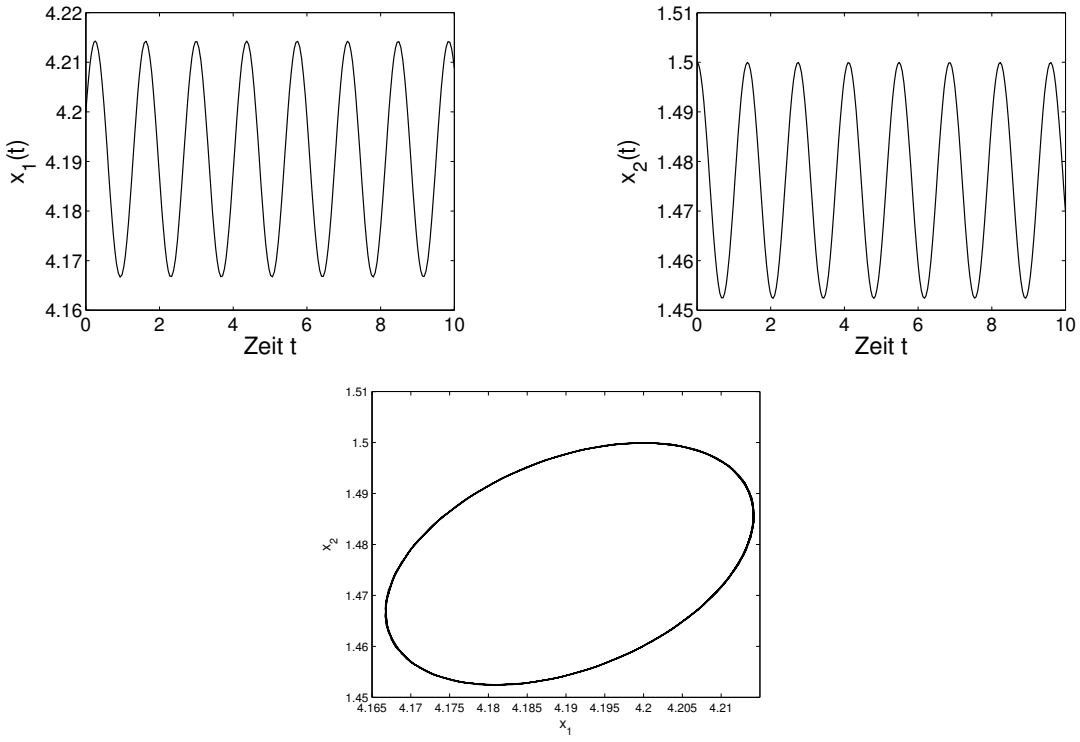


Abbildung 2.7: Oszillationen können bereits durch ein einziges System linearer Differenzialgleichungen hervorgerufen werden. Hier wird für das Beispiel der Bereich gewählt, der durch A_5 und b_5 beschrieben wird, also $\dot{x}(t) = A_5 x(t) + b_5$. Mit der Anfangsbedingung $x(0) = (4.2, 1.5)^t$ erhält man Oszillationen, die keine weiteren Bereiche im Zustandsraum erreichen. Oben links sieht man $(t, x_1(t))$, oben rechts $(t, x_2(t))$ und unten die Trajektorie $x(t) = (x_1(t), x_2(t))^t$.

Differenzialgleichungen reicht hier bereits aus, um periodisches Verhalten zu erzeugen. Dies liegt daran, dass $A = \begin{pmatrix} -2 & 5 \\ -5 & 2 \end{pmatrix}$ zwei rein komplexe Eigenwerte, nämlich $\lambda_1 = \sqrt{21}i$ und $\lambda_2 = -\sqrt{21}i$ besitzt. Anhand der Eigenwerte der Matrix kann bereits das Verhalten abgelesen werden, siehe dazu Walter (1993); Hirsch u. Smale (1974). Die Matrizen aller übrigen Bereiche besitzen reelle Eigenwerte, so dass in dem Fall Oszillationen dadurch hervorgerufen werden, dass die Gleichgewichtspunkte in einem anderen Bereich als dem jeweils ausgewählten liegen.

Kapitel 3

Komponentenauswahl für Modelle genregulatorischer Netzwerke

Im ersten und zweiten Kapitel wurde der zentralen Frage nachgegangen, welcher Modellierungsansatz der gegebenen Datenlage und der jeweils übergeordneten Fragestellung gerecht wird. Es wird also angenommen, dass die Komponenten, die das interessierende Subsystem definieren, schon bekannt sind. Unter Komponenten werden hier all diejenigen Bestandteile einer Zelle gefasst, die im Modell als Variable aufgenommen werden können, also beispielsweise Gene, mRNA oder Proteine. Sieht man sich verschiedene Anwendungen von Modellen auf Subsysteme in einer Zelle an, so wird deutlich, dass die wichtigsten Komponenten zwar als bekannt angenommen werden können, jedoch nicht vorhergesehene Verhaltensweisen üblicherweise durch Einflüsse von weiteren Komponenten erklärt werden. Neben der Frage nach dem Modellierungsansatz ist demnach auch die folgende Frage bedeutsam: Welche Auswahl von Komponenten wird für die Modellierung verwendet? Welches sind also neben den bereits bekannten Komponenten diejenigen Komponenten, die im System ebenfalls eine wichtige Rolle spielen?
In Hashimoto u. a. (2004) wurde eine Methode vorgestellt, die ebenfalls die Wahl der Komponenten in genregulatorischen Netzwerken betrifft. Mithilfe von Genexpressionsdaten wird hier eine Quelle bestehend aus einem oder mehreren Genen derart erweitert, dass die Autonomie des Subnetzwerkes iterativ vergrößert wird. Autonomie wird hierbei definiert über zwei Eigenschaften von genregulatorischen Subnetzwerken. Die Gene müssen stark miteinander interagieren und sie dürfen nicht stark von anderen Genen außerhalb des Netzwerkes abhängig sein. Die Methode wurde für probabilistische Boolesche Netzwerke entwickelt und ist erweiterbar auf endliche Quantifizierungen von Genexpressionen. Eine Anwendung der Methode mit probabilistischen Booleschen Netzwerken erfolgte bereits auf Glioma-Genexpressionsdaten (Hashimoto u. a., 2004). Wenig später wurde ein Ansatz in Bansal u. a. (2006) vorgestellt, der ebenfalls ein Subnetzwerk um ein ausgewähltes Gen aufbaut. Hierbei werden Störungsexperimente des ausgewählten Gens durchgeführt und die Antwort der Genexpressionsprofile über mehrere darauf folgende Zeitpunkte beobachtet, um ein genregulatorisches Netzwerk abzuleiten. Ein entgegengesetzter Ansatz ist ebenfalls möglich. Anstelle aus einigen wenigen Genen das Netzwerk zu erweitern kann auch versucht werden, aus allen Genen sämtliche regulatorischen Module zu bestimmen. Die meisten Clusteralgorithmen finden zwar Gene, die coreguliert sind, jedoch reicht eine Clusterung nicht aus, um die regulatorischen Module zu bestimmen. Gene werden bei den meisten Clusterungen nur einem Cluster zugeordnet, obwohl sie durchaus verschiedenen regulatorischen Modulen zugehörig sein können, und sie spielen oft nur in wenigen Experimenten

eine Rolle, so dass das Rauschen sehr stark ist (Kloster u. a., 2005). Signaturalgorithmen wie in Ihmels u. a. (2002) können diese Probleme beheben. Sie bestimmen eine Menge von coregulierten Genen zusammen mit Bedingungen, unter denen sie coreguliert sind. Hierbei kann das Wissen über bereits gefundene regulatorische Module für den nächsten Iterationsschritt mit verwendet werden, wie in Kloster u. a. (2005) beschrieben. Beide Ansätze beschränken sich auf die Verwendung von Genexpressionsdaten.
Im Folgenden werden zwei Ansätze vorgestellt, die Gene von Subnetzwerken bestimmen und die als Datengrundlage neben Genexpressionsdaten auch Interaktionstabellen zur Lösung des Problems heranziehen. $G = \{g_1, ..., g_n\}$ sei die Menge aller Gene eines Organismus und $S \subseteq G$ seien die Quellgene, also diejenigen Gene, die das Subnetzwerk definieren. Diesen Quellgenen sollen Gene hinzugefügt werden, die starke Interaktionen mit den Quellgenen eingehen. Wir bezeichnen diese hinzuzufügende Menge von Genen mit $\tilde{S} \subseteq G \setminus S$. Es sollen nicht nur die Gene des Subnetzwerkes bestimmt werden, sondern auch die Paare von Genen, die sich beeinflussen, sowie möglichst auch die Richtung der Regulierung. Neben G und S ist außerdem eine Relation $R \subseteq G \times G$ und eine $(n \times T)$-Matrix E gegeben. Interagieren g_i und g_j für $i, j \in \{1, ..., n\}$ miteinander, aktiviert also beispielsweise g_i das Gen g_j oder inhibiert g_j das Gen g_i, so ist $(g_i, g_j) \in R$. Die Relation ist symmetrisch, da die Richtung der Interaktion hier nicht vorgegeben ist. Die Matrix E besteht aus Expressionszeitreihen aller Gene G der Länge T. Es ist $E = (e_1, ..., e_n)^t$, wobei $e_i = (e_{i1}, ..., e_{iT})$ die Genexpressionsmessungen von Gen i zu den Zeitpunkten $1, .., T$ beschreibt. Für beide der folgenden Methoden wird ein empirischer Korrelationskoeffizient $\tau(e_i, e_j)$ für je zwei Genexpressionszeitreihen $e_i, e_j \in \mathbf{R}^T$ benötigt, um gleiche bzw. gegenläufige Entwicklungen in den Zeitreihen zu erfassen. Es sei $X = e_i$ und $Y = e_j$ für zwei Zeitreihen $e_i, e_j \in \mathbf{R}^T$. Wir verwenden den Kendallschen Korrelationskoeffizienten, da dieser sehr intuitiv ein Maß dafür bereitstellt, wie stark zwei Genexpressionsreihen dazu tendieren, sich gegen- oder gleichläufig zu verhalten. Die Definition basiert auf Kendall (1938):

Definition 3.1. *(Kendallscher Korrelationskoeffizient)*
Für $X, Y \in \mathbf{R}^T$ mit $X, Y \neq \alpha(1, ..., 1)^t$ für $\alpha \in \mathbf{R}$ definieren wir den Kendallschen Korrelationskoeffizienten durch

$$\tau_K(X,Y) = \frac{P - I}{\sqrt{\left(\frac{n(n-1)}{2} - T_X\right)\left(\frac{n(n-1)}{2} - T_Y\right)}}, \tag{3.1}$$

wobei P die Anzahl der Proversionen, I die Anzahl der Inversionen, T_X die Anzahl der Bindungen in X und T_Y die Anzahl der Bindungen in Y beschreibt. Jedes Paar (X_i, Y_i) wird dabei mit jedem anderen Paar (X_j, Y_j) verglichen. Eine Proversion liegt vor, wenn X und Y sich gleich verändern, das heißt

$$(X_i > X_j) \wedge (Y_i > Y_j) \text{ oder } (X_i < X_j) \wedge (Y_i < Y_j).$$

Eine Inversion liegt vor, wenn sich X und Y gegenläufig verändern, das heißt

$$(X_i > X_j) \wedge (Y_i < Y_j) \text{ oder } (X_i < X_j) \wedge (Y_i > Y_j).$$

Eine Bindung in X bzw. Y liegt vor, wenn $X_i = X_j$ bzw. $Y_i = Y_j$ gilt. Für konstantes X bzw. konstantes Y definieren wir wieder $\tau_K(X,Y) = 0$.

Der Kendallsche Korrelationskoeffizient ist bei größeren Stichproben rechnerisch aufwändig, da insgesamt $\frac{n(n-1)}{2}$ Vergleiche durchzuführen sind, jedoch ist er unempfindlich gegenüber Ausreißern und erkennt monoton steigende beziehungsweise monoton fallende Zusammenhänge. Des Weiteren ist er derart erweiterbar, dass Fehler in den Daten mit berücksichtigt werden können, siehe folgende Definition:

Definition 3.2. *(Erweiterter Kendallscher Korrelationskoeffizient)*
Für $X,Y \in \mathbf{R}^T$ ist der erweiterte Kendallsche Korrelationskoeffizient definiert durch

$$\tau_{\tilde{K}}(X,Y) \quad = \quad \frac{\tilde{P}-\tilde{I}}{\sqrt{\left(\frac{n(n-1)}{2}-\tilde{T}_X\right)\left(\frac{n(n-1)}{2}-\tilde{T}_Y\right)}}. \tag{3.2}$$

Für ein gegebenes $\rho > 0$, das entsprechend der Varianz des Datensatzes gewählt wird, entspricht $\tilde{P}$ der Anzahl der robusten Proversionen, das heißt es wird jedes Paar (X_i,Y_i) mit jedem anderen Paar (X_j,Y_j) verglichen und gezählt, wie oft

$$(X_i > X_j + \rho) \wedge (Y_i > Y_j + \rho) \text{ oder } (X_i < X_j - \rho) \wedge (Y_i < Y_j - \rho)$$

gilt. Entsprechend bezeichnet $\tilde{I}$ die Anzahl der robusten Inversionen, wobei eine robuste Inversion vorliegt, wenn

$$(X_i > X_j + \rho) \wedge (Y_i < Y_j - \rho) \text{ oder } (X_i < X_j - \rho) \wedge (Y_i > Y_j + \rho)$$

gilt. $\tilde{T}_X$ bzw. $\tilde{T}_Y$ bezeichnen die robusten Bindungen in X bzw. Y. Für die robusten Bindungen in X bzw. Y zählt man die Anzahl der Paare (X_i,X_j) bzw. (Y_i,Y_j), für die

$$X_i - \rho < X_j < X_i + \rho \text{ beziehungsweise } Y_i - \rho < Y_j < Y_i + \rho$$

gilt. Für Zeitreihen $X,Y \in \mathbf{R}^T$, für die gilt, dass entweder $\tilde{T}_X = \frac{n(n-1)}{2}$ oder $\tilde{T}_Y = \frac{n(n-1)}{2}$ ist, wird $\tau_{\tilde{K}}(X,Y) = 0$ definiert.

Kurze Zeitreihen oder Zeitreihen, die zeitlich sehr lange Abstände zwischen den Messungen haben, so dass zeitlich verschobene Antworten nicht mehr zu erkennen sind, sollten mit dem erweiterten Kendallschen Korrelationskoeffizienten analysiert werden. Hierbei kann man jedoch keine Aussage über die Richtung der Regulierung machen, sondern nur die Aussage tätigen, dass entweder eines der Gene das andere reguliert, dass beide coreguliert sind oder dass nur zufällig dasselbe Expressionsmuster zu erkennen ist und die Gene unabhängig voneinander reguliert werden. Bei langen Zeitreihen, in denen zeitlich verschobene Antworten auf die Hoch- oder Herunterregulierung eines Gens zu erkennen sind, gibt der zeitlich verschobene, erweiterte Kendallsche Korrelationskoeffizient Auskunft über die Richtung der Regulierung:

Definition 3.3. *(zeitlich verschobener, erweiterter Kendallscher Korrelationskoeffizient)*
Für $X,Y \in \mathbf{R}^T$ ist der zeitlich verschobene, erweiterte Kendallsche Korrelationskoeffizient definiert durch

$$\tau_{\tilde{K},t^*}(X,Y) \quad = \quad \tau_{\tilde{K}}(X^*,Y^*), \tag{3.3}$$

wobei $X^* = (X_1, ..., X_{T-k})^t$ *und* $Y^* = (Y_{k+1}, ..., Y_T)^t$ *für ein* $k \in \mathbb{N}^+$ *ist. Wieder wird für Zeitreihen* $X, Y \in \mathbf{R}^T$*, für die gilt, dass entweder* $\tilde{T}_{X^*} = \frac{n(n-1)}{2}$ *oder* $\tilde{T}_{Y^*} = \frac{n(n-1)}{2}$ *ist, der Korrelationskoeffizient als Null definiert.*

Der zeitlich verschobene Koeffizient verschiebt die Zeitpunkte der Reihe von Y derart, dass der neue Wert zum Zeitpunkt t jeweils dem alten Wert zum Zeitpunkt $t+k$ entspricht. Allerdings birgt dieser Korrelationskoeffizient durch die Verschiebung einige Nachteile. Die üblicherweise schon kurzen Zeitreihen, die bei Genexpressionsmessungen entstehen, werden bei diesem Korrelationskoeffizienten um weitere k Zeitpunkte gekürzt. Des Weiteren müssen die Messungen mit konstanten Zeitabständen durchgeführt worden sein, da sonst auch k nicht konstant gewählt werden kann.

Im Folgenden bezeichnen wir einen Korrelationskoeffizienten mit $\tau(X,Y)$, der je nach Qualität, Art und Umfang der Datenlage geeignet gewählt werden muss.

Bemerkung 3.4. *Für den Kendallschen Korrelationskoeffizienten und den erweiterten Kendallschen Korrelationskoeffizienten* $\tau : \mathbf{R}^T \times \mathbf{R}^T \to \mathbf{R}$ *und Expressionszeitreihen* $X, Y \in \mathbf{R}^T$ *gilt wie vom Pearsonschen Korrelationskoeffizienten bekannt*

1. $-1 \leq \tau(X,Y) \leq 1$,
2. $\tau(X,X) = 1$,
3. $\tau(X,-X) = -1$.

3.1 Graphentheoretischer Ansatz zur Komponentenauswahl

Die Methode dieses Kapitelabschnitts resultiert aus einem Forschungsaufenthalt in der biowissenschaftlichen Abteilung der Los Alamos National Laboratories, USA, im September und Oktober 2006. Hier wurde in einer Zusammenarbeit mit Nicole Radde und Christian Forst die folgende Methode entwickelt, welche in Radde u. a. (2006) publiziert wurde. Der erste graphentheoretische Teil der Methode basiert auf einem von Christian Forst entwickelten Algorithmus (Cabusora u. a., 2005; Mawuenyega u. a., 2005), der zweite und dritte Teil der Methode wurde zusammen mit Nicole Radde konzipiert und anhand von Daten für das *M. tuberculosis* umgesetzt (Radde u. a., 2006).

3.1.1 Graphentheoretische Auswahl einiger Komponenten

Aus den Genen G und der Relation R, die Aufschluss über Interaktionen zwischen den Genen gibt, wird ein Interaktionsgraph $\mathbb{G} = (G, R^*)$ gebildet.

Definition 3.5. *(Interaktionsgraph, (Cabusora u. a., 2005; Mawuenyega u. a., 2005))*
Gegeben ist eine Genmenge G und eine symmetrische Relation R. Der Interaktionsgraph $\mathbb{G} = (G, R^*)$ *besteht aus Knoten, die den Genen G entsprechen, und aus einer Kantenmenge* R^**, die folgendermaßen definiert ist.* R^* *enthält die ungerichtete Kante* $\{g_i, g_j\}$ *mit* $g_i, g_j \in G$*, wenn* (g_i, g_j) *und* $(g_j, g_i) \in R$ *sind.*

Um eine Gewichtsfunktion $l : R^* \rightarrow \mathbf{R}^+$ auf den ungerichteten Kanten zu bestimmen, wird die Genexpressionsmatrix E sowie ein Korrelationskoeffizient $\tau : \mathbf{R}^T \times \mathbf{R}^T \rightarrow \mathbf{R}$ verwendet. Da ein betragsmäßig hoher Korrelationskoeffizient zwischen zwei Genexpressionszeitreihen ein niedriges Kantengewicht verursachen soll, wird

$$l(g_i, g_j) \quad = \quad \Phi^{-1}(1 - |\tau(e_i, e_j)|) \tag{3.4}$$

gesetzt für $\{g_i, g_j\} \in R^*$ und e_i, e_j Genexpressionszeitreihen von Gen g_i bzw. g_j. $\Phi^{-1} : \mathbf{R}^+ \rightarrow \mathbf{R}$ ist dabei die Inverse der Verteilungsfunktion $\Phi : \mathbf{R} \rightarrow \mathbf{R}^+$ der Standardnormalverteilung. Aus diesem Interaktionsgraphen mit gewichteten Kanten wird ein Teilgraph berechnet, indem zwischen gegebenen Knoten $S \subseteq G$ k-kürzeste Wege bestimmt werden. Zwischen je zwei Knoten g_i und g_j mit $g_i, g_j \in S$ werden also der kürzeste, der zweitkürzeste, und so weiter, bis hin zum K-kürzesten Weg berechnet. Formal definiert sich das Problem wie folgt:

Problemstellung 3.6. *(k-kürzeste-Wege-Problem, (Jimenez u. Marzal, 1999))*
Es sei $\mathbb{G} = (G, R^*)$ *ungerichteter Interaktionsgraph mit Gewichtsfunktion* $l : R^* \rightarrow \mathbf{R}$*. Des Weiteren seien* $s, t \in G$ *gegeben sowie ein* $K \in \mathbf{N}^+$*. Ein Weg von s nach t ist definiert als Folge* $\pi = \pi_1, ..., \pi_a$*, so dass* $\pi_1 = s$*,* $\pi_a = t$ *und* $(\pi_i, \pi_{i+1}) \in R^*$ *für* $1 \leq i < a$ *gilt. Die Länge des Weges ist definiert als* $L(\pi) = \sum_{1 \leq i < a} l(\pi_i, \pi_{i+1})$*.* $P(t)$ *sei die Menge sämtlicher Wege mit* $\pi_1 = s$ *und* $\pi_a = t$*. Wir definieren* $P^k(t)$ *für* $k \geq 0$ *als die Menge der k-kürzesten Wege in* $P(t)$*. Der k-kürzeste Weg ist definiert als* $\pi^k(t) = \arg\min_{\pi \in P(t) \setminus P^{k-1}(t)} L(\pi)$ *und seine Länge ist* $L^k(t) = L(\pi^k(t)) = \min_{\pi \in P(t) \setminus P^{k-1}(t)} L(\pi)$*. Gesucht werden nun* $\pi^1(t), \pi^2(t), ..., \pi^K(t)$*.*

Für $K = 1$ läßt sich dieses Problem mit dem Dijkstra-Algorithmus (siehe z. B. Thulasiraman u. Swamy (1992)) lösen, da nur nichtnegative Kanten vorhanden sind. Der Dijkstra-Algorithmus sucht ausgehend von einem Startknoten s sukzessive den nächstbesten Knoten, der einen kürzesten Weg von s ausgehend besitzt. Für allgemeines $K \in \mathbf{N}^+$ löst ein rekursiver Enummerationsalgorithmus wie in Jimenez u. Marzal (1999) beschrieben das Problem. Zunächst berechnet man mithilfe des Dijkstra-Algorithmus kürzeste Wege für einen Startknoten s. Iterativ wird nun $L^k(t)$, also die Länge des k-kürzesten Weges von s nach t, und $\pi^k(t)$, also der Weg selbst, für $k = 2, ..., K$ berechnet. Man kennt die $(k-1)$-kürzesten Wege von s zu sämtlichen Knoten $u \in G$, die adjazent zu t sind. Die Länge des k-kürzesten Weges von s nach t läßt sich dann berechnen, indem man sämtliche $(k-1)$-kürzesten, $(k-2)$-kürzesten, bis hin zu den kürzesten Wegen von s zu den zu t adjazenten Knoten u um $l(u,t)$ verlängert und den k-kürzesten Weg unter den erhaltenen Wegen heraussucht.

Neben einer Angabe für K wird auch noch die maximale Länge der Wege L bestimmt, da Interaktionen über sehr viele Gene unwahrscheinlich sind. Implementiert und angewendet auf

Interaktionsdaten sowie auf Genexpressionsdaten des Fettsäuremetabolismus wurde der Algorithmus in Cabusora u. a. (2005); Mawuenyega u. a. (2005). Um eine Einschätzung der Parameter K und L zu erhalten, sollen folgende Intervalle erwähnt werden. Für einen Interaktionsgraphen des *M. tuberculosis*, der über verschiedene Genexpressionszeitreihen gewichtete Kanten enthält, konnten für weitere Analysen verwertbare Ergebnisse für folgende Werte erzielt werden: $1 \leq K \leq 5$, $5 \leq L \leq 10$.
Der gesuchte Teilgraph $\mathbb{G}^*$ des Interaktionsgraphen enthält alle Quellgene S sowie sämtliche k-kürzeste Wege zwischen Genen g_i und g_j mit $g_i, g_j \in S$ für $k \leq K$, die eine Länge kleiner oder gleich L haben. Zusätzlich zu den Quellgenen S erhalten wir also noch eine Menge S' an Genen, die auf diesen k-kürzesten Wegen liegen.

3.1.2 Statistische Analyse der ausgewählten Komponenten

Im vorherigen Schritt berücksichtigt der Teilgraph $\mathbb{G}^*$ sämtliche Gene, die auf k-kürzesten Wegen maximaler Länge L für $k \leq K$ liegen. Im Allgemeinen ist die Menge S' der neu zu den Quellgenen S hinzugekommenen Genen recht groß. Für eine Modellbildung möchte man jedoch nur die wichtigsten einflussreichen Gene mit in die Variablenliste aufnehmen. Daher wird im zweiten Schritt die Korrelation zwischen Genen aus S' und Genen aus S überprüft. Hierzu verwenden wir vorzugsweise den zeitlich verschobenen erweiterten Korrelationskoeffizienten $\tau_{\tilde{K},t^*}$. Nur Gene aus S', die betragsmäßig besonders hohe Korrelationskoeffizienten zu Genen aus S aufweisen, werden anschließend in der Modellbildung berücksichtigt. Ein besonders hoher Korrelationskoeffizient läßt sich nur in Relation zu allen übrigen Korrelationskoeffizienten definieren. Insbesondere ist es auch von der Wahl des Korrelationskoeffizienten abhängig, ab welchem Schwellwert ein Korrelationskoeffizient als signifikant hoch beschrieben werden kann. Bei der Messung der Transkription sehr vieler Gene, wie beispielsweise bei der Messung des gesamten Genoms, wird bei manchen Normalisierungsmethoden von Mikroarrays angenommen, dass auf einem Mikroarray genauso viele Gene hoch- wie herunterreguliert sind. Es wird angenommen, dass im Mittel das Verhältnis von behandelter Stichprobe zu unbehandelter Stichprobe gleich 1 ist beziehungsweise sich bei logarithmierten Werten ein Mittelwert von 0 ergibt. Ähnliches erwartet man bei den Korrelationskoeffizienten. Berechnet man die Korrelationen $\tau(e_i, e_j)$ für sämtliche Paare aus Expressionszeitreihen $\{e_i, e_j\}$, $i, j = 1, ..., n$, so werden insgesamt $\frac{n(n-1)}{2}$ Korrelationskoeffizienten $\tau_1, ..., \tau_{n(n-1)/2}$ berechnet und wir können analog zur Mikroarrayanalyse die Annahme machen, dass die Werte im Mittel ungefähr bei Null liegen. Die Korrelationskoeffizienten seien derart geordnet, dass $\tau_1, ..., \tau_l$ mit $l \in N^+$, $l \leq n(n-1)/2$ die voneinander verschiedenen Korrelationswerte sind. Wir betrachten nun die aus den Korrelationskoeffizienten resultierende diskrete Verteilung $\mathscr{D}$, die man erhält, indem man jedem Korrelationswert $\tau_1, ..., \tau_l$ eine Wahrscheinlichkeit zuordnet, die gleich der relativen Häufigkeit des Auftretens dieses Wertes in $\tau_1, ..., \tau_{n(n-1)/2}$ ist und durch $p(\tau_i)$ für $i = 1, ..., l$ bezeichnet wird. Der Raum (Ω, Σ, P) besteht also aus der Ergebnismenge $\Omega = \{\tau_1, ..., \tau_l\}$, der $\sigma-$Algebra $\Sigma = \mathscr{P}(\Omega)$ als Potenzmenge von Ω und dem Wahrscheinlichkeitsmaß P auf Σ, das über die relativen Häufigkeiten definiert ist, also $P(A) = p(\tau_{i_1}) + ... + p(\tau_{i_A})$ für $A \in \Sigma$, wobei $A = \{\tau_{i_1}, ..., \tau_{i_A}\}$ für verschiedene $i_1, ..., i_A \in \{1, ..., l\}$ sei.

Mithilfe des Chi-Quadrat-Tests läßt sich herausfinden, ob aufgrund der Werte $\tau_1, ..., \tau_{n(n-1)/2}$ sogar eine Normalverteilung für die Verteilung der Korrelationskoeffizienten angenommen werden kann (Bronstein u. Semendjajew, 1991).

Wir gehen nun von einem gerichteten vollständigen Graphen $K = (S \cup S', E')$ aus, in dem zwischen jedem geordneten Paar aus Genen (g_i, g_j) mit $g_i, g_j \in S \cup S'$ eine gerichtete Kante e_{ij} existiert.
Gegeben sei ein Signifikanzniveau $\alpha > 0$, dann kann jede Kante e_{ij} aufgrund des entsprechenden Korrelationswertes $\tau_{\tilde{K},t^*}(e_i, e_j)$ auf Signifikanz überprüft werden.

Definition 3.7. *(signifikante Kante)*
Eine Kante e_{ij}, die g_i und g_j verbindet, heißt signifikant bezüglich eines Signifikanzniveaus $\alpha > 0$ und der Wahl eines Korrelationskoeffizienten τ genau dann, wenn der entsprechende Korrelationswert $\tau(e_i, e_j)$ signifikant bezüglich der zugrunde gelegten Verteilung der Korrelationskoeffizienten ist.

Wir betrachten zunächst den Fall, dass eine Normalverteilung der Korrelationskoeffizienten mit Mittelwert m und Standardabweichung σ angenommen werden kann. Dann gilt:

Bemerkung 3.8. *Für die Korrelationskoeffizienten gelte die Verteilung $X \sim \mathcal{N}(m, \sigma)$. Die Kante e_{ij} ist signifikant bezüglich α, wenn für den entsprechenden Korrelationskoeffizienten $\tau(e_i, e_j)$ gilt:*

$$P(X \geq \tau(e_i, e_j)) \leq \alpha/2. \tag{3.5}$$

Es ist e_{ij} somit signifikant, wenn die Wahrscheinlichkeit, den Korrelationswert $\tau(e_i, e_j)$ oder eine noch größere Abweichung vom Mittelwert zu erreichen, kleiner oder gleich α ist.

Alternativ kann die diskrete Verteilung aus den Koeffizienten selbst für die Berechnung von signifikanten Kanten verwendet werden.

Bemerkung 3.9. *Die Kante e_{ij} ist signifikant bezüglich α, wenn für den zugehörigen Korrelationskoeffizienten $\tau(e_i, e_j)$ gilt:*

$$\tau(e_i, e_j) \leq \tau_{\min} \tag{3.6}$$

oder

$$\tau(e_i, e_j) \geq \tau_{\max}, \tag{3.7}$$

mit $\tau_{\min} = \max_{\tau \in T_{\min}}\{\tau\}$ und $\tau_{\max} = \min_{\tau \in T_{\max}}\{\tau\}$ ist. $T_{\min}$ enthält dabei die $(\alpha/2)(n(n-1)/2)$ kleinsten Werte der Reihe $\tau_1, ..., \tau_{n(n-1)/2}$, $T_{\max}$ die $(\alpha/2)(n(n-1)/2)$ größten Werte.

Ziel dieses Kapitels war die Bestimmung der Gene eines regulatorischen Subnetzwerkes. Dieses besteht aus den Quellgenen und soll um diejenigen Gene erweitert werden, die starke Interaktionen zu den Quellgenen aufweisen. Aus dem gerichteten, vollständigen Graphen $K = (S \cup S', E')$ werden zunächst die nicht signifikanten Kanten entfernt. Den hierdurch entstehenden Graphen nennen wir K'. Anschließend werden auch diejenigen Knoten $g' \in S'$ entfernt, für die gilt, dass kein $g \in S$ existiert, so dass (g', g) Kante in dem Graphen K' ist. Das heißt, es werden diejenigen Knoten entfernt, die keine gerichtete Kante mehr zu einem der Quellgene aufweisen (siehe auch Abbildung 3.1). Es sei betont, dass nur gerichtete Kanten von einem Knoten g' zu einem Quellgen g dazu führen, den Knoten g' im Subnetzwerk zu behalten. Eine gerichtete Kante von einem Quellgen g zu einem Knoten g' ist jedoch unbedeutend. Dies heißt nur, dass das Quellgen g dieses Gen g' reguliert, jedoch nicht von diesem beeinflusst wird. Die Kanten erhalten für entsprechende negative Korrelationskoeffizienten ein Minuszeichen als Markierung, für entsprechende positive Korrelationskoeffizienten ein Pluszeichen. Der entstandene Graph wird K^* genannt und enthält die Knoten $S \cup \tilde{S}$, die den Ausgangspunkt für die Modellierung der Dynamik des genregulatorischen Netzwerkes bilden. Für das in Kapitel 2 vorgestellte Modell be-

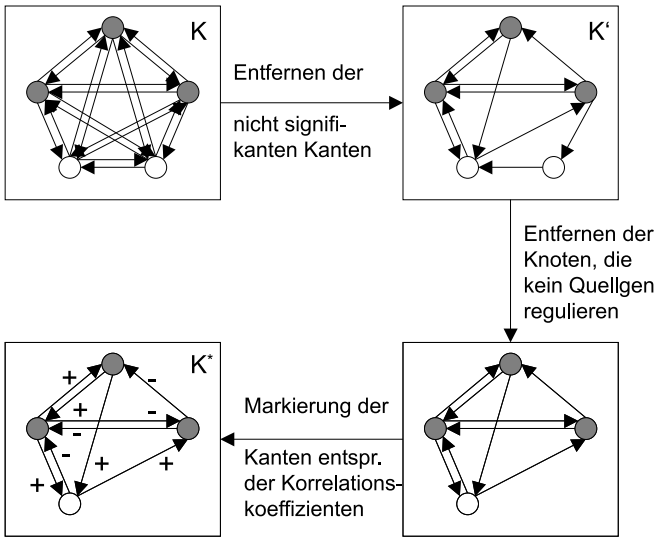


Abbildung 3.1: Ein gerichteter vollständiger Graph K mit den Genen $S \cup S'$ als Knoten ist unser Ausgangspunkt. Die Quellgene S sind durch graue Knoten, die zusätzlichen Gene durch weiße Knoten dargestellt. Zuerst werden nicht signifikante Kanten entfernt und man erhält den Graphen K'. Anschließend werden die Knoten entfernt, die keine Kante zu einem der Quellgene besitzen. Man erhält K^*, indem man den Kanten des entstandenen Graphen positive bzw. negative Markierungen entsprechend des zugehörigen Korrelationskoeffizienten zuteilt. Die verbleibenden Gene sind die gesuchten Gene der Menge $S \cup \tilde{S}$, die die Variablen in der anschließenden Modellierung bestimmen.

deutet dies, dass die Genmenge des zu modellierenden genregulatorischen Netzwerkes aus $S \cup \tilde{S}$ besteht. Wie bereits im vorherigen Kapitel eingeführt bezeichne x_i die Konzentration der mRNA

von Gen $g_i \in S \cup \tilde{S}$, $i = 1,...,|S \cup \tilde{S}|$. Dann lautet die Differenzialgleichung zu Beschreibung von x_i:

$$\dot{x}_i(t) \quad = \quad c_i - \gamma_i x_i(t) + f_i(x_1(t),...,x_{|S\cup\tilde{S}|}(t)), \tag{3.8}$$

wobei $c_i, \gamma_i \in \mathbf{R}^+$ und $f_i : \mathbf{R}^{|S\cup\tilde{S}|} \to \mathbf{R}$ Summe derjenigen Regulierungsfunktionen ist, die durch den entstandenen Graphen K^* bestimmt sind. Seien $g_{j_1},...,g_{j_\nu}$ mit $j_1,...,j_\nu \in \{1,...,|S \cup \tilde{S}|\}$ diejenigen Gene, die eine positiv markierte Kante zu Gen x_i besitzen, und $g_{j_{\nu+1}},...,g_{j_s}$ mit $j_{\nu+1},...,j_s \in \{1,...,|S \cup \tilde{S}|\}$ diejenigen Gene, die eine negativ markierte Kante zu Gen x_i besitzen. Dann hat die Regulierungsfunktion die Form

$$f_i(x_1(t),...,x_{|S\cup\tilde{S}|}(t)) \quad = \quad \sum_{j=j_1}^{j_\nu} rf_{i,j}^{+}(x_j(t)) + \sum_{j=j_{\nu+1}}^{j_s} rf_{i,j}^{-}(x_j(t)). \tag{3.9}$$

Der Vorteil dieser Komponentenauswahl liegt darin, dass K^* als Basis für verschiedene Modelle genregulatorischer Netzwerke verwendet werden kann. Nicht nur Differenzialgleichungsmodelle, sondern beispielsweise auch Boolesche und Bayessche Netzwerke, können auf K^* aufgebaut werden. Sowohl Vor- als auch Nachteil ist die Erstellung eines Interaktionsgraphen, der es ermöglicht, zusätzliches Wissen zu den experimentellen Messungen zu verwenden, jedoch bei lückenhaftem Wissen über die Interaktionen das Ergebnis der Methode dadurch verschlechtert, dass völlig unbekannte Gene nicht mit den Quellgenen über einen Pfad verbunden sind und somit nicht bei der Auswahl berücksichtigt werden.

Hat man kurze Zeitreihen vorliegen, die lange und unregelmäßige Zeitabstände zwischen den Messungen beinhalten, wie es beispielsweise in der Anwendung in Kapitel 6 vorkommt, so kann auch der erweiterte Kendallsche Korrelationskoeffizient $\tau_{\tilde{K}}$ anstelle von $\tau_{\tilde{K},t^*}$ gewählt werden. Die Richtung der Regulierung kann dann nicht aus den Daten bestimmt werden. Anstelle eines gerichteten vollständigen Graphen K geht man dann von einem ungerichteten vollständigen Graphen K_u aus, der also nur eine ungerichtete Kante zwischen jeweils zwei Genen, die den Knoten entsprechen, besitzt. Die Definition signifikanter Kanten bleibt für $\tau = \tau_{\tilde{K}}$ unverändert. Mit derselben Vorgehensweise wie oben erhält man aus K_u den ungerichteten Graphen K_u^*, dessen Unterschied zu K^* darin liegt, dass ebenfalls ungerichtete Kanten statt gerichteter Kanten vorliegen. Um die Richtung der Regulierung zu bestimmen, muss weiteres biologisches Wissen einbezogen werden. Dies kann beispielsweise das Wissen über spezifische Bindungsstellen sein, siehe dazu auch Kapitel 6.

3.2 Formale Begriffsanalyse als Methode zur Komponentenauswahl

Die Methode dieses Kapitelabschnitts wurde zusammen mit Susanne Motameny entwickelt. Hierbei ist die Erstellung des formalen Kontextes mit Mustern der Arbeit von Frau Motameny zuzuordnen, die statistischen Analysen mit den Genen aus den Begriffen bzw. Zwischenbegriffen sowie die Vorverarbeitung der Genlisten sind von mir durchgeführt worden. Im ersten Teil

dieses Abschnitts werden diejenigen Grundlagen der Begriffsanalyse eingeführt, die für das Verständnis der Methode nötig sind, welche im zweiten Teil beschrieben wird.

3.2.1 Grundlagen der formalen Begriffsanalyse

Die Ursprünge der formalen Begriffsanalyse finden sich in Wille (1982), der in seiner ersten Veröffentlichung zur formalen Begriffsanalyse wiederum auf eine Arbeit von Birkhoff (1938) aufbaut. Die formale Begriffsanalyse sei

> 'ein Gebiet der Angewandten Mathematik, das sich auf eine Mathematisierung von Begriff und Begriffshierarchie gründet und damit mathematisches Denken für die begriffliche Datenanalyse und Wissensverarbeitung aktiviert' ((Ganter u. Wille, 1996), Vorwort).

Die folgenden Definitionen stammen ebenfalls aus Ganter u. Wille (1996).

Definition 3.10. *(formaler Kontext, (Ganter u. Wille, 1996))*
Ein formaler Kontext $\mathbb{K} := (G, M, I)$ *besteht aus zwei Mengen* G *und* M *sowie einer Inzidenzrelation* $I \subseteq G \times M$*. Elemente aus* G *heißen Gegenstände, Elemente aus* M *Merkmale von* $\mathbb{K}$*. Gilt* $(g,m) \in I$ *für einen Gegenstand* $g \in G$ *und ein Merkmal* $m \in M$*, so liest man dies als 'der Gegenstand* g *hat das Merkmal* m*'.*

Endliche Kontexte lassen sich daher in einer Kreuzchentabelle darstellen, in der jede Zeile einem Gegenstand und jede Spalte einem Merkmal zugeordnet ist. Es wird genau dann ein Kreuzchen in der Zeile des Gegenstands $g \in G$ und in der Spalte des Merkmals $m \in M$ gemacht, wenn $(g,m) \in I$ gilt. Um eine Ordnung auf diesen Daten zu erhalten, betrachtet man die gemeinsamen Merkmale einer Menge von Gegenständen sowie diejenigen Gegenstände, die sämtliche Merkmale einer Menge von Merkmalen aufweisen. Greift man sich nun eine spezielle Menge $A \subseteq G$ heraus, so listet man die gemeinsamen Merkmale der Elemente aus A in der Menge A' auf, greift man sich eine spezielle Menge $B \in M$ heraus, so betrachtet man diejenigen Gegenstände $B' \subseteq G$, die sämtliche Merkmale aus B aufweisen. (A,B) wird formaler Begriff genannt, falls die Menge A' mit B und die Menge B' mit A übereinstimmt. Ein formaler Begriff ist folgendermaßen definiert:

Definition 3.11. *(formaler Begriff, (Ganter u. Wille, 1996))*
Gegeben sei ein formaler Kontext $\mathbb{K} = (G, M, I)$*. Ein Paar* (A,B) *mit* $A \subseteq G$*,* $B \subseteq M$ *ist genau dann ein formaler Begriff von* $\mathbb{K}$*, wenn* $A' = B$ *und* $B' = A$ *gilt. Hierbei ist für jede Menge* $A \subseteq G$ *die Menge* A' *definiert durch*

$$A' := \{m \in M : (g,m) \in I \text{ für alle } g \in A\}$$

und für jede Menge $B \subseteq M$ *ist die Menge* B' *definiert durch*

$$B' := \{g \in G : (g,m) \in I \text{ für alle } m \in B\}.$$

Für einen formalen Begriff (A,B) heißt A der Umfang und B der Inhalt des Begriffs. $\mathscr{B}(\mathbb{K})$ und $\mathscr{B}(G,M,I)$ bezeichnen die Menge aller Begriffe des Kontextes $\mathbb{K} = (G,M,I)$.

In einer Kreuzchentabelle erhält man alle formalen Begriffe durch das Auffinden von maximalen Rechtecken aus Kreuzchen, wobei Umordnungen von Spalten und Zeilen erlaubt sind.

Lemma 3.12. *(Ganter u. Wille, 1996) Für jedes $A \subseteq G$ läßt sich ein formaler Begriff durch das Paar (A'',A') bilden und für jedes $B \subseteq M$ läßt sich ein formaler Begriff durch das Paar (B',B'') bilden.*

Definition 3.13. *(Gegenstandsbegriff, Merkmalsbegriff, (Ganter u. Wille, 1996))*
Für einen Gegenstand $g \in G$ heißt $\gamma g := (\{g\}'',\{g\}')$ der Gegenstandsbegriff zum Gegenstand g und für ein Merkmal $m \in M$ heißt $\mu m := (\{m\}',\{m\}'')$ der Merkmalsbegriff zum Merkmal m.

Über die Teilmengenbeziehung der Gegenstände (oder über die Teilmengenbeziehung der Merkmale aufgrund der Eigenschaft $A_1 \subseteq A_2 \Leftrightarrow A_2' \subseteq A_1'$) lässt sich eine Ordnung auf der Menge der Begriffe $\mathscr{B}(\mathbb{K})$ einführen:

Definition 3.14. *(Unterbegriff, Oberbegriff, Ordnung der Begriffe, (Ganter u. Wille, 1996))*
Es seien $(A_1,B_1),(A_2,B_2) \in \mathscr{B}(\mathbb{K})$. (A_1,B_1) heißt Unterbegriff von (A_2,B_2) und (A_2,B_2) heißt Oberbegriff von (A_1,B_1), falls $A_1 \subseteq A_2$ gilt. Es wird in diesem Fall $(A_1,B_1) \leq (A_2,B_2)$ geschrieben. Die Relation '$\leq$' wird Ordnung der Begriffe genannt. Die derart geordnete Menge $\mathscr{B}(\mathbb{K})$ wird mit $\underline{\mathscr{B}}(\mathbb{K})$ bezeichnet.

Bemerkung 3.15. *(Ganter u. Wille, 1996) Die partiell geordnete Menge der Begriffe $\underline{\mathscr{B}}(\mathbb{K}) = (\mathscr{B}(\mathbb{K}),\leq)$ ist ein Verband, da zu je zwei Elementen $(A_1,B_1),(A_2,B_2) \in \mathscr{B}(\mathbb{K})$ stets das Supremum $(A_1,B_1) \vee (A_2,B_2)$ und das Infimum $(A_1,B_1) \wedge (A_2,B_2)$ zweier Begriffe existiert. Die im weiteren betrachteten endlichen Verbände sind aufgrund der Endlichkeit auch vollständige Verbände, das heißt, dass zu jeder Teilmenge $X \subseteq \mathscr{B}(\mathbb{K})$ das Supremum $\vee X$ und Infimum $\wedge X$ existiert.*

Aufgrund der Ordnung der Begriffe sind die Begriffe nun in Form eines Liniendiagramms (Hasse-Diagramm) darstellbar. Die Knoten sind die Begriffe, wobei sie durch eine Linie von unten nach oben verbunden werden, wenn der untenstehende Begriff Unterbegriff des obenstehenden Begriffs ist. In den Kapiteln 6 und 7 werden reduzierte Liniendiagramme verwendet, da sonst das Liniendiagramm aufgrund der langen Bezeichnungen der Begriffe sehr unübersichtlich wird. Im reduzierten Liniendiagramm wird jeder Gegenstand $g \in G$ und jedes Merkmal $m \in M$ nur einmal eingetragen, und zwar am zugehörigen Gegenstandsbegriff γg beziehungsweise am zugehörigen Merkmalsbegriff μm. Die Gegenstände des Begriffs sind dann diejenigen Gegenstände, die am Knoten des Begriffs oder an absteigenden Linienzügen liegen. Die Merkmale des Begriffs sind dann diejenigen Merkmale, die am Knoten des Begriffs oder an aufsteigenden

Linienzügen liegen.

In dieser Arbeit wird die formale Begriffsanalyse zum einen zur Visualisierung der Dynamik von genregulatorischen Netzwerken in Kapitel 7 und zum anderen zur Auswahl von Genen für Modelle genregulatorischer Subnetzwerke in diesem Kapitel und in Kapitel 6 verwendet. Darüber hinaus hat die formale Begriffsanalyse jedoch weitere, vielfältige Anwendungsmöglichkeiten, beispielsweise über Ausnutzung von Implikationen (Fakler, 2006).

3.2.2 Komponentenauswahl genregulatorischer Netzwerke

Wie in Kapitel 3.1 sind die Menge aller Gene G, die Quellgene S, die Relation von Interaktionen R und die Genexpressionsmatrix E gegeben. Zunächst wird in einem ersten Schritt eine Vorauswahl von Genen aufgrund der Relation R getroffen. Dann wird ein Begriffsverband aufgrund dieser Gene und der Genexpressionsmatrix E aufgestellt und die Zwischenbegriffe und Begriffe der Quellgene für den letzten Schritt einer statistischen Analyse verwendet.

Vorauswahl der Gene

Als Erstes wird die Menge G mithilfe der Relation R und eines fest gewählten Wertes $K \in \mathbb{N}^+$ verkleinert. Dies bewirkt, dass der Begriffsverband im zweiten Schritt übersichtlicher ist. Kann man jedoch nicht auf R vertrauen, weil eventuell wichtige Interaktionen fehlen, oder ist R nicht vorhanden, so kann dieser Schritt übersprungen werden. Die neue Genliste S_K wird sukzessive über Genlisten S_k, $k = 1,...,K$ aufgrund der Relation R folgendermaßen bestimmt:

$$\begin{aligned} S_1 &:= \{g \in G : \exists s \in S \text{ mit } (s,g) \in R\} \\ S_k &:= S_{k-1} \cup \{g \in G : \exists s \in S_{k-1} \text{ mit } (s,g) \in R\} \text{ für } 2 \leq k \leq K. \end{aligned}$$

Bilden wir einen Interaktionsgraphen aus R und G wie in Kapitel 3.1.1 beschrieben, so können die Genlisten S_k folgendermaßen erläutert werden: Die Menge S_k enthält all diejenigen Gene, die Knoten eines Weges von maximaler Länge k sind, der als Anfangsknoten eines der Quellgene besitzt. S_1 besitzt also nur die Nachbarn der Quellgene im Interaktionsgraphen, S_2 besitzt zusätzlich noch die Nachbarn der Nachbarn der Quellgene und so fort, siehe auch Abbildung 3.2.

Begriffsverband der Genexpressionen

Für den Begriffsverband ist die Erstellung eines formalen Kontextes $\mathbb{K} = (G,M,I)$ notwendig. Abhängig von den Quellgenen werden Muster der Form $\mathscr{M} = (M_1,...,M_T)$ mit $M_i \in \mathbb{N}_0^+$ erzeugt, die interessante Verhaltensweisen bezüglich dieser Quellgene aufweisen. Die Muster haben ebenfalls wie die Genexpressionsmessungen T Zeitpunkte. Beispielsweise kann die Hochregulierung eines Gens in einem Muster dargestellt werden, indem für die ersten Zeitpunkte eine Null angenommen wird und für alle darauf folgenden Zeitpunkte eine Eins, also durch ein Muster der Form $\mathscr{M}_1 = (0,0,...,0,1,1,...,1)^t$. Hochregulierung mit anschließender Herunterregulierung kann dargestellt werden, indem in dem vorherigen Muster für die letzten Zeitpunkte wieder eine Null angenommen wird, also durch ein Muster der Form $\mathscr{M}_2 = (0,0,...,0,1,1,...,1,0,0,...,0)^t$.

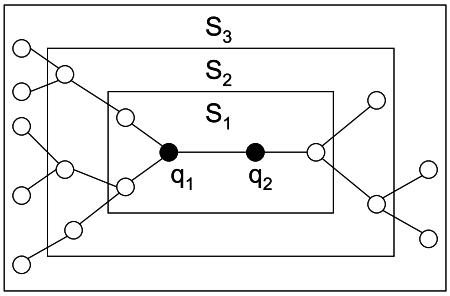


Abbildung 3.2: Die Gene q_1 und q_2 seien Quellgene und sind durch schwarze Kreise im Interaktionsgraph dargestellt. Die Rechtecke in der Abbildung enthalten dann die entsprechenden Genlisten S_1, S_2 und S_3. Dabei ist $S_1 \subseteq S_2 \subseteq S_3$.

Bei sehr wenigen gemessenen Zeitpunkten kann man die Genexpression der Quellgene selbst für die Muster verwenden, bei sehr vielen Zeitpunkten müssen vorher die Werte geglättet werden. Der formale Kontext $\mathbb{K}$ besteht dann aus den Gegenständen $S_{\mathbb{K}} = (g_1, ..., g_r)$ und den Merkmalen $M = (\mathscr{M}_1, ...\mathscr{M}_m)$. Es gilt $(g, \mathscr{M}) \in I$ mit $g \in S_K$ und $\mathscr{M} \in M$ genau dann, wenn g und $\mathscr{M}$ stark korreliert sind, das heißt, dass der Korrelationskoeffizient von der Genexpressionszeitreihe des Gens g und dem Muster $\mathscr{M}$ größer als ein bestimmter Schwellwert $\eta > 0$ ist. Als Korrelationskoeffizienten kommen wiederum die Korrelationsfunktionen in Frage, die zu Beginn des Kapitels vorgestellt wurden, also Definitionen (3.1), (3.2) und (3.3). Für den Kontext $\mathbb{K}$ gilt dann:

$$(g, \mathscr{M}) \in I \Leftrightarrow |\tau(e, \mathscr{M})| > \eta, \tag{3.10}$$

wobei e die Genexpressionszeitreihe von Gen g ist. Der Begriffsverband zu diesem Kontext offenbart die Struktur der Gegenstände und ihrer Merkmale. Der Inhalt eines Begriffs enthält eine maximal mögliche Menge von Genen, die mit denselben Mustern stark korreliert sind.

Auswahl der Gene, statistische Analyse und Modellierung

Der Begriffsverband erlaubt zunächst eine Aussage darüber, welche Gene welchem Muster zugeordnet werden. Im reduzierten Liniendiagramm sind die Muster an den Merkmalsbegriffen eingezeichnet, die Gene an den Gegenstandsbegriffen, siehe auch das Beispiel eines Kontextes mit Begriffsverband in Tabelle 3.1 und Abbildung 3.3.

In der Abbildung 3.3 ist zu sehen, dass der obere linke Begriff aus den Genen g_1, g_3 und g_5 besteht, die man erhält, indem man sämtliche Gene an absteigenden Linienzügen berücksichtigt, sowie aus dem Muster $\mathscr{M}_1$, das an dem Begriff selbst eingezeichnet ist. Der untere rechte Begriff besteht dagegen aus dem am Begriff selbst eingezeichneten Gen g_2 sowie den Mustern $\mathscr{M}_2$, $\mathscr{M}_3$ und $\mathscr{M}_4$, die man erhält, indem man sämtliche Muster an aufsteigenden Linienzügen berücksichtigt. Der oberste Begriff ist das Paar $(\{g_1, ..., g_6\}, \emptyset)$, welches sämtliche Gegenstände als Umfang hat, der unterste Begriff ist das Paar $(\emptyset, \{\mathscr{M}_1, ..., \mathscr{M}_4\})$, welches sämtliche Merkmale als Inhalt hat.

Tabelle 3.1: Kontext mit 6 Genen $g_1,...,g_6$ und 4 Mustern $\mathscr{M}_1,...,\mathscr{M}_4$. Die Inzidenzrelation ist durch eine Kreuzchentabelle dargestellt. Ein 'x' bedeutet, dass der Gegenstand aus der entsprechenden Zeile zu dem Merkmal aus der entsprechenden Spalte inzident ist, ein '-' bedeutet, dass keine Inzidenz vorliegt.

Gen \ Muster	$\mathscr{M}_1$	$\mathscr{M}_2$	$\mathscr{M}_3$	$\mathscr{M}_4$
g_1	x	x	-	-
g_2	-	x	x	x
g_3	x	-	-	-
g_4	-	-	x	-
g_5	x	-	x	-
g_6	-	-	x	x

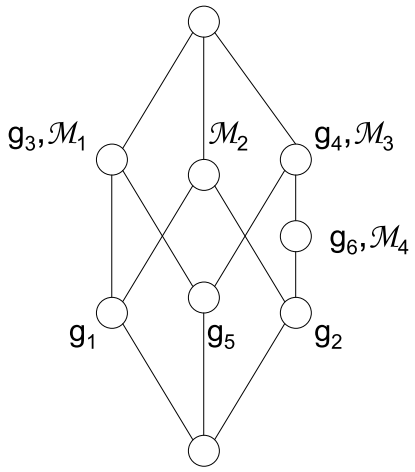


Abbildung 3.3: Begriffsverband zu dem Beispielkontext aus Tabelle 3.1.

Man erhält aus dem Begriffsverband außerdem eine Aussage über Ober- bzw. Unterbegriffe von Begriffen, das heißt wie die Begriffe mit anderen Begriffen zusammenhängen. Der Begriff $(\{g_2,g_6\},\{\mathscr{M}_3,\mathscr{M}_4\})$ ist beispielsweise Oberbegriff zu $(\{g_2\},\{\mathscr{M}_2,\mathscr{M}_3,\mathscr{M}_4\})$, da er eine Obermenge der Gegenstände aus letzterem Begriff umfasst und dadurch auch weniger Mustern zugeordnet ist. Andererseits ist $(\{g_2,g_6\},\{\mathscr{M}_3,\mathscr{M}_4\})$ Unterbegriff zu $(\{g_2,g_4,g_5,g_6\},\{\mathscr{M}_3\})$, da er nur eine Teilmenge der Gegenstände des letzten Begriffs umfasst, dadurch jedoch auch mehr Eigenschaften besitzt, also mehr Mustern zugeordnet ist. Andererseits wird Gen g_5 auch dem Muster $\mathscr{M}_1$ zugeordnet. Dies ist eine Eigenschaft, die weder Gen g_2 noch Gen g_4 haben, auch wenn sie zusammen mit g_5 in einem Begriff sein können. Bilden nun g_2 und g_4 die Quellgene S, so möchte man g_4 nicht in die Menge S' aufnehmen, also in die Menge der Gene, die mit dem zuvor beschriebenen statistischen Vorgehen weiter analysiert werden. Um zu beschreiben, welche Gene in S' aufgenommen werden sollen, führen wir hier die Definition eines Zwischenbegriffs ein:

Definition 3.16. *(Zwischenbegriff)*
In einem Begriffsverband nennen wir einen Begriff (A,B) Zwischenbegriff von zwei Begriffen (A_1,B_1), (A_2,B_2), wenn (A_1,B_1) Oberbegriff oder Unterbegriff von (A_2,B_2) ist und zusätzlich (A,B) an einem absteigenden oder aufsteigenden Linienzug von (A_1,B_1) nach (A_2,B_2) liegt. Der Begriff darf auch am Anfang oder Ende des Linienzugs liegen, dass heißt, (A,B) ist auch dann Zwischenbegriff von (A_1,B_1), (A_2,B_2), wenn $(A,B) = (A_1,B_1)$ oder $(A,B) = (A_2,B_2)$ gilt und (A_1,B_1) Ober- oder Unterbegriff von (A_2,B_2) ist.

Die Genmenge S' wird nun folgendermaßen bestimmt:

Bemerkung 3.17. *Es werden genau diejenigen Gene $g_i \in G \setminus S$ in die Menge S' aufgenommen, deren Gegenstandsbegriff γg_i mit einem der Gegenstandsbegriffe γg_j mit $g_j \in S$ zusammenfällt oder deren Gegenstandsbegriff γg_i ein Zwischenbegriff der Gegenstandsbegriffe γg_j mit $g_j \in S$ ist.*

Ist im Beispiel $S = \{g_2, g_4\}$, so besteht die weiter zu analysierende Menge S' nur aus dem einzigen Gen g_6, also $S' = \{g_6\}$. Die statistische Analyse erfolgt analog zu Kapitel 3.1.2 und ergibt die gesuchte Genmenge $\tilde{S}$.

Die hier vorgestellte Methode hat im Gegensatz zu dem graphentheoretischen Ansatz den Vorteil, dass der Begriffsverband bei geeigneter Wahl der Muster $\mathscr{M}$ einen genaueren Überblick über die korrelationsbedingte Struktur der verschiedenen Expressionszeitreihen gibt. Biologisches Wissen kann hier einfließen, indem im Beispiel möglicherweise $\mathscr{M}_2$ zu einem biologisch sehr wichtigen Muster erklärt wird. Dann kann im Begriffsverband direkt abgelesen werden, dass zusätzlich zu Gen g_6 noch Gen g_1 in der weiteren Analyse berücksichtigt werden sollte. Außerdem ist es möglich, in größeren Begriffsverbänden 'Stränge' herauszufinden, also wiederum relativ abgetrennte Begriffsverbände im Gesamtbegriffsverband, deren Gene starke Ähnlichkeiten in ihren zugeordneten Merkmalen haben, jedoch zu den Genen des übrigen Verbandes wenig gemeinsame Merkmale aufweisen. Solche 'Stränge' erhält man aus der Kreuzchentabelle, wenn durch Umordnung der Spalten und Zeilen eine Kreuzchentabelle mit der Form $A = diag(A_1,...,A_l)$ entsteht. Für weitere Analysemöglichkeiten siehe Ganter u. Wille (1996). Beide Methoden haben den weiteren Vorteil, dass die Verwendung von Interaktionsdaten möglich, jedoch die Existenz solcher Daten nicht Voraussetzung für die Durchführung der Methode ist. Üblicherweise sind Interaktionsdaten nur in den seltensten Fällen für das gesamte Genom bekannt. Sind daher Analysen ohne Einbezug der Interaktionsdaten notwendig, so kann in der ersten Methode ein vollständiger Graph als Interaktionsgraph gewählt werden, in der zweiten Methode wird die gesamte Genliste als Gegenstandsmenge zur Erstellung des Begriffsverbandes gewählt. Eine Anwendung beider hier vorgestellter Methoden ist in Kapitel 6 zu finden, in dem Experimente an *M. tuberculosis* analysiert werden.

Kapitel 4

Parameterschätzung

Nach dem in Kapitel 2 eingeführten Modell für genregulatorische Netzwerke kann die Dynamik solch eines Netzwerkes beschrieben werden durch

$$\dot{x}(t) \quad = \quad A_q x(t) + b_q \tag{4.1}$$

mit $x(t) = (x_1(t), ..., x_n(t))^t \in \mathbf{R}^n$, Matrizen $A_q = (a_{q_{ij}})_{i,j=1,...,n} \in \mathbf{R}^{n\times n}$ und Vektoren $b_q = (b_{q_1}, ..., b_{q_n})^t \in \mathbf{R}^n$, $q = 1, ..., \xi$, siehe Gleichung (2.35). Gegeben seien Messdaten $\hat{x}(t)$ zu den Zeitpunkten $t = 0, ..., l$. Dann kann jede Zeile des Systems einzeln beschrieben werden

$$\dot{x}_i(t) \quad = \quad \sum_{j=1}^{n} a_{q_{ij}} x_j(t) + b_{q_i}. \tag{4.2}$$

Die Parameter $a_{q_{ij}} \in \mathbf{R}$ und $b_{q_i} \in \mathbf{R}$ sollen aus den Messdaten geschätzt werden. In diesem Kapitel werden zwei Lösungsmethoden vorgestellt, die abhängig davon sind, ob Fragestellung 1.3 oder 1.4 gelöst werden muss - also, ob die zugrunde liegende Struktur des genregulatorischen Netzwerkes bekannt ist. Im ersten Teil dieses Kapitels wird zunächst der Differenzialquotient geschätzt. Im zweiten Teil werden die Parameter des Systems bestimmt. Dabei wird in Kapitel 4.2 eine detaillierte Kenntnis der biochemischen Prozesse im betrachteten System vorausgesetzt, also insbesondere wird die Struktur des genregulatorischen Netzwerkes benötigt. Wir nehmen an, dass wir zur Schätzung der Parameter der Gleichung (4.2) die $l+1$ Messpaare $(\hat{x}(t_0), y_i(t_0)), ..., (\hat{x}(t_l), y_i(t_l)) \in \mathbf{R}^{n+1}$ gegeben haben, wobei $y_i(t_k) := \hat{\dot{x}}_i(t_k)$ sei.

4.1 Bestimmung der y-Werte durch Schätzung des Differenzialquotienten

Wir wollen zunächst aus den Messwerten $x_i(t_0), ..., x_i(t_l)$ die entsprechenden Ableitungen nach der Zeit $y_i(t_k) = \hat{\dot{x}}_i(t_k)$ schätzen. Eine sehr einfache Art der y-Wertebestimmung geschieht über Differenzenquotienten, siehe Gebert u. a. (2006):

$$y_i(t_k) = \hat{\dot{x}}_i(t_k) \quad = \quad \begin{cases} \frac{\hat{x}_i(t_{k+1}) - \hat{x}_i(t_k)}{t_{k+1} - t_k} & \text{für } k = 0, ..., l-1 \\ \\ \frac{\hat{x}_i(t_l) - \hat{x}_i(t_{l-1})}{t_l - t_{l-1}} & \text{für } k = l. \end{cases} \tag{4.3}$$

In der praktischen Anwendung - siehe Kapitel 5 und 6 - erweist sich diese Herangehensweise jedoch als zu stark vereinfacht. Das Problem liegt beispielsweise in Kapitel 5 darin, dass in den sehr kurzen Zeitreihen drei Extrema in kurzem zeitlichen Abstand aufeinanderfolgen. In Abbildung 4.1 (links) wird solch ein Beispiel verdeutlicht. Hier ist zu erkennen, dass die

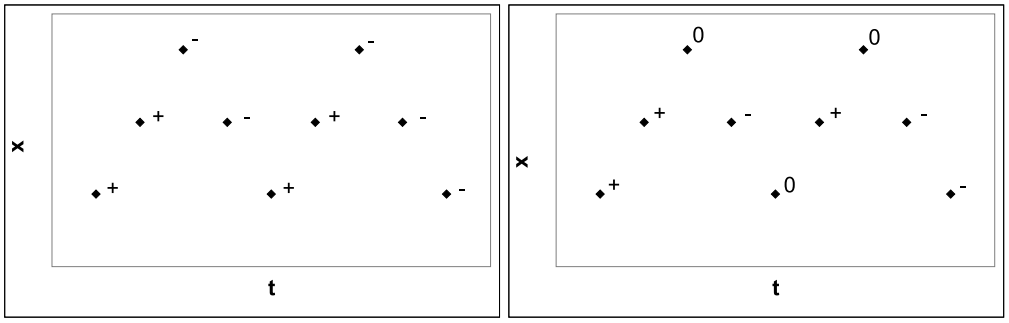


Abbildung 4.1: Beispiel für zwei mögliche Schätzungen der y-Werte, '+' steht hierbei für $y > 0$, '-' für $y < 0$ und '0' für $y = 0$. Links ist eine Schätzung der Differenzialquotienten mit Differenzenquotienten zu sehen, rechts wird eine quadratische Regression verwendet.

Ableitungen in sämtlichen Extremalpunkten größer oder kleiner als Null sind, wenn man nach (4.3) abschätzt. Schätzt man hingegen die y-Werte über quadratische Regression, so erhält man in Beispiel 4.1 (rechts) bessere Schätzungen des Differenzialquotienten in den Extremalpunkten. Bei diesem Vorgehen wird zunächst eine quadratische Regression für die $\hat{x}$-Werte von drei aufeinanderfolgenden Zeitpunkten t_{k-1}, t_k, t_{k+1} durchgeführt und das Polynom anschließend bezüglich t differenziert und an der Stelle t_k, $k = 1, ..., l-1$ ausgewertet. Das Polynom, das durch $\hat{x}_i(t_{k-1}), \hat{x}_i(t_k)$ und $\hat{x}_i(t_{k+1})$, $k = 1, ..., l-1$, gelegt wird, erhält man durch

$$\begin{aligned} p(t) &= \hat{x}_i(t_{k-1}) \frac{(t-t_k)(t-t_{k+1})}{(t_{k-1}-t_k)(t_{k-1}-t_{k+1})} \\ &\quad + \hat{x}_i(t_k) \frac{(t-t_{k-1})(t-t_{k+1})}{(t_k-t_{k-1})(t_k-t_{k+1})} \\ &\quad + \hat{x}_i(t_{k+1}) \frac{(t-t_{k-1})(t-t_k)}{(t_{k+1}-t_{k-1})(t_{k+1}-t_k)}. \end{aligned} \tag{4.4}$$

Differenziert man dieses Polynom nach t, so erhält man

$$\begin{aligned} \frac{\mathrm{d}p(t)}{\mathrm{d}t} &= \frac{\hat{x}_i(t_{k-1})}{(t_{k-1}-t_k)(t_{k-1}-t_{k+1})}(2t - t_k - t_{k+1}) \\ &\quad + \frac{\hat{x}_i(t_k)}{(t_k-t_{k-1})(t_k-t_{k+1})}(2t - t_{k-1} - t_{k+1}) \\ &\quad + \frac{\hat{x}_i(t_{k+1})}{(t_{k+1}-t_{k-1})(t_{k+1}-t_k)}(2t - t_{k-1} - t_k). \end{aligned} \tag{4.5}$$

Einsetzen von (t_k) ergibt die Schätzungen für $y_i(t_k)$:

$$\begin{aligned} y_i(t_k) = \hat{\dot{x}}_i(t_k) = \frac{\mathrm{d}p}{\mathrm{d}t}(t_k) &= \frac{\hat{x}_i(t_{k-1})}{(t_{k-1}-t_k)(t_{k-1}-t_{k+1})} \cdot (t_k - t_{k+1}) \\ &+ \frac{\hat{x}_i(t_k)}{(t_k-t_{k-1})(t_k-t_{k+1})} \cdot (2t_k - t_{k-1} - t_{k+1}) \\ &+ \frac{\hat{x}_i(t_{k+1})}{(t_{k+1}-t_{k-1})(t_{k+1}-t_k)} \cdot (t_k - t_{k-1}) \\ &\text{für } k = 1,...,l-1. \end{aligned} \tag{4.6}$$

Die Randpunkte $y_i(t_0)$ und $y_i(t_l)$ werden hierbei folgendermaßen bestimmt: Für $k = 0$ bzw. $k = l$ wird dieselbe quadratische Regression wie für $k = 1$ bzw. $k = l-1$ verwendet, jedoch an der Stelle t_0 bzw. t_l nach t differenziert, das heißt

$$\begin{aligned} y_i(t_0) &= \frac{\hat{x}_i(t_0)}{(t_0-t_1)(t_0-t_2)}(2t_0 - t_1 - t_2) \\ &+ \frac{\hat{x}_i(t_1)}{(t_1-t_0)(t_1-t_2)}(t_0 - t_2) \\ &+ \frac{\hat{x}_i(t_2)}{(t_2-t_0)(t_2-t_1)}(t_0 - t_1) \end{aligned} \tag{4.7}$$

und

$$\begin{aligned} y_i(t_l) &= \frac{\hat{x}_i(t_{l-2})}{(t_{l-2}-t_{l-1})(t_{l-2}-t_l)}(t_l - t_{l-1}) \\ &+ \frac{\hat{x}_i(t_{l-1})}{(t_{l-1}-t_{l-2})(t_{l-1}-t_l)}(t_l - t_{l-2}) \\ &+ \frac{\hat{x}_i(t_l)}{(t_l-t_{l-2})(t_l-t_{l-1})}(2t_l - t_{l-2} - t_{l-1}). \end{aligned} \tag{4.8}$$

4.2 Methode der kleinsten Quadrate mit linearen Nebenbedingungen

Nachdem zunächst die Ableitungen y_i für jeden Zeitpunkt und jedes Gen $i = 1,...,n$ bestimmt wurden, muss anschließend eine Zuordnung der Daten $(\hat{x}(t_0), y_i(t_0)),...,(\hat{x}(t_l), y_i(t_l))$ zu den verschiedenen Bereichen im Zustandsraum erfolgen. Wir gehen hier in diesem Unterkapitel von Problemstellung 1.4 aus, so dass bekannt ist, welches Gen durch welche anderen Gene reguliert wird. Im Idealfall sind auch die Schwellwerte der Regulierungsfunktionen bekannt, so dass die Einteilung des Zustandsraumes bereits gegeben ist. Andernfalls müssen diese Schwellwerte ebenfalls über die Messdaten geschätzt werden. Wie solch eine Aufteilung des Zustandsraums im speziellen Fall gefunden werden kann, ist in Kapitel 5.4.2 beschrieben. Im Folgenden betrachten wir also eine Teilmenge $(\hat{x}(t_{k_1}), y_i(t_{k_1})),...,(\hat{x}(t_{k_s}), y_i(t_{k_s}))$, $\{k_1,...,k_s\} \subseteq \{1,...,l\}$ der

Messdaten, die einer einzigen linearen Differenzialgleichung zugeordnet werden kann. Wir betrachten nun also ein festes i. Um die Notation zu vereinfachen, wählen wir die zwei neuen Notationen $X = (x_{lj}) \in \mathbf{R}^{s\times(n+1)}$ und $y \in \mathbf{R}^s$, wobei $x_{lj} = \hat{x}_l(t_{k_{j-1}})$ für $l = 1,...,s$, $j = 2,...,n+1$, $x_{l1} = 1$ für $l = 1,...,s$ und $y_l = y_i(t_{k_j})$ für $l = 1,...,s$, also

$$X = \begin{pmatrix} 1 & \hat{x}_1(t_{k_1}) & \cdots & \hat{x}_n(t_{k_1}) \\ 1 & \hat{x}_1(t_{k_2}) & \cdots & \hat{x}_n(t_{k_2}) \\ \vdots & \vdots & & \vdots \\ 1 & \hat{x}_1(t_{k_s}) & \cdots & \hat{x}_n(t_{k_s}) \end{pmatrix} \tag{4.9}$$

gilt. Die erste Spalte der Matrix X wird im Modell für die Schätzung des Parameters b_q benötigt. Außerdem gehen wir zunächst davon aus, dass 'genügend' Messdaten vorhanden sind, dass also $n+1 \leq s$ gilt, und dass die Matrix X den maximal möglichen Rang $n+1$ habe.

4.2.1 Beschreibung des Verfahrens

Wir legen ein Gauss-Markov-Modell (siehe Faigle u. a. (2002))

$$y = Xa + \varepsilon \tag{4.10}$$

zugrunde, so dass $y \in \mathbf{R}^s$ nicht nur linear abhängig ist von dem Regressor $a \in \mathbf{R}^{n+1}$ durch die Matrix $X \in \mathbf{R}^{s\times(n+1)}$, sondern auch von einem Fehler, der durch den Zufallsvektor $\varepsilon = (\varepsilon_1,...,\varepsilon_s)^t$ beschrieben wird. Für ε nehmen wir $E(\varepsilon) = 0$ und $E(\varepsilon\varepsilon^t) = \sigma^2 I_s$ an[1], dass also der Erwartungswert jeder Zufallsvariablen ε_i gleich Null ist, die Fehler unkorreliert sind und jeweils dieselbe Varianz $\sigma^2 \geq 0$ haben. Zur Schätzung von a für das Gauss-Markov-Modell (4.10) verwenden wir den Kleinste-Quadrate-Schätzer, der gegeben ist durch

$$\hat{a} = (X^tX)^{-1}X^ty. \tag{4.11}$$

Der Satz von Gauss-Markov besagt, dass dieser Schätzer zum einen erwartungstreu ist, das heißt, der Erwartungswert $E(\hat{a})$ des Schätzers ist gleich dem zu schätzenden Parameter a, zum anderen ist er effizient, das heißt, es gibt in der Klasse der linearen, erwartungstreuen Schätzer keinen Schätzer mit einer kleineren Varianz.

Der Name 'Kleinste-Quadrate' stammt daher, dass

$$\hat{a} = \arg\min_{a\in\mathbf{R}^{n+1}} \sum_{l=1}^{s} \left(y_l - a_1 - \sum_{k=2}^{n+1} a_k x_{lk} \right)^2 = \arg\min_{a\in\mathbf{R}^{n+1}} (y-Xa)^t(y-Xa) \tag{4.12}$$

[1] Wir verwenden in diesem Zusammenhang die Bezeichnung I_s für die $(s\times s)$-Einheitsmatrix, da im Fall der sonst üblichen Bezeichnung E_s Verwechslungsgefahr mit dem Erwartungswert E besteht.

gilt, dass also $\hat{a}$ die Restsumme von Abweichungsquadraten minimiert. Über die Minimierung dieser Summe lässt sich der Kleinste-Quadrate-Schätzer folgendermaßen berechnen: Zur Minimierung wird $g(a) = (y - Xa)^t(y - Xy)$ nach a differenziert und gleich Null gesetzt, also

$$\frac{\mathrm{d}g(a)}{\mathrm{d}a} = 2X^tXa - 2X^ty = 0 \tag{4.13}$$

$$\Leftrightarrow \quad 2X^tXa = 2X^ty. \tag{4.14}$$

Da X Rang $n+1$ hat, existiert $(X^tX)^{-1}$ und Gleichung (4.13) löst sich nach a auf zu dem Schätzer (4.11).

4.2.2 Parameterschätzung mit Nebenbedingungen

Minimiert man die Restsumme der Abweichungsquadrate, so kann zusätzlich der Lösungsraum, also der Vektorraum $\mathbf{R}^{n+1}$, aus dem bisher die Vektoren a zur Minimierung der Summe gewählt wurden, eingeschränkt werden. Dies ist vorteilhaft, wenn biologisches Wissen verfügbar ist oder wenn einige Datenpunkte 'vertrauenswürdiger' sind als andere. Dieser zweite Aspekt kann entweder durch unterschiedliche Gewichtung der Datenpunkte geschehen, das heißt, man minimiert die Summe

$$\sum_{l=1}^{s} \left(m_l \left(y_l - a_1 - \sum_{k=2}^{n+1} a_k x_{lk} \right) \right)^2 \tag{4.15}$$

mit entsprechenden Gewichten $m_l \in \mathbf{R}^+$, $l = 1,...,s$, oder man möchte diese 'vertrauenswürdigen' Datenpunkte exakt interpolieren. Dies ist beispielsweise dann der Fall, wenn Experimente gemacht werden, in denen ein Bakterium einem Medikament oder anderen Nährstoffen bzw. einer anderen Umgebung ausgesetzt wird. Kann man zu Beginn des Experiments davon ausgehen, dass sich das System in einem Fixpunkt befindet und gegen Ende des Experiments, dass sich das System wiederum in einem Fixpunkt befindet, so kann man die Anfangs- und Endmessung zur Reduktion der Variablen verwenden. Es sei $(x(t_0), y_i(t_0)) \in \mathbf{R}^{n+1}$ die Messung zu Beginn des Experiments und $(x(t_l), y_i(t_l)) \in \mathbf{R}^{n+1}$ die Messung gegen Ende des Experiments. Können wir von Fixpunkten zu Beginn und Ende des Experiments ausgehen und werden diese beiden Fixpunkte beschrieben durch die Differenzialgleichungen $\dot{x}_i(t) = a_1 + a_2x_1(t) + ... + a_{n+1}x_n(t)$ bzw. $\dot{x}_i(t) = a'_1 + a'_2x_1(t) + ... + a'_{n+1}x_n(t)$ mit $a_i, a'_i \in \mathbf{R}$ für $i = 1,...,n+1$, so kann jeweils eine Nebenbedingung für die Optimierung aufgestellt werden:

$$\min_{a \in \mathbf{R}^{n+1}} \sum_{l=1}^{s} (y_l - a_1 - \sum_{k=2}^{n+1} a_k x_{lk})^2, \tag{4.16}$$

so dass

$$y_i(t_0) = a_1 + a_2x_1(t) + ... + a_{n+1}x_n(t)$$

und analog für den Zeitpunkt t_l:

$$\min_{a' \in \mathbf{R}^{n+1}} \sum_{l=1}^{s} (y_l - a'_1 - \sum_{k=2}^{n+1} a'_k x_{lk})^2, \tag{4.17}$$

so dass

$$y_i(t_l) = a'_1 + a'_2x_1(t) + ... + a'_{n+1}x_n(t).$$

Wir können jedoch auch unser bereits vorhandenes Wissen über die biologischen Prozesse integrieren, indem die parametrisierten Regulierungsfunktionen direkt geschätzt werden. Wir gehen weiterhin davon aus, dass die Struktur des genregulatorischen Netzwerkes bekannt ist, das heißt, wir wissen, welche Gene $j_{1_i},...,j_{u_i}$ das Gen i positiv regulieren und welche Gene $j_{u_i+1},...,j_{v_i}$ das Gen i negativ regulieren. Die Gleichungen für Gen i lauten dann

$$\begin{aligned} y(t_0) &= c_i - \gamma_i x_i(t_0) + rf_{i,j_{1_i}}(x_{j_{1_i}}(t_0)) + ... + rf_{i,j_{v_i}}(x_{j_{v_i}}(t_0)) \\ &\vdots \\ y(t_l) &= c_i - \gamma_i x_i(t_l) + rf_{i,j_{i,1}}(x_{j_{1_i}}(t_l)) + ... + rf_{i,j_{v_i}}(x_{j_{v_i}}(t_l)). \end{aligned} \tag{4.18}$$

Diese Gleichungen können wieder in die Form $A\vartheta - b$ transformiert werden, so dass der Parametervektor $\vartheta = (c_1, \gamma_1, k_{1j_{1_1}}, ..., k_{1j_{v_1}}, ...c_n, \gamma_n, k_{1j_{1_n}}, ..., k_{1j_{v_n}})^t \in \mathbf{R}^z$ mit $z = 2n + \sum_{p=1}^{n} v_p$ geschätzt werden muss. Gilt für einen Zeitpunkt $\tilde{t}$, dass für Gen 1 sämtliche Regulierungsfunktionen über den zweiten Schwellwerten liegen, so erhält man folgende Zeile $(a_{w,1},...,a_{w,z})$ der Matrix A:

$$a_{w,1} = 1, a_{w,2} = -x_i(\tilde{t}), a_{w,3} = 1, \cdots, a_{w,u_i+2} = 1, a_{w,u_i+3} = -1, \cdots, a_{w,u_i+v_i+2} = -1 \tag{4.19}$$

und alle übrigen Einträge dieser Zeile sind gleich Null. Außerdem erhält man den Eintrag $y_1(\tilde{t})$ für den w-ten Eintrag des Vektors b. Entsprechend erhält man die Zeilen der Matrix für weitere Bereiche im Zustandsraum. Für eine Regulierung von j nach i unterhalb des Schwellwertes erhält man eine Null in der Matrix, für eine Regulierung zwischen zwei Schwellwerten den Wert $(x_j(\tilde{t}) - \theta_{ij1})/(\theta_{ij2} - \theta_{ij1})$.
Bei dieser Formulierung des Minimierungsproblems lässt sich das biologische Wissen integrieren, indem man Grenzen für die Parameter setzt. Für alle genregulatorischen Netzwerke muss gelten, dass $\vartheta \geq 0$ ist. Außerdem kann es sein, dass für einige Parameter auch eine obere Grenze bekannt ist. Des Weiteren bezieht sich die Integration von Wissen auf Relationen der Parameter untereinander. Beispielsweise kann bekannt sein, dass eine Syntheserate c_i größer als die Syntheserate c_j ist, jedoch beim Abbau $\gamma_j \leq \gamma_i$ gilt. Zusammen mit den Fixpunktgleichungen, die in der Form $G\vartheta = f$ beschrieben werden können, ergibt sich folgendes Optimierungsproblem:

$$\begin{aligned} &\min_{\vartheta \in \mathbf{R}^z} ||A\vartheta - b||^2 \\ &\text{so dass} \\ &C\vartheta \geq d \\ &\text{und } G\vartheta = f \\ &\text{und } \vartheta \geq 0 \end{aligned} \tag{4.20}$$

mit $A, C, G \in \mathbf{R}^{(l+1)n \times z}$ und $b, d, f \in \mathbf{R}^z$. Solch ein quadratisches Minimierungsproblem mit linearen Nebenbedingungen kann beispielsweise mit Aktive-Mengen-Strategien oder Innere-Punkt-Verfahren gelöst werden, siehe hierzu Faigle u. a. (2002); Lawson u. Hanson (1974). In später folgenden Anwendungen werden wir uns auf die in (4.20) aufgestellten Nebenbedingungen beschränken, also auf

- die Beschreibung der Fixpunkte in Gleichungen $G\vartheta = f$,

- die Beschreibung der Relationen zwischen den Parametern durch $C\vartheta \geq d$ und
- die Beschränkung der Parameter durch $\vartheta \geq 0$.

Ist die Struktur des Netzwerkes nicht gegeben, so kann eine andere Einschränkung des Lösungsraums $\mathbf{R}^{n+1}$ bei der Minimierung der Funktion (4.12) dadurch vorgenommen werden, dass der Eingangsgrad der Knoten beschränkt wird. Ist also $||y - Xa||^2$ zu minimieren, so können zusätzliche Variablen

$$\rho_k = \begin{cases} 0 & \text{für } a_k = 0 \\ 1 & \text{für } a_k \neq 0 \end{cases} \quad \text{für } k = 2, ..., n+1 \tag{4.21}$$

dazu verwendet werden, den Eingangsgrad durch $\sum_{k=1}^{n} \rho_k \leq K_{\max}$ mit einem vorher bestimmten maximalen Eingangsgrad $K_{\max}$ zu beschränken. Da dieser Eingangsgrad schwierig zu bestimmen ist, wird diese Einschränkung des Lösungsraums in den folgenden Kapiteln nicht weiter verfolgt, kann jedoch bei entsprechend vorhandenem Wissen bei speziellen Netzwerken angewendet werden.

4.3 Diskussion

Die Methode der kleinsten Quadrate mit linearen Nebenbedingungen hat den Nachteil, dass bereits eine Zuordnung der Messdaten zu den Differenzialgleichungen erfolgt sein muss. Dies ist zwar bei kleineren Netzwerken durchaus möglich (siehe Kapitel 5 und 6), ist jedoch nicht allgemein anwendbar. Geht man von der allgemeinen Gleichung $\dot{x} = Ax + b$ aus und möchte die Stetigkeit der rechten Seite des Differenzialgleichungsmodells wahren, müssen eventuell weitere lineare Gleichungen für manche der Bereiche des Zustandsraums eingeführt werden, in denen keine Messdaten liegen. Weiterhin gehen wir davon aus, dass die x-Werte ohne Fehler gemessen wurden und nur die y-Werte fehlerbehaftet sind.
Geht man dagegen direkt von der Gleichung $\dot{x}_i = c_i - \gamma_i x_i + \sum r f(x_j)$ aus, so bewahrt man mit jeglichen Parametern die Stetigkeit der rechten Seite und hat die Möglichkeit, sämtliches biologisches Wissen über Parameter als Nebenbedingungen einzubauen.
Der Vorteil dieser Methode besteht also insbesondere darin, dass sämtliches Wissen - über Abbauraten, Fixpunkte etc. - in die Parameterschätzung einfließen kann.
Soll jedoch Problem 1.3 gelöst werden, für das weder die Struktur des genregulatorischen Netzwerkes noch die Einteilung des Zustandsraums gegeben ist, sind wesentlich mehr Daten erforderlich als für die Methode aus Kapitel 4.2. Für diese Problemstellung kann eine Verallgemeinerung der Methode der kleinsten Quadrate auf mehrere Sektionen beispielsweise durch Splines oder durch Multiphasenregression (Seber u. Wild, 1989; Lerman, 1980) verwendet werden. Eine Methode, die in den Bereich der Multiphasenregression eingeordnet werden kann, ist von Breiman (1993) eingeführt worden und beruht auf Scharnierhyperebenen (siehe auch Pucar u. Sjöberg (1998)). Es beruht darauf, zunächst zwei Hyperebenen zur Beschreibung der Daten zu verwenden und dann iterativ weitere Scharniere in die bereits vorhandenen Hyperebenen einzufügen. Die Frage ist jedoch, wieviele Hyperebenen gewählt werden sollen. Schließlich wird sich der quadratische Fehler immer weiter verringern, je mehr Hyperebenen man zulässt. Ein weiteres Problem bei solchen Verfahren sind die großen Datenmengen, die benötigt werden, wenn die

Einteilung des Zustandsraum ebenso wie die Schätzung der Parameter in den einzelnen Bereichen gefordert ist. Um mit einer relativ geringen Anzahl von Datenpunkten Resultate erzielen zu können, müssen alternativ wiederum Annahmen über die Struktur gemacht werden, so dass man ein niedrigdimensionaleres Problem erhält. Ist Wissen über die Regulatoren eines Gens vorhanden, so kann eine Teilmenge der Variablen in dem Scharnierhyperebenenverfahren zugrunde gelegt werden. Andere Methoden machen hierzu unter anderem Annahmen über die Anzahl der Regulierungen. In Radde u. Kaderali (2007) wird beispielsweise ein Bayessches Lernverfahren verwendet, in dem a priori Verteilungen über die Regulierungsstärken angenommen werden.
In den beiden Anwendungen aus Kapitel 5 und 6 wird aufgrund der geringen Datenmenge auf das in diesem Kapitel 4.2 beschriebene Verfahren zurückgegriffen und Schwellwerte als bekannt vorausgesetzt.

Kapitel 5

Anwendung des Modells auf die Stickstoffaufnahme von *Corynebacterium glutamicum*

In diesem Kapitel wird die Regulierung der Stickstoffaufnahme in *Corynebacterium glutamicum* modelliert und simuliert. Nach einer Einführung in die relevanten biologischen Vorgänge werden die Differenzialgleichungen aufgestellt, die Parameter aus experimentellen Daten geschätzt und Simulationen für verschiedene Anfangsbedingungen diskutiert. Unter anderem wurde der Fragestellung nachgegangen, wie der Zusammenhang zwischen Proteasenaktivitäten und der Konzentration eines der involvierten Proteine aussieht. Des Weiteren wurde auch das Verhalten von Mutanten simuliert. Die Ergebnisse dieses Kapitels werden in Gebert u. a. (2007a) veröffentlicht.

Stickstoff ist eine wesentliche Komponente, um die Lebensfähigkeit einer Zelle zu sichern, da für den Aufbau fast aller Makromoleküle in einer Zelle - wie Proteine, Nukleinsäuren und Zellwandkomponenten - Stickstoff benötigt wird. Um auch im Fall eines niedrigen Stickstoffgehalts in der Umgebung für eine optimale Versorgung mit Stickstoff zu sorgen, haben viele Bakterien gute Kontrollmechanismen entwickelt. Da diese Mechanismen von Organismus zu Organismus variieren können, ist es nötig, einen Organismus für die Modellbildung auszuwählen. Prof. Dr. Andreas Burkovski konnte als Kooperationspartner gewonnen werden, der seit über zehn Jahren die Vorgänge der Stickstoffaufnahme in dem Bakterium *C. glutamicum* untersucht und somit über ein tiefgehendes Wissen in diesem Bereich verfügt, siehe Burkovski (2003a,b, 2005); Jakoby u. a. (2000); Nolden u. a. (2001); Siewe u. a. (1996). Neben zahlreichen Gesprächen über die biologischen Prozesse gab es den weiteren Vorteil, dass experimentelle Daten vorhanden waren. Aufgrund dieser Gegebenheiten fiel die Wahl des Organismus auf das einzellige Bakterium *C. glutamicum*.
Das Gram-positive Bodenbakterium, das 1957 erstmals isoliert wurde, ist von großer Bedeutung für die industrielle Aminosäureproduktion, z. B. werden zur Zeit über 1.500.000 Tonnen L-Glutamat und über 560.000 Tonnen L-Lysin pro Jahr mithilfe verschiedener *C. glutamicum*-Stämme hergestellt (Burkovski, 2003a). Neben Ammonium kann *C. glutamicum* eine Vielzahl anderer Stoffe als Stickstoffquelle verwerten, zum Beispiel Allantoin, Glutamat, Glutamin, Ornithin und Harnstoff. Unter Verwendung dieser Substrate bewerkstelligt es das Bakterium, die Zellvermehrung mit nahezu gleicher Wachstumsrate aufrecht zu erhalten (Burkovski, 2003b).

Im Folgenden wird die Regulierung der Stickstoffaufnahme von *C. glutamicum* beschrieben. Der Schwerpunkt liegt auf der Modellierung und Simulation für Übergänge zwischen Stickstoffüberschuss und Stickstoffmangel in der Umgebung des Bakteriums. Außerdem wird auch auf die Fragestellung eingegangen, Proteasenaktivitäten aufgrund der Konzentration des Proteins GlnK zu schätzen.
Die Regulierungsmechanismen der Stickstoffaufnahme sind vielfältig und ereignen sich nicht nur auf genregulatorischer Ebene. Zu den weiteren Mechanismen zählen posttranslationale Modifizierungen, die erst nach erfolgter Proteinsynthese stattfinden, Protein-Protein-Interaktionen und Regulierung des Abbaus bestimmter Proteine. Somit sind neben den Genregulationen, die über das Grundmodell beschrieben werden können, Erweiterungen des Modells für jede darüber hinausgehende Regulierung notwendig.

5.1 Erweiterungen auf allgemeinere biochemische Netzwerke

Das bisherige Modell wurde für genregulatorische Prozesse gebildet. Ihm unterliegt die Annahme, dass die Konzentration des Proteins proportional zur Konzentration der mRNA sei. Im Folgenden sollen auch Proteinkonzentrationen in das Modell eingebunden werden. Den Graphen mit den Makromolekülen als Knoten und den Interaktionen zwischen diesen Makromolekülen als Kanten bezeichnen wir als biochemisches Netzwerk. Wir sind nun daran interessiert, verschiedene, bisher vernachlässigte Interaktionsmöglichkeiten in das Grundmodell aus Kapitel 2 mit einzubeziehen. Sind $x_1(t),...,x_n(t)$ die Konzentrationen der mRNA zum Zeitpunkt t und $y_1(t),...,y_n(t)$ die entsprechenden Proteinkonzentrationen, so wird unter Annahme der Proportionalität von Protein- und mRNA-Konzentrationen die zeitliche Veränderung der Konzentrationen beschrieben durch

$$\begin{aligned} \dot{x}_1(t) &= c_1 - \gamma_1 x_1(t) + f_1(y_1(t),...,y_n(t)) \\ y_1(t) &= \alpha_1 x_1(t) \\ &\vdots \quad\quad \vdots \\ \dot{x}_n(t) &= c_n - \gamma_n x_n(t) + f_n(y_1(t),...,y_n(t)) \\ y_n(t) &= \alpha_n x_n(t), \end{aligned} \tag{5.1}$$

wobei die Regulierungsfunktionen $f_i : \mathbf{R}^n \to \mathbf{R}$, $i = 1,...,n$ die Summe der Aktivierungs- und Inhibierungsfunktionen ist, die jeweils abhängig von den Proteinkonzentrationen sind, und $c_i, \gamma_i \in \mathbf{R}$, $i = 1,...,n$, die Grundsynthese und den Grundabbau der i-ten mRNA beschreiben. Dies wurde in Kapitel 2 derart zusammengefasst, dass der Zwischenschritt der Translation entfernt wurde und die Regulierungsfunktionen f_i, $i = 1,...,n$, direkt abhängig sind von den mRNA-Konzentrationen. Dies ist möglich, wenn die Proportionalitätsannahme zugrunde gelegt wird. Die Funktion $f_i(y_1(t),...,y_n(t))$ entspricht dann der Funktion $f_i(\alpha_1 x_1(t),...,\alpha_n y_n(t))$ für $i = 1,...,n$. So lässt sich das Modell für mRNA- und Proteinkonzentration vereinfachen zu einem Modell, dessen Grundlage einzig die mRNA-Konzentrationen sind. Das entsprechende Differenzialgleichungssystem enthält jeweils eine Gleichung für $\dot{x}_1,...,\dot{x}_n$. Ebenso kann man sich nur

auf Proteinkonzentrationen beschränken. Die Gleichungen beschreiben dann die zeitliche Veränderung der Proteinkonzentrationen $\dot{y}_1, ..., \dot{y}_n$ durch

$$\dot{y}_i(t) = \alpha_i c_i - \gamma_i y_i(t) + \alpha_i f_i(y_1(t), ..., y_n(t)) \text{ für } i = 1, ..., n \tag{5.2}$$

mit $\alpha_i, c_i, \gamma_i \in \mathbb{R}^+$.

5.1.1 Posttranskriptionale Regulierungen

Posttranskriptionale Regulierungen spielen in Prokaryonten eine eher geringfügige Rolle. Die bedeutendsten Regulierungen wie alternatives Spleißen oder alternative Polyadenylierungen sind hauptsächlich bei Eukaryonten zu finden. Alternatives Spleißen bewirkt, dass aus einer mRNA mehrere verschiedene Proteine gebildet werden können, indem jeweils verschiedene Teile der mRNA vor der Translation entfernt werden. Für das Modell bedeutet dies, dass für eine mRNA-Konzentration x_i mehrere Variablen für verschiedene Proteinkonzentrationen $y_{i_1}, ..., y_{i_k}$ eingeführt werden müssen. Polyadenylierung beeinflußt die mRNA durch Anhängen mehrerer Adenin-Nukleotide derart, dass sie unter anderem weniger schnell abgebaut wird. Die Polyadenylierung kann nicht nur am Ende der mRNA stattfinden, sondern auch an alternativen mRNA-Stücken. Die Regulierung der Stabilität der mRNA ist jedoch beispielsweise auch über das Entfernen von uracilreichen Regionen möglich. Hier muss im Modell die Abbaurate γ_i der entsprechenden mRNA-Gleichung verändert werden. Je stabiler die mRNA ist, desto kleiner muss γ_i ausfallen. Um die Abhängigkeit von anderen Faktoren zu berücksichtigen, kann statt der Konstanten γ_i eine Funktion $\gamma_i(v)$ betrachtet werden, wobei v ein Vektor ist, der die beeinflussenden Faktoren enthält. Weitere vielfältige Regulierungsmechanismen sind möglich, die den Rahmen dieses Kapitels jedoch sprengen würden. Für weitere Regulierungsmechanismen siehe Janning u. Knust (2004). Wir beschränken uns bei den möglichen Erweiterungen unseres Grundmodells weiterhin auf prokaryontische Modelle, so dass die in diesem Abschnitt beschriebenen Regulierungen im Folgenden, insbesondere in den folgenden Anwendungskapiteln, außer Acht gelassen werden, jedoch für Erweiterungen von eukaryontischen Modellen berücksichtigt werden müssen.

5.1.2 Posttranslationale Regulierungen

Regulierter Abbau

Häufig ist die Proportionalitätsannahme zwischen mRNA- und Proteinkonzentrationen schon aus dem Grund hinfällig, weil Proteine wesentlich länger in der Zelle verbleiben können als die mRNA-Moleküle. In Bakterien beträgt die Halbwertszeit von RNAs beispielsweise im Durchschnitt ca. 3-5 Minuten (Janning u. Knust, 2004). Die Halbwertszeiten von Proteinen schwanken sehr viel stärker. Der Abbau ist insbesondere nicht nur abhängig von der Konzentration des Makromoleküls selbst, sondern kann durch das Vorhandensein von anderen Proteinen begünstigt werden. Gleichung (5.2) wird dann erweitert, indem neben dem Grundabbau, der durch γ_i beschrieben wird, eine Funktion der Form

$$s_{i,j}(x_i(t), x_j(t)) = \begin{cases} 0 & \text{für } x_j(t) \leq \theta_{i,j,1} \\ \lambda \gamma_{i,j} x_i(t) & \text{für } \theta_{i,j,1} < x_j(t) < \theta_{i,j,2} \text{ mit } 0 \leq \lambda \leq 1 \\ \gamma_{i,j} x_i(t) & \text{für } \theta_{i,j,2} \leq x_j(t) \end{cases}$$

eingefügt wird, das heißt, $\dot{y}_i$ wird beschrieben durch $\dot{y}_i(t) = \alpha_i c_i - \gamma_i y_i(t) + \alpha_i f_i(y_1(t),...,y_n(t)) - s_{i,j}(x_i(t),x_j(t))$. Die Abhängigkeit von x_j wird also als Stufenfunktion angenommen. Der Grund liegt darin, dass eine weitere Abhängigkeit von x_i besteht, das Modell jedoch weiterhin die stückweise lineare Form beibehalten soll. Um die dreistufige Approximierung zu verfeinern, können entsprechend mehr Stufen eingefügt werden. Drei Stufen sind jedoch für eine relativ grobe Modellierung ausreichend, da zwei Stufen benötigt werden, um im Wildtyp das Verhalten bei Grundexpression von Gen j und bei verstärkter Expression von Gen j modellieren zu können, und drittens das Verhalten bei völliger Abwesenheit des Transkripts (z. B. in einer Mutante) erfasst werden kann.

Modifikationen und Reifungsprozesse

Proteine werden in der Zelle häufig noch weiter verändert, bevor ihnen Regulierungen möglich sind. Hierbei unterscheidet man zwischen Modifikation und Reifung des Proteins, siehe auch Janning u. Knust (2004), wobei der erste Prozess reversibel, der zweite irreversibel ist. Modifikationen können beispielsweise Acetylierungen, Phosphorylierungen, Methylierungen und Glykosylierungen sein. Ersteres ist das Hinzufügen einer Acetylgruppe an einem Ende des Proteins, wodurch die Lebensdauer des Proteins erhöht werden kann, da Proteasen acetylierte Proteine weniger schnell abbauen. Bei der Phosphorylierung handelt es sich um das Anfügen einer Phosphatgruppe, bei der Methylierung um das Hinzufügen einer Methylgruppe und bei der Glykosylierung entsprechend um das Hinzufügen von Zuckerresten. In Kapitel 6 wird eine Adenylylierung beschrieben, das heißt, das Hinzufügen eines Adenylrests. All diese Veränderungen der Proteine haben zur Folge, dass sich Eigenschaften des Proteins ändern, wie beispielsweise die Bindungseigenschaft. Es sei y_k für ein $k \in \{1,...,n\}$ die Konzentration desjenigen Proteins, das erst durch Veränderung seines Moleküls an das entsprechende DNA-Stück binden kann. In dem Modell, das über (5.2) die Proteinkonzentrationen beschreibt, wird dann zusätzlich die Variable $\tilde{y}_k(t)$ mit aufgenommen, die die Konzentration des auf irgendeine Weise modifizierten Proteins beschreibt. Die bisherigen Funktionen f_i, $i = 1,...,n$, sind nun zusätzlich abhängig von $\tilde{y}_k(t)$, das heißt $f_i : \mathbf{R}^{n+1} \to \mathbf{R}$. Mit Ausnahme von $\dot{y}_k(t)$ und der Änderung der Funktionen f_i bleiben die Differenzialgleichungen für $\dot{y}_1,...,\dot{y}_n$ wie in Gleichung (5.2) beschrieben bestehen. Die Gleichung für $\dot{\tilde{y}}_k(t)$ ändert sich zu $\dot{\tilde{y}}_k(t) = mod(y_k(t),z_1(t),...,z_s(t))$. Hierbei sind die Variablen $z_1(t),...,z_s(t)$ entweder aus der Variablenmenge $\{y_1(t),...,y_n(t)\}$ oder sie sind externe Variablen, die Indikatoren entsprechen, welche beispielsweise über die Zellmembran ins Innere der Zelle gelangen können und Proteine modifizieren können. Entsprechend muss die Gleichung für $y_k(t)$ modifiziert werden durch $\dot{y}_k(t) = \alpha_k c_k - \gamma_k y_k(t) + \alpha_k f_k(y_1(t),...,y_n(t)) - mod(y_k(t),z_1(t),...,z_s(t))$. Im einfachsten Fall ist die Funktion nur abhängig von $y_k(t)$ und der Konzentration eines einzigen Indikators $z(t)$. Setzt man wiederum im einfachsten Fall nur einen einzigen Schwellwert θ für die Indikatorkonzentration, so dass also eine Umwandlung von y_k in $\tilde{y}_k$ für $z(t) > \theta$ erfolgt und eine Rückumwandlung von $\tilde{y}_k$ in $y_k(t)$ für $z(t) \leq \theta$, so erhält man die Funktion *mod* über

$$mod(y_k(t),z(t)) \quad = \quad \begin{cases} k_1 y_k(t) & \text{für } z(t) > \theta \\ -k_2 \tilde{y}_k(t) & \text{für } z(t) \leq \theta \end{cases} \quad ,k_1,k_2 > 0. \tag{5.3}$$

Die Schnelligkeit der Umwandlungen kann demnach verschieden sein und wird durch die Parameter k_1 und k_2 bestimmt. Im allgemeinen Fall, dass die Umwandlungsfunktion *mod* von der Variablen y_k sowie mehreren anderen Variablen $z_1(t),...,z_s(t)$ abhängt, kann ebenfalls im einfachsten Fall jeweils ein Schwellwert für jede Variable $z_1,...,z_s$ angenommen werden und verschiedene Umwandlungsraten $k_1,...,k_{2^s}$ für die 2^s verschiedenen Fälle verwendet werden.

Reifungen von Proteinen sind dagegen irreversible Prozesse. Beispielsweise können manche Proteine erst dann an die DNA binden und dadurch andere Gene regulieren, wenn sie durch Proteasen gespalten wurden. Gehen wir wieder vom einfachsten Fall aus, dass die Konzentration des gereiften Proteins nur von der Konzentration des nicht gereiften Proteins selbst sowie von der Konzentration eines einzigen Indikators $z(t)$ abhängt, so kann solch ein irreversibler Prozess durch die Funktion

$$r(y_k(t),z(t)) \quad = \quad \begin{cases} k_1 y_k(t) & \text{für } z(t) > \theta \\ 0 & \text{für } z(t) \leq \theta \end{cases} \quad ,k_1 > 0 \tag{5.4}$$

beschrieben werden. Die Differenzialgleichungen der beiden Variablen sind dann wieder durch $\dot{\tilde{y}}_k(t) = r(y_k(t),z(t))$ und $\dot{y}_k(t) = \alpha_k c_k - \gamma_k y_k(t) + \alpha_k f_k(y_1(t),...,y_n(t)) - r(y_k(t),z(t))$ gegeben. Wie bei der Modifikation ergibt sich bei Abhängigkeit von mehreren Variablen $z_1,...,z_s$ und der Einführung der Schwellwerte $\theta_1,...,\theta_s$ die Reifungsfunktion, indem 2^{s-1} Umwandlungsraten $k_1,...,k_{2^{s-1}}$ für die verschiedenen Fälle, in denen die Variablen z_i für $i = 1,...,n$ ober- bzw. unterhalb ihres Schwellwerts θ_i liegen, bestimmt werden.

Das ursprüngliche, in Kapitel 2 entwickelte Modell wird also folgendermaßen erweitert:

Definition 5.1. *(Modell für biochemische Netzwerke)*
Die Konzentration x_i eines Makromoleküls i in der Zelle wird durch

$$\begin{aligned} \frac{dx_i(t)}{dt} \quad = \quad & c_i - \gamma_i x_i(t) + \sum_{(i,j)\in\Omega_+} rf^+_{i,j}(x_j(t)) + \sum_{(i,j)\in\Omega_-} rf^-_{i,j}(x_j(t)) \\ & + \sum_{(i,j)\in\Omega_p} f_{i,j}(x_j(t)) \end{aligned} \tag{5.5}$$

beschrieben. Hierbei ist c_i wie im Grundmodell für genregulatorische Netzwerke die Grundsyntheserate, $\gamma_i x_i(t)$ die Abbaurate und $rf^+_{i,j}(x_j(t))$ bzw. $rf^-_{i,j}(x_j(t))$ stückweise lineare Aktivierungs- bzw. Inhibierungsfunktionen der Form (2.32). Ω_+ bzw. Ω_- enthält geordnete Paare (i,j) mit $i,j \in \{1,...,n\}$, so dass Makromolekül j einen aktivierenden bzw. inhibierenden Einfluss auf Makromolekül i hat. Ω_p enthält ebenfalls geordnete Paare (i,j) mit $i,j \in \{1,...,n\}$, so dass Paare aus Ω_p die posttranskriptionalen Regulierungen wie Modifikationen oder Reifungen mithilfe der Funktionen $f_{i,j}(x_j(t))$ beschreiben. Des Weiteren wird im Fall von zusätzlichem Abbau von x_i durch x_j der Term $s_{i,j}(x_j(t),x_i(t))$ auf der rechten Seite der Gleichung subtrahiert.

5.2 Qualitative Beschreibung der Stickstoffaufnahme

Die Hauptkomponenten des in diesem Kapitel betrachteten Systems sind die Proteine AmtR, AmtB, GlnD, GlnK und das durch GlnD modifizierte Protein GlnK, das durch GlnD adenylyliert wird und daher im Folgenden den Namen GlnK∼AMP erhält. Der Transkriptionsfaktor AmtR spielt auf genregulatorischer Ebene eine wichtige Rolle (siehe Abbildung 5.1), da er neben dem *amtB-glnK-glnD*-Operon noch mindestens 33 weitere Gene reguliert (Beckers u. a., 2005).

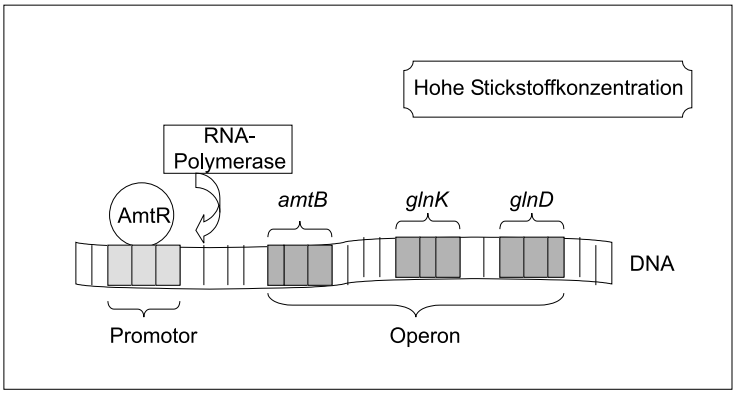


Abbildung 5.1: Im Fall von hoher Stickstoffkonzentration in der Umgebung des Bakteriums inhibiert der Transkriptionsfaktor AmtR die Expression des *amtB-glnK-glnD*-Operons.

Das Protein AmtB ist ein Ammoniumtransporter, der Ammonium in die Zelle hineintransportiert, sich also in der Zellmembran befindet. Neben AmtB gibt es noch den Transporter AmtA, der Ammonium und Methylammonium in die Zelle befördert. AmtB akzeptiert jedoch ausschließlich Ammonium, also NH_4^+ (Burkovski, 2003a). Es wird vermutet, dass es einen weiteren, dritten Transportweg gibt, der jedoch noch unbekannt ist (Meier-Wagner u. a., 2001). Des Weiteren kann Stickstoff auch in die Zelle gelangen, indem NH_3 durch die Zellwand diffundiert (siehe Abbildung 5.2). Ist das Ammonium mittels einer der vermutlich drei Transportwege in

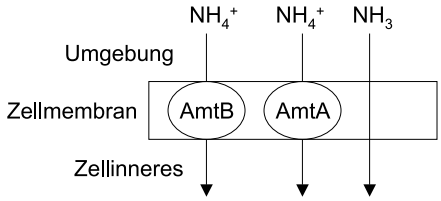


Abbildung 5.2: Verschiedene Wege, auf denen Stickstoff in die Zelle gelangen kann: NH_4^+ kann mittels der Proteine AmtB oder AmtA in das Zellinnere transportiert werden, und NH_3 ist es möglich, durch die Zellwand zu diffundieren (Burkovski, 2003a).

die Zelle gelangt, wird es durch das sogenannte GS/GOGAT-System bzw. durch den GDH-Weg

aufgenommen. Es wird Glutamin bzw. Glutamat aus NH_4^+ produziert, da diese zwei Moleküle besser von der Zelle gespeichert werden können als NH_4^+. Bei Stickstoffmangel ist hauptsächlich das GS/GOGAT-System aktiv, bei Stickstoffüberschuss der GDH-Weg, wobei weiterhin 28% des Ammoniums über den GS-Weg verarbeitet werden (Burkovski, 2003a).
Einen Überblick über die Assimilierung von Ammonium in *C. glutamicum* im Vergleich zu *Escherichia coli* und *Bacillus subtilis* bietet der Übersichtsartikel (Burkovski, 2003a). Bei der Modellierung wird jedoch nur die Stickstoffaufnahme durch AmtB in die Zelle betrachtet, so dass die Assimilierung außer Acht gelassen wird, die keine Rückkopplung auf das betrachtete System verursacht.
Unterschreitet die Stickstoffkonzentration in der Umgebung der Zelle einen bisher unbekannten Schwellwert, so gibt es einen bisher ebenfalls unbekannten Signalweg zu GlnD, dessen Konzentration sich daraufhin erhöht. Die nachfolgenden Prozesse sind jedoch gut untersucht (Burkovski, 2003b). Eine hohe Konzentration des Proteins GlnD bewirkt eine Adenylylierung des Proteins GlnK, welches in niedriger Konzentration in der Zelle vorhanden ist. GlnK∼AMP bindet an den Hauptregulator der Stickstoffaufnahme, AmtR, der bisher den Promotor des *amtB-glnK-glnD*-Operons blockierte und bewirkte, dass nur geringe Mengen mRNA transkribiert wurden (siehe Abbildung 5.1). Durch die Bindung von AmtR durch GlnK∼AMP löst sich AmtR von dem Promotor ab, so dass der Platz für die RNA-Polymerase frei ist und damit eine Transkription für die Gene *amtB, glnK* und *glnD* erreicht wird (siehe Abbildung 5.3). Zu den weiteren

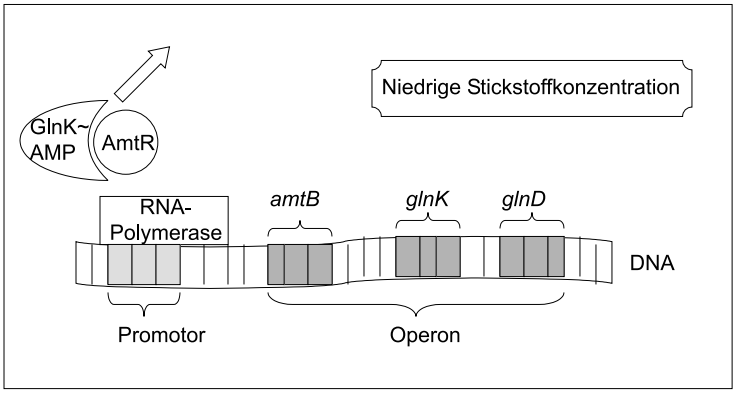


Abbildung 5.3: Im Fall von niedriger Stickstoffkonzentration in der Umgebung des Bakteriums löst GlnK∼AMP den Transkriptionsfaktor AmtR vom Promotor ab, und die Gene *amtB, glnK* und *glnD* werden verstärkt exprimiert.

Genen, die AmtR reguliert, gehören zum Beispiel auch das *amtA*-Gen und das *gltBD*-Operon. Das *amtA*-Gen kodiert den Ammoniumtransporter AmtA, der jedoch aufgrund der Regulierung hauptsächlich bei niedriger Stickstoffkonzentration aktiv ist. Die *gltBD*-Gene kodieren das GOGAT-System, welches bei Stickstoffüberschuss mehr Energie zur Verarbeitung von Ammonium verbraucht als der GDH-Weg und daher bei Stickstoffüberschuss weniger exprimiert wird. Erhöht man nach einer Stickstoffmangelsituation die Stickstoffkonzentration in der Umgebung des Bakteriums sprunghaft, gibt man also einen Stickstoffpuls, so wird wieder ein unbekann-

ter Signalweg zum Protein GlnD aktiv, der die Konzentration von GlnD verringert. Außerdem ändert sich auch die Funktion des Proteins GlnD. Es bewirkt nun in diesem Szenario des Stickstoffüberschusses, dass GlnK∼AMP demodifiziert wird, also deadenylyliert wieder als GlnK vorliegt. Dadurch wird die Bindungsfähigkeit von AmtR an den Promotor des *amtB-glnK-glnD*-Operons nicht länger blockiert, so dass die Expression der Gene *amtB, glnK* und *glnD* erneut unterdrückt wird. Das im Fall von Stickstoffüberschuss in hoher Konzentration vorliegende Protein AmtB verringert sich in seiner Konzentration nun langsam. Das Protein AmtB sorgt dafür, dass sich GlnK an der Zellmembran befindet, solange AmtB noch vorhanden ist. FtsH-Proteasen bauen an dieser Stelle das Protein GlnK ab (siehe Abbildung 5.4).

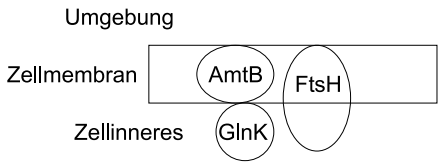


Abbildung 5.4: GlnK bindet an das Protein AmtB und wird von FtsH-Proteasen abgebaut. Ein Teil dieser Makromoleküle sind ebenfalls in der Membran verankert.

Es wird vermutet, dass die Bindung von GlnK an AmtB bewirkt, dass AmtB inaktiv wird und - schon bevor es abgebaut wird - kein Ammonium mehr transportiert. Abbildung 5.5 fasst die Vorgänge während Stickstoffmangel und -überschuss zusammen.

5.3 Datengrundlage und Datenvorverarbeitung

Datengrundlage für die spätere Parameterschätzung waren Westernblots von Wildtyp- und *amtB*-Mutanten-Experimenten, die in Stroesser u. a. (2004) veröffentlicht sind.
Die Erläuterung der Herstellung von Westernblots ist in Kapitel 1.1.2 in dem Abschnitt über Proteinarrays und weitere Messmethoden beschrieben. Die Abbildung 5.6 zeigt die Westernblots für die verwendeten Wildtyp- und *amtB*-Mutanten-Experimente.

Beide Bakterienstämme befanden sich zunächst in einem stickstoffarmen Medium, als zum Zeitpunkt $t = 0$ Minuten ein Stickstoffpuls gegeben wurde. Es wurden Messungen für die Proteine GlnK und GlnK∼AMP zu den Zeitpunkten $t = 0, 1, 2, 4, 6, 8, 10$ und 20 Minuten durchgeführt. Im Folgenden bezeichnen $t_0, ..., t_7$ diese acht Zeitpunkte.
In der Abbildung 5.6 lässt sich schon ohne weitere Auswertungen feststellen, dass mit großen Schwankungen in den Experimenten gerechnet werden muss. Beispielsweise wurde dasselbe Experiment für alle drei Westernblots auf der linken Seite der Abbildung 5.6 durchgeführt. Dennoch erkennt man schon ohne weitere Analyse beim Vergleich des dritten Westernblots mit den ersten beiden, dass GlnK∼AMP im dritten Westernblot nur innerhalb der ersten zwei Minuten nachweisbar war, wohingegen das Protein in den ersten beiden Experimenten bis einschließlich sechs Minuten nach Stickstoffpuls vorliegt. Diese Schwankungen sind typisch für biologische Experimente und haben ihre Ursache beispielsweise in den Ungenauigkeiten der Pipette, Pipet-

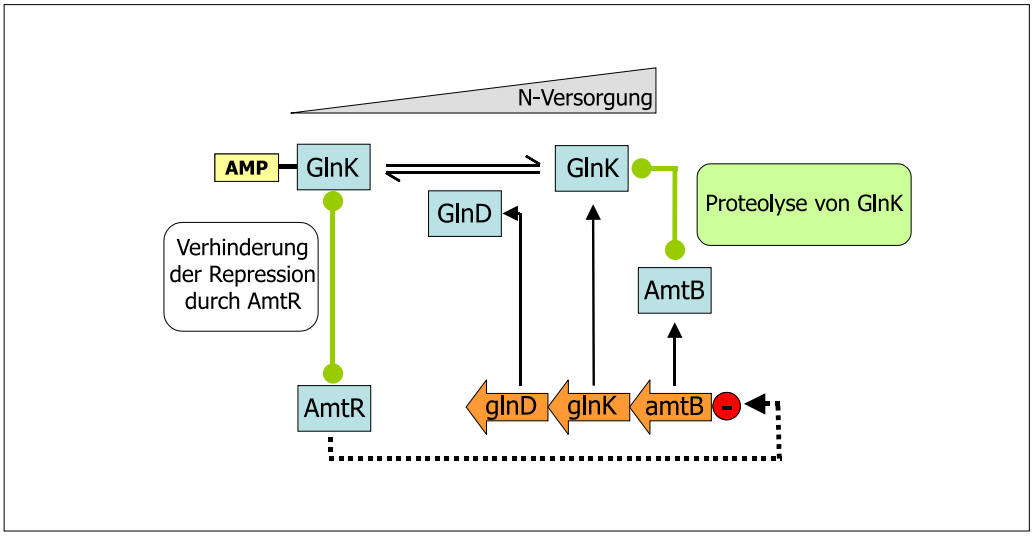


Abbildung 5.5: Die Abbildung erfolgt mit freundlicher Genehmigung von Andreas Burkovski. Sie zeigt ein Schema der Regulierungen bei der Stickstoffaufnahme in *C. glutamicum*. GlnD bewirkt bei Stickstoffmangel eine Modifizierung von GlnK in GlnK∼AMP, bei Stickstoffüberschuss eine Demodifizierung zurück in GlnK. GlnK∼AMP interagiert mit AmtR, welches daraufhin nicht mehr als Repressor der Gene *amtB, glnK* und *glnD* wirken kann. GlnK wird bei Stickstoffüberschuss durch ein Zusammenwirken von AmtB und verschiedenen Proteasen abgebaut.

tierfehlern oder Messungenauigkeiten. Eventuell wurde GlnK∼AMP jedoch durch einen noch unbekannten, weiteren Prozess schneller abgebaut als in den ersten beiden Experimenten.
Um die Westernblots auszuwerten wurde die Software PCBAS 2.0 verwendet, die kalibrierte Intensitäten aufgrund ausgewählter Bildfenster angibt. Die Konzentration von GlnK∼AMP zum Zeitpunkt t_0 wurde von uns auf 100% gesetzt, alle weiteren Werte sind relativ zu dieser Messung zu sehen. Es wurden zwei Durchgänge mit PCBAS 2.0 durchgeführt, je ein Durchgang von Nicole Radde und mir. Die einzelnen Werte und die Mittelwerte für die Westernblots des Wildtyps sind in der Tabelle 5.1, diejenigen für die Westernblots der *amtB*-Mutante in der Tabelle 5.2 zu finden.

Betrachtet man in der Tabelle 5.1 jeweils den Mittelwert der drei Experimente, so liegen die Abweichungen von diesem Mittelwert in beiden Auswertungen bei bis zu 11%. Dies erscheint zwar sehr viel, allerdings lässt sich eine klare Tendenz aus den Daten herauslesen, so dass es gerechtfertigt scheint, mit dem Mittelwert zu arbeiten. In den Wildtypexperimenten nimmt die GlnK∼AMP-Konzentration in den ersten Minuten annähernd exponentiell ab und ist nach spätestens 8 Minuten vollständig abgebaut. Liegt die GlnK∼AMP-Konzentration zu Beginn des Experiments bei 100%, so sieht man nach einer Minute im Mittelwert nur noch ca. 39%. Nach 8 Minuten und im weiteren Verlauf des Experiments lässt sich kein GlnK∼AMP mehr nachweisen. Auch die GlnK-Konzentration wird relativ zu der anfänglichen GlnK∼AMP-Konzentration gemessen. Die GlnK-Konzentration ist zu Beginn des Experiments sehr gering und wird auf 1%

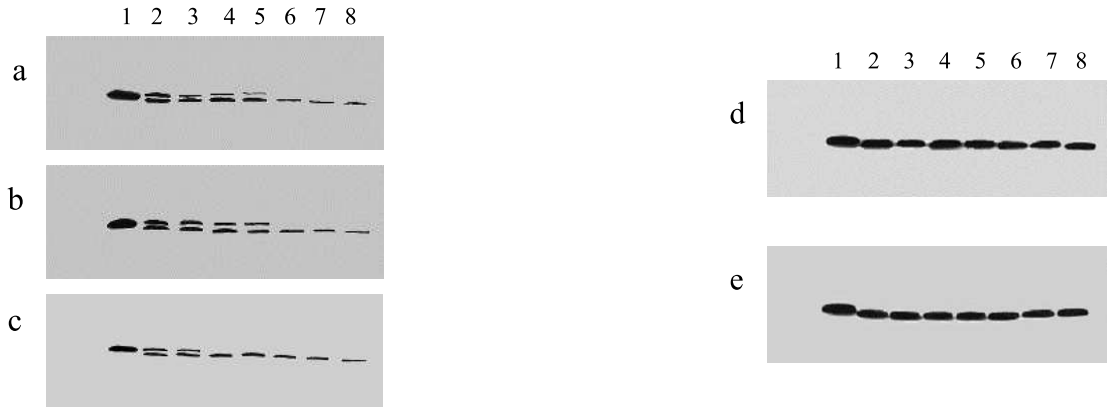


Abbildung 5.6: Westernblots für GlnK und GlnK∼AMP der drei Wildtyp- (a-c)und der zwei *amtB*-Mutantenexperimente (d,e). Mithilfe von GlnK-spezifischen Antikörpern wurde der Abbau von GlnK und GlnK∼AMP in Zellextrakten des Wildtyps (links) und der *amtB*-Mutante (rechts) beobachtet, die in stickstoffreichem Minimalmedium angezogen wurden, dann für zwei Stunden in stickstofffreiem Medium inkubiert wurden (1), und nach 1, 2, 4, 6, 8, 10 und 20 Minuten (2-8) nach Zugabe von 100 mM $(NH_4)_2SO_4$ gemessen wurde. Die obere Reihe zeigt jeweils GlnK∼AMP, die untere Reihe jeweils GlnK.

gesetzt[1]. In den ersten Minuten erreicht sie ihr Maximum, das im Mittel bei über 39% liegt, und wird anschließend langsam abgebaut, bis sie ungefähr 19% der ursprünglichen GlnK∼AMP-Konzentration erreicht. Auf diesem Niveau verbleibt die Konzentration im weiteren Verlauf. Bei zwei der drei Experimente lässt sich feststellen, dass nach dem Maximum innerhalb von ca. 3-4 Minuten noch ein zweites lokales Maximum erreicht wird, das im Mittel bei etwa 35% liegt.

Auch in der *amtB*-Mutante, siehe Tabelle 5.2, kann man ein ähnliches Verhalten von GlnK∼AMP und GlnK beobachten wie im Wildtyp. Allerdings wird in der *amtB*-Mutante GlnK kaum abgebaut und verbleibt nach dem Abbau auf einem wesentlich höheren Niveau als im Wildtyp. GlnK∼AMP verschwindet dagegen noch schneller als im Wildtyp. Die Anfangskonzentration von GlnK∼AMP ist zu Beginn des Experiments auf 100% gesetzt worden, doch schon nach einer Minute lässt sich das Protein nicht mehr nachweisen und bleibt bei 0% für den Rest des Experiments. GlnK startet wieder bei 1% und erreicht innerhalb der ersten Minuten im Mittel wieder ein lokales Maximum, das wesentlich höher liegt als im Wildtyp. Dann nimmt die Konzentration wieder etwas ab, erreicht ein zweites Maximum und sinkt zum Ende des Experiments hin auf ca. $66-70\%$. Der Grund für das höhere Maximum von GlnK und den verringerten Abbau ist in dem Abbauprozess von GlnK zu finden, siehe auch Abbildung 5.4. Das Protein GlnK wird durch das Protein AmtB an der Membran fixiert, so dass die Proteasen das Protein GlnK abbauen können. Findet diese Fixierung aufgrund fehlenden AmtBs nicht statt, so kann auch der Abbau nicht erfolgen.

[1] GlnK ist bei Stickstoffmangel in geringer Konzentration immer in der Zelle vorhanden.

Tabelle 5.1: GlnK∼AMP (1) und GlnK (2) im Wildtyp nach Stickstoffpuls zu den Zeitpunkten $t_0,...,t_7$. Intensitätswerte und Mittelwerte der Westernblots sind aufgelistet. $W1\_WT_1,...,W3\_WT_1$ und μ_WT_1 entspricht den drei Westernblots sowie dem Mittelwert der von mir durchgeführten Auswertung, $W1\_WT_2,...,W3\_WT_2$ und μ_WT_2 der von Nicole Radde durchgeführten Auswertung.

(1)	$W1\_WT_1$	$W2\_WT_1$	$W3\_WT_1$	μ_WT_1	$W1\_WT_2$	$W2\_WT_2$	$W3\_WT_2$	μ_WT_2
t_0	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
t_1	32.00	44.00	38.00	38.00	39.00	42.00	40.00	40.33
t_2	11.00	37.00	25.00	24.33	15.00	37.00	25.00	25.67
t_3	17.00	27.00	0.00	14.67	19.00	27.00	0.00	15.33
t_4	13.00	27.00	0.00	13.33	13.00	27.00	0.00	13.33
t_5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
t_6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
t_7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
(2)	$W1\_WT_1$	$W2\_WT_1$	$W3\_WT_1$	μ_WT_1	$W1\_WT_2$	$W2\_WT_2$	$W3\_WT_2$	μ_WT_2
t_0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
t_1	42.00	35.00	41.00	39.33	37.00	38.00	41.00	38.67
t_2	31.00	31.00	43.00	35.00	26.00	31.00	43.00	33.33
t_3	35.00	35.00	44.00	38.00	35.00	35.00	44.00	38.00
t_4	28.00	24.00	45.00	32.33	28.00	25.00	46.00	33.00
t_5	17.00	21.00	34.00	24.00	17.00	21.00	34.00	24.00
t_6	14.00	17.00	38.00	23.00	14.00	17.00	38.00	23.00
t_7	15.00	17.00	25.00	19.00	15.00	17.00	25.00	19.00

Der Grund für die unterschiedlichen Ergebnisse bei der Auswertung der Westernblots liegt in den Bildfenstern, die über die Westernblots gelegt werden. Da die Proteine GlnK und GlnK∼AMP ähnlich groß sind, ist in manchen Fällen die Trennung im Gel nicht vollständig, was zu den Schwankungen in den Ergebnissen führt. Im Folgenden werden die Mittelwerte aus beiden Auswertungen verwendet, siehe Tabellen 5.3 und 5.4.

5.4 Modellierung, Parameterschätzung und Simulationen

5.4.1 Modellierung

Der Signaltransduktionsweg von einer veränderten Stickstoffkonzentration zu einer veränderten Konzentration des Proteins GlnD ist bisher unbekannt. Daher wird für die Modellbildung die GlnD-Konzentration als Indikator für die Stickstoffkonzentration angenommen. Aufgrund

Tabelle 5.2: GlnK∼AMP (1) und GlnK (2) in der *amtB*-Mutante nach Stickstoffpuls zu den Zeitpunkten $t_0,...,t_7$. Intensitätswerte und Mittelwerte der Westernblots sind aufgelistet. $W1\_M_1, W2\_M_1$ und μ_M_1 entspricht den zwei Westernblots sowie dem Mittelwert der von mir durchgeführten Auswertung, $W1\_M_2, W2\_M_2$ und μ_M_2 der von Nicole Radde durchgeführten Auswertung.

(1)	$W1\_M_1$	$W2\_M_1$	μ_M_1	$W1\_M_2$	$W2\_M_2$	μ_M_2
t_0	100.00	100.00	100.00	100.00	100.00	100.00
t_1	0.00	0.00	0.00	0.00	0.00	0.00
t_2	0.00	0.00	0.00	0.00	0.00	0.00
t_3	0.00	0.00	0.00	0.00	0.00	0.00
t_4	0.00	0.00	0.00	0.00	0.00	0.00
t_5	0.00	0.00	0.00	0.00	0.00	0.00
t_6	0.00	0.00	0.00	0.00	0.00	0.00
t_7	0.00	0.00	0.00	0.00	0.00	0.00
(2)	$W1\_M_1$	$W2\_M_1$	μ_M_1	$W1\_M_2$	$W2\_M_2$	μ_M_2
t_0	1.00	1.00	1.00	1.00	1.00	1.00
t_1	62.00	89.00	75.50	59.00	86.00	72.50
t_2	65.00	71.00	68.00	66.00	59.00	62.50
t_3	68.00	95.00	81.50	60.00	82.00	71.00
t_4	60.00	76.00	68.00	61.00	73.00	67.00
t_5	67.00	68.00	67.50	68.00	59.00	63.50
t_6	54.00	66.00	60.00	53.00	64.00	58.50
t_7	68.00	70.00	69.00	67.00	66.00	66.50

fehlender Messwerte wird hier eine Boolesche Variable gewählt:

$$I_{\text{GlnD}} = \begin{cases} 1 & \text{bei hoher Stickstoffkonzentration in der Umgebung des Bakteriums} \\ 0 & \text{bei niedriger} \qquad\qquad \text{“} \end{cases} \tag{5.6}$$

Bei niedriger Stickstoffkonzentration ist die Konzentration von GlnD hoch und bewirkt die Modifikation von GlnK in GlnK∼AMP (Adenylylierung). Ist die Stickstoffkonzentration hoch, so sinkt die Konzentration von GlnD, was eine Demodifizierung von GlnK∼AMP zurück in GlnK nach sich zieht (Deadenylylierung). GlnK∼AMP inhibiert den Transkriptionsfaktor AmtR, der wiederum das *amtB-glnK-glnD*-Operon inhibiert. Diese zweifache Inhibierung wird im Modell als einfache positive Regulierung zusammengefasst mit dem Resultat, dass AmtR aus dem Modell entfernt werden kann. Genauer gesagt wird im Modell aus der Inhibierung von AmtR durch GlnK∼AMP sowie der Inhibierung von GlnK∼AMP durch AmtR eine Aktivierung von GlnK∼AMP auf sich selbst. Außerdem wird aus der Inhibierung von AmtR durch GlnK∼AMP sowie der Inhibierung von AmtB durch AmtR eine Aktivierung von GlnK∼AMP auf AmtB.
Dem Protein GlnK∼AMP wird im Folgenden der Index 1 zugeteilt, dem Protein GlnK der Index 2 und dem Protein AmtB der Index 3. Entsprechend bezeichnet x_1 die Konzentration von GlnK∼AMP, x_2 die Konzentration von GlnK und x_3 die Konzentration von AmtB.

Tabelle 5.3: Aus den Westernblots der Wildtypexperimente nach Stickstoffpuls gemittelte Daten für GlnK∼AMP und GlnK (μ_WT) sowie Standardabweichungen (σ_WT) zu den Zeitpunkten $t_0,...,t_7$.

t	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
μ_WT([GlnK∼AMP])	100.00	39.17	25.00	15.00	13.33	0.00	0.00	0.00
σ_WT([GlnK∼AMP])	0.00	3.76	9.87	11.24	11.03	0.00	0.00	0.00
μ_WT([GlnK])	1.00	39.00	34.17	38.00	32.67	24.00	23.00	19.00
σ_WT([GlnK])	0.00	2.52	6.49	4.24	9.20	7.26	10.68	4.32

Tabelle 5.4: Aus den Westernblots der *amtB*-Mutante nach Stickstoffpuls gemittelte Daten für GlnK∼AMP und GlnK (μ_M) sowie Standardabweichungen (σ_M) zu den Zeitpunkten $t_0,...,t_7$.

t	0	1	2	4	6	8	10	20
μ_M([GlnK∼AMP])	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
σ_M([GlnK∼AMP])	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
μ_M([GlnK])	1.00	74.00	65.25	76.25	67.50	65.50	59.25	67.75
σ_M([GlnK])	0.00	15.68	4.92	15.46	8.19	4.36	6.70	1.71

Die bereits beschriebenen Aktivierungen im Modell finden auf Genregulationsebene statt und werden daher mit den in Kapitel 2 hergeleiteten Regulierungsfunktionen beschrieben:

$$rf_{1,1}(x_1) = \begin{cases} 0 & \text{für } x_1 \leq \theta_{1,1,1} \\ \frac{k_{1,1}}{\theta_{1,1,2}-\theta_{1,1,1}}(x_1-\theta_{1,1,1}) & \text{für } \theta_{1,1,1} < x_1 < \theta_{1,1,2} \\ k_{1,1} & \text{für } \theta_{1,1,2} \leq x_1 \end{cases} \tag{5.7}$$

und

$$rf_{3,1}(x_1) = \begin{cases} 0 & \text{für } x_1 \leq \theta_{3,1,1} \\ \frac{k_{3,1}}{\theta_{3,1,2}-\theta_{3,1,1}}(x_1-\theta_{3,1,1}) & \text{für } \theta_{3,1,1} < x_1 < \theta_{3,1,2} \\ k_{3,1} & \text{für } \theta_{3,1,2} \leq x_1 \end{cases} . \tag{5.8}$$

Alle weiteren Prozesse in dem betrachteten System finden auf Proteinebene statt.
Die Modifizierung und Demodifizierung des Proteins GlnK kann durch

$$mod(GlnD,x_1,x_2) = \begin{cases} -k_1x_1 & \text{für } I_{\text{GlnD}} = 1 \\ k_2x_2 & \text{für } I_{\text{GlnD}} = 0 \end{cases} \tag{5.9}$$

charakterisiert werden. Die Funktion *mod* wird in derjenigen Differenzialgleichung hinzuaddiert, die das zeitliche Verhalten von GlnK∼AMP beschreibt, und subtrahiert in der Differenzialgleichung, die GlnK beschreibt.
Der Abbau von GlnK durch Interaktion mit AmtB und nachfolgende Proteolyse ist ebenfalls

eine posttranskriptionale Regulierung und wird als proportional zur Konzentration von GlnK angenommen. Bisher wurde AmtB noch nicht mit Westernblots gemessen, weil hierzu kein Antikörper existiert, der zum Nachweis benötigt wird. Man weiß jedoch, dass der Abbau durch die Proteasen verhindert wird, wenn kein AmtB vorliegt, dass der Abbau bei hoher AmtB-Konzentration ebenfalls hoch ist, und dass er bei niedriger AmtB-Konzentration ebenfalls niedrig ist. Somit bietet sich eine dreistufige Abbauregulierung an:

$$s^*_{2,3}(x_2,x_3) = \begin{cases} 0 & \text{für } x_3 \leq \theta_{2,3,1} \\ \lambda\gamma_{2,3}x_2 & \text{für } \theta_{2,3,1} < x_3 < \theta_{2,3,2} \text{ mit } 0 \leq \lambda \leq 1 \\ \gamma_{2,3}x_2 & \text{für } \theta_{2,3,2} \leq x_3 \end{cases} . \tag{5.10}$$

Grundsätzlich wäre diese Funktion auch von der Proteasekonzentration abhängig. Im Modell wird jedoch aufgrund fehlender Daten über Proteasen angenommen, dass die Proteasen immer in hinreichender Menge in der Zelle vorhanden sind, so dass die Funktion demnach nur abhängig von x_2 und x_3 ist. Simulationen für eine reduzierte Proteasenaktivität können dennnoch ausgeführt werden, indem $s^*_{2,3}$ mit einem entsprechenden Faktor zwischen 0 und 1 multipliziert wird, welcher der Proteasenaktivität entspricht.
Das dynamische Verhalten von x_1, x_2 und x_3 kann mithilfe der Funktionen (5.6)-(5.10) wie folgt beschrieben werden (siehe Gebert u. a. (2007a)):

$$\dot{x}_1(t) = -\gamma_1 x_1(t) + rf_{1,1}(x_1(t)) + mod(GlnD, x_1(t), x_2(t)) \tag{5.11}$$
$$\dot{x}_2(t) = c_2 - \gamma_2 x_2(t) - mod(GlnD, x_1(t), x_2(t)) - s^*_{2,3}(x_2(t), x_3(t)) \tag{5.12}$$
$$\dot{x}_3(t) = c_3 - \gamma_3 x_3(t) + rf_{3,1}(x_1(t)) \tag{5.13}$$

Im nächsten Schritt wird eine Vereinfachung vorgenommen, welche die ersten zwei Gleichungen von der dritten entkoppelt, da aufgrund der Experimente nur über GlnK und GlnK∼AMP Aussagen gemacht werden können. GlnK∼AMP wirkt über AmtR und über AmtB auf den Abbau von GlnK ein, so dass im Folgenden die Funktion $s^*_{2,3}$ von x_2 und x_1 abhängig gemacht wird anstelle von x_2 und x_3:

$$s_{2,3}(x_2,x_1) = \begin{cases} 0 & \text{für } \textit{amtB}\text{-Mutanten} \\ \lambda\gamma_{2,3}x_2 & \text{für } \theta_{2,1,1} < x_1 < \theta_{2,1,2} \text{ mit } 0 \leq \lambda \leq 1 \\ \gamma_{2,3}x_2 & \text{für } \theta_{2,1,2} \leq x_1 \end{cases} . \tag{5.14}$$

Für die Parameterschätzung und die Simulationen wird im Folgenden also $s^*_{2,3}(x_2,x_3)$ durch $s_{2,3}(x_2,x_1)$ ersetzt.

5.4.2 Parameterschätzung

Fixpunktgleichungen

Man kann davon ausgehen, dass sowohl zum Zeitpunkt t_0 als auch zum Zeitpunkt t_7 ein Fixpunkt vorliegt, daher können für den Wildtyp Gleichungen für diese Fixpunkte aufgestellt werden. Diese Gleichungen entsprechen den Differenzialgleichungen, die bei den jeweiligen Fixpunkten

gelten und enthalten die zugehörigen Werte für x_1 und x_2 aus Tabelle 5.3:

$$\begin{aligned} k_2 + k_{1,1} - 100\gamma_1 &= 0 && \text{GlnK}\sim\text{AMP im Wildtyp bei Stickstoffmangel,} && (5.15)\\ -k_2 + c_2 - \gamma_{2,3} - \gamma_2 &= 0 && \text{GlnK im Wildtyp bei Stickstoffmangel,} && (5.16)\\ -0k_1 - 0\gamma_1 &= 0 && \text{GlnK}\sim\text{AMP im Wildtyp bei Stickstoffüberschuss,} && (5.17)\\ c_2 - 19\lambda\gamma_{2,3} - 19\gamma_2 &= 0 && \text{GlnK im Wildtyp bei Stickstoffüberschuss.} && (5.18) \end{aligned}$$

Für jeden Fixpunkt erhält man zwei Gleichungen, jeweils eine für GlnK und eine für GlnK$\sim$AMP gemäß der Differenzialgleichungen (5.11) und (5.12). Die ersten beiden Gleichungen (5.15) und (5.16) beschreiben x_1 und x_2 in ihrem Fixpunkt bei Stickstoffmangel, wohingegen die beiden letzten Gleichungen (5.17) und (5.18) den Fixpunkt bei Stickstoffüberschuss beschreiben. Da Gleichung (5.17) keine Aussage beinhaltet, stehen den sieben Parametern nur drei Gleichungen gegenüber. Diesem unterbestimmten System müssen weitere Bedingungen oder Informationen zugeführt werden, wir wählen hierzu die Wildtypmessungen aus Tabelle 5.3 zu den Zeitpunkten $t_0, ..., t_7$. Die Messungen der *amtB*-Mutante aus Tabelle 5.4 sollen später zur Evaluierung herangezogen werden und sollen daher nicht weiter in die Parameterschätzung eingehen. Doch aus den bisher gewählten Daten alleine lässt sich zwischen dem grundsätzlichen Abbau von GlnK, γ_2, und dem Abbau von GlnK durch AmtB, $\gamma_{2,3}$ bzw. $\lambda\gamma_{2,3}$, nicht unterscheiden. Daher wird auch der Fixpunkt für Stickstoffüberschuss der *amtB*-Mutante zum Zeitpunkt t_7 für die Parameterschätzung verwendet, siehe Tabelle 5.4:

$$\begin{aligned} -0k_1 - 0\gamma_1 &= 0 && \text{GlnK}\sim\text{AMP in der } amtB\text{-Mutante bei Stickstoffüberschuss,} && (5.19)\\ c_2 - 67.75\gamma_2 &= 0 && \text{GlnK in der } amtB\text{-Mutante bei Stickstoffüberschuss.} && (5.20) \end{aligned}$$

Im Folgenden werden nur die vier Gleichungen (5.15), (5.16), (5.18) und (5.20) verwendet. Dabei ist jedoch zu beachten, dass die Zahl 67.75 in Gleichung (5.20) relativ zu der GlnK$\sim$AMP-Konzentration der *amtB*-Mutante im Fixpunkt bei Stickstoffmangel zu setzen ist. In Bezug auf die GlnK$\sim$AMP-Konzentration des Wildtyps im Fixpunkt bei Stickstoffmangel ist dieser Wert unbekannt. Der Wert 67.75 wird zunächst als erste Schätzung für x_2 angenommen, dann jedoch iterativ verbessert. Das Verfahren wird im Unterkapitel „Schätzung der restlichen Parameter“ näher erläutert.

Schätzung der Schwellwerte

Aufgrund der kurzen Zeitreihen wählen wir das Verfahren aus Kapitel 4.2. Die Daten müssen also vor der eigentlichen Parameterschätzung den verschiedenen linearen Differenzialgleichungen zugeordnet werden, das heißt, im ersten Schritt werden die Schwellwerte $\theta_{1,1,1}$, $\theta_{1,1,2}$ und $\theta_{2,1,2}$ geschätzt. Gibt man dem System einen Stickstoffpuls, so nehmen wir an, dass folgende Reihenfolge für die biologischen Prozesse und somit auch für die Differenzialgleichungen zutrifft: Zunächst wird GlnK$\sim$AMP zurückverwandelt in GlnK, x_1 nimmt also ab, x_2 wächst. Unterschreitet x_1 seinen Schwellwert $\theta_{1,1,2}$, so verringert sich die Aktivierung von GlnK$\sim$AMP auf sich selbst, so dass auch weniger GlnK produziert wird, x_2 nimmt ab. Das erste Maximum von x_2 liegt zwischen t_0 und t_2, und führt somit zu dem Schwellwert $\theta_{1,1,2}$. Unterschreitet x_1 den Schwellwert $\theta_{2,1,2}$, so wird der Abbau von GlnK durch AmtB verlangsamt, man erhält also

einen leichten Anstieg von x_2. Das lokale Minimum liegt zwischen t_1 und t_3 und führt zu dem Schwellwert $\theta_{2,1,2}$. Unterschreitet x_1 den letzten Schwellwert $\theta_{1,1,1}$ so wird wiederum weniger GlnK∼AMP produziert mit der Folge, dass auch x_2 abnimmt. Dieses zweite lokale Maximum liegt zwischen t_2 und t_4 und führt zu dem Schwellwert $\theta_{1,1,1}$. Die genauen Werte für die Schwellwerte werden berechnet, indem eine Schätzung für die lokalen Maxima und Minima von x_2 vorgenommen wird und dann die jeweiligen Werte für t in die polynomielle Regression von x_1 eingesetzt wird, siehe Tabelle 5.5. Mithilfe der Schwellwerte können die einzelnen Messwerte

Tabelle 5.5: Schätzung der Schwellwerte. Zunächst wird ein Polynom $p(t)$ der Ordnung zwei durch drei Wertepaare für x_2 gelegt, die das Extremum einschliessen. Dieses Polynom $p(t)$ für die x_2-Werte befindet sich in der dritten Zeile. Die Ableitung $p'(t)$ wird gleich Null gesetzt, und die Lösung dieser Gleichung findet sich in der vierten Zeile. Für dieselben Zeitpunkte wird ein Polynom $f(t)$ für die x_1-Werte berechnet, das in der fünften Zeile steht. Für das erste Polynom $p(t)$ sei das Extremum an der Stelle $t = \hat{t}$, dann wird $f(\hat{t})$ als Schätzung für den Schwellwert gesetzt. Man erhält also jeweils einen Wert für x_1, der einem Extremum von x_2 entspricht.

Schwellwert	$\theta_{1,1,2}$	$\theta_{2,1,2}$	$\theta_{1,1,1}$
Zeitpunkte der Wertepaare	t_0, t_1, t_2	t_1, t_2, t_3	t_2, t_3, t_4
Polynom $p(t)$ für x_2-Werte	$-21.4t^2 + 59.4t + 1$	$2.23t^2 - 11.5t + 48.27$	$-1.1375t^2 + 8.725t + 21.3$
Art des Extremums	Maximum	Minimum	Maximum
$\hat{t}$ *mit* $p'(\hat{t}) = 0$	1.39	2.58	3.84
Polynom $f(t)$ für x_1-Werte	$23.33t^2 - 84t + 100$	$3.067t^2 - 23.4t + 59.53$	$1.0375t^2 - 11.225t + 43.3$
Schwellwert-schätzung $f(\hat{t})$	28.1	19.6	15.5

nun den linearen Differenzialgleichungen zugeordnet werden, so dass im Anschluss die übrigen Parameter unter Verwendung der linearen Gleichungen geschätzt werden können. Messwerte zu den Zeitpunkten t_0 und t_1 werden den Gleichungen

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = -k_1 x_1(t) + k_{1,1} - \gamma_1 x_1(t) \tag{5.21}$$

$$\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} = k_1 x_1(t) + c_2 - \gamma_{2,3} x_2(t) - \gamma_2 x_2(t) \tag{5.22}$$

aufgrund von $x_1 > \theta_{1,1,2}$ zugeordnet, Messwerte zum Zeitpunkt t_2 aufgrund von $\theta_{2,1,2} < x_1 < \theta_{1,1,2}$ den Gleichungen

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = -k_1 x_1(t) + \frac{k_{1,1}}{12.6}(x_1(t) - 15.5) - \gamma_1 x_1(t) \tag{5.23}$$

$$\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} = k_1 x_1(t) + c_2 - \gamma_{2,3} x_2(t) - \gamma_2 x_2(t). \tag{5.24}$$

Für $\theta_{1,1,1} < x_1 < \theta_{2,1,2}$ liegen keine Messwerte vor. Die restlichen Daten zu den Zeitpunkten $t_3, ..., t_7$ werden durch die Gleichungen

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = -k_1 x_1(t) - \gamma_1 x_1(t) \tag{5.25}$$

$$\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} = k_1 x_1(t) + c_2 - \lambda \gamma_{2,3} x_2(t) - \gamma_2 x_2(t) \tag{5.26}$$

beschrieben, da $x_1 < \theta_{1,1,1}$, und führen zum Fixpunkt bei Stickstoffüberschuss.
Um eine Gleichgewichtung jeder Differenzialgleichung in der Parameterschätzung zu erhalten, werden die Messwerte zu den Zeitpunkten t_0 und t_1 jeweils fünffach, die zu den Zeitpunkten t_2 jeweils zehnfach und die zu den übrigen Zeitpunkten jeweils doppelt gewichtet.

Schätzung der restlichen Parameter

Wie in Kapitel 4.2 beschrieben, werden zunächst mit Gleichung (4.6) die Differenzialquotienten $\frac{dx_i(t)}{dt}$ zu den acht Zeitpunkten $t_0, ..., t_7$ geschätzt. Tabelle 5.6 listet die berechneten Werte auf.

Tabelle 5.6: Schätzungen für $\frac{dx_1(t)}{dt}$ und $\frac{dx_2(t)}{dt}$ zu den Zeitpunkten $t_0, ..., t_7$.

t	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
$\frac{d\hat{x}_1(t)}{dt}$ in $[min^{-1}]$	-84.17	-37.5	-11.11	-2.92	-3.75	-3.33	0.00	0.00
$\frac{d\hat{x}_2(t)}{dt}$ in $[min^{-1}]$	59.42	16.58	-2.58	-0.375	-3.5	-2.42	-0.48	-0.32

Zur Parameterschätzung des Vektors

$$x = (k_1, \gamma_1, k_{1,1}, c_2, \gamma_{2,3}, \lambda\gamma_{2,3}, \gamma_2, k_2)^t \tag{5.27}$$

wird unter Verwendung der beiden Matrix-Vektor-Paare

$$A = \begin{pmatrix} 0 & -100 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & 0 & -19 & -19 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -67.75 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

und

$$C = \begin{pmatrix}
-100 & -100 & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & 5mal & & & & \\
-100 & -100 & 1 & 0 & 0 & 0 & 0 & 0 \\
100 & 0 & 0 & 1 & -1 & 0 & -1 & 0 \\
 & & & 5mal & & & & \\
100 & 0 & 0 & 1 & -1 & 0 & -1 & 0 \\
-39.17 & -39.17 & 1 & 0 & 0 & 0 & 0 & 0 \\
 & & & 5mal & & & & \\
-39.17 & -39.17 & 1 & 0 & 0 & 0 & 0 & 0 \\
39.71 & 0 & 0 & 1 & -39 & 0 & -39 & 0 \\
 & & & 5mal & & & & \\
39.71 & 0 & 0 & 1 & -39 & 0 & -39 & 0 \\
-25 & -25 & 0.72 & 0 & 0 & 0 & 0 & 0 \\
 & & & 10mal & & & & \\
-25 & -25 & 0.72 & 0 & 0 & 0 & 0 & 0 \\
25 & 0 & 0 & 1 & -34.17 & 0 & -34.17 & 0 \\
 & & & 10mal & & & & \\
25 & 0 & 0 & 1 & -34.17 & 0 & -34.17 & 0 \\
-15 & -15 & 0 & 0 & 0 & 0 & 0 & 0 \\
-15 & -15 & 0 & 0 & 0 & 0 & 0 & 0 \\
15 & 0 & 0 & 1 & 0 & -38 & -38 & 0 \\
15 & 0 & 0 & 1 & 0 & -38 & -38 & 0 \\
-13.33 & -13.33 & 0 & 0 & 0 & 0 & 0 & 0 \\
-13.33 & -13.33 & 0 & 0 & 0 & 0 & 0 & 0 \\
13.33 & 0 & 0 & 1 & 0 & -32.67 & -32.67 & 0 \\
13.33 & 0 & 0 & 1 & 0 & -32.67 & -32.67 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -24 & -24 & 0 \\
0 & 0 & 0 & 1 & 0 & -24 & -24 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -23 & -23 & 0 \\
0 & 0 & 0 & 1 & 0 & -23 & -23 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -19 & -19 & 0 \\
0 & 0 & 0 & 1 & 0 & -19 & -19 & 0
\end{pmatrix}, \quad d = \begin{pmatrix}
-84.17 \\
5mal \\
-84.17 \\
47.42 \\
5mal \\
47.42 \\
-37.5 \\
5mal \\
-37.5 \\
12.58 \\
5mal \\
12.58 \\
-11.11 \\
10mal \\
-11.11 \\
-2.58 \\
10mal \\
-2.58 \\
-2.92 \\
-2.92 \\
-0.375 \\
-0.375 \\
-3.75 \\
-3.75 \\
-3.5 \\
-3.5 \\
-3.33 \\
-3.33 \\
-2.42 \\
-2.42 \\
0 \\
0 \\
-0.483 \\
-0.483 \\
0 \\
0 \\
0 \\
0
\end{pmatrix}$$

das Optimierungsproblem

$$\min_x ||Cx - d||_2^2 \tag{5.28}$$

mit den Nebenbedingungen

$$Ax = b \tag{5.29}$$

gelöst (siehe Kapitel 4.2).
In der Gleichung $Ax = b$ wird festgelegt, dass die Fixpunktgleichungen exakt erfüllt sein sollen und dass die Abbauraten von GlnK∼AMP und GlnK gleich hoch sein sollen.
Die Zahl 67.75 in der Matrix A bedeutet, dass die GlnK-Konzentration in der *amtB*-Mutante nach 20 Minuten noch 67.75% der GlnK∼AMP-Konzentration zu Beginn des Experiments entspricht. Der Wert der GlnK∼AMP-Konzentration liegt jedoch höher als 100, wenn man ihn im Vergleich zum Wildtyp betrachtet, zu dem alle sonstigen Werte in Bezug gesetzt wurden. Daher wird mit 67.75 als vorläufigem Wert eine Parameterschätzung durchgeführt, die zu dem Vektor

$$\hat{x} \quad = \quad (0.61445, 0.09176, 3.5685, 6.2167, 0.51745, 0.23544, 0.09176, 5.6075)^{\,t} \tag{5.30}$$

führt.
In der *amtB*-Mutante gilt im Stickstoffmangel folgende Gleichung

$$0 \quad = \quad k_2\bar{x}_2 + k_{1,1} - \gamma_1\bar{x}_1 \tag{5.31}$$
$$0 \quad = \quad -k_2\bar{x}_2 + c_2 - \gamma_2\bar{x}_2. \tag{5.32}$$

Setzt man die entsprechenden geschätzten Werte aus Gleichung (5.30) ein, so erhält man die Fixpunktwerte für x_1 und x_2 in der Mutante bei Stickstoffmangel:

$$\bar{x}_1 \quad = \quad 105.502 \tag{5.33}$$
$$\bar{x}_2 \quad = \quad 1.091. \tag{5.34}$$

Nun beginnt man die Parameterschätzung erneut mit der veränderten Matrix

$$\tilde{A} \quad = \quad \begin{pmatrix} 0 & -100 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & 0 & -19 & -19 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -71.477 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}, \tag{5.35}$$

die anstelle von A verwendet wird. Hierbei gilt $71.477 = 0.6775 \cdot 105.502$, der neue Wert in der Matrix entspricht also 67.75% des mit den geschätzten Parametern berechneten GlnK∼AMP-Wertes für Stickstoffmangel in der Mutante. Nach sechs Iterationen erhält man den Wert 71.430, nach sieben 71.484, nach acht Iterationen wiederum 71.430. Somit konvergiert dieses Verfahren gegen einen Grenzzyklus, in dessen Bereich sich die Schätzungen für die Parameter befinden.

Folgende Schranken sind für die Lösungen durch den Grenzzyklus gegeben:

$$\begin{aligned} k_1: &\quad 0.6090 - 0.6091/min \\ \gamma_1: &\quad 0.0944 - 0.0945/min \\ k_{1,1}: &\quad 3.3133 - 3.3174/min \\ c_2: &\quad 6.7452 - 6.7525/min \\ \gamma_{2,3}: &\quad 0.5250 - 0.5251/min \\ \lambda: &\quad 0.4964 - 0.4969 \\ \lambda\gamma_{2,3}: &\quad 0.2606 - 0.2609/min \\ \gamma_2: &\quad 0.0944 - 0.0945/min \\ k_2: &\quad 6.1258 - 6.1329/min \end{aligned}$$

Da im Folgenden nur mit zwei Stellen hinter dem Komma gerechnet wird, lautet das Ergebnis der Parameterschätzung zusammengefasst

$$k_1 = 0.61/min, \gamma_1 = 0.09/min, k_{1,1} = 3.32/min, c_2 = 6.75/min, \gamma_{2,3} = 0.53/min, \quad (5.36)$$

$$\lambda = 0.50 \text{ bzw. } \lambda\gamma_{2,3} = 0.26/min, \gamma_2 = 0.09/min \text{ und } k_2 = 6.13/min. \quad (5.37)$$

5.4.3 Simulationen

In diesem Kapitel werden drei unterschiedliche Typen von Simulationen diskutiert. Zunächst wird das Experiment simuliert, aus dem die Parameterschätzung hervorgegangen ist, um zu überprüfen, ob die parametrisierten Differenzialgleichungen als Grundlage angenommen werden können. Zweitens werden verschiedene Bedingungen simuliert, für die es keine Messdaten gibt und die nur auf biologische Plausibilität geprüft werden können. Sie können jedoch Vorhersagen möglich machen, um weitere Experimente zu planen. Zum dritten werden Simulationen für Experimente vorgenommen, die bisher nicht in die Parameterschätzung eingeflossen sind. Diese sind besonders wertvoll für die Evaluierung des Modells.
In der ersten Simulation werden die Bedingungen gewählt, die den experimentellen Daten zur Parameterschätzung zugrunde liegen. Die Bakterien werden in stickstoffarmem Medium angezogen und einem Stickstoffpuls zum Zeitpunkt t_0 ausgesetzt. Das Ergebnis der Simulation mit Anfangsbedingungen $x_1(t_0) = 100$ und $x_2(t_0) = 1$ ist in Abbildung 5.7 dargestellt. Die anfänglich hohe GlnK∼AMP-Konzentration fällt im Laufe der Simulation innerhalb von 5 bis 10 Minuten auf Null. GlnK wächst von $x_2(t_0) = 1$ innerhalb von 2 Minuten auf knapp 42 an, fällt leicht auf ca. 40, um sich dann nach einem erneuten Maximum bei ca. 41 dem Fixpunkt anzunähern, der bei $\bar{x}_2 = 19$ liegt. Im Vergleich mit den experimentellen Daten ergibt sich eine zufriedenstellende Simulation, da die Simulation sowohl qualitativ als auch quantitativ den Messergebnissen nahe kommt.
Im nächsten Schritt wird für den Wildtyp der entgegengesetzte Übergang simuliert, also der Übergang von Stickstoffüberschuss zu Stickstoffmangel, siehe Abbildung 5.8. Die Simulation startet in dem Fixpunkt, der sich aus den Differenzialgleichungen ergibt, die für stickstoffreiche Umgebungen gelten, d.h. $x_1(t_0) = 0$ und $x_2(t_0) = 19$. Die GlnK∼AMP-Konzentration steigt

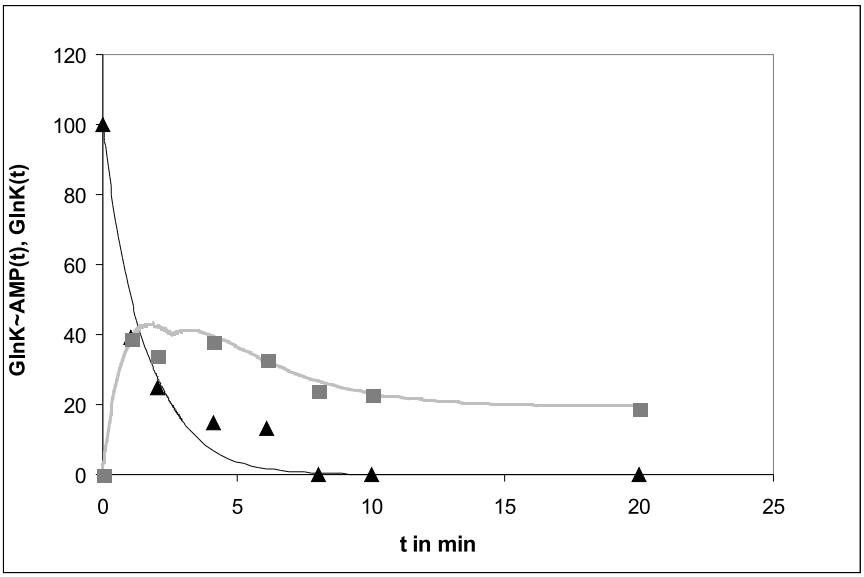


Abbildung 5.7: Wildtypsimulation und -experiment. Das Bakterium befindet sich vor den Messungen im Stickstoffmangel. Dann wird ein Stickstoffpuls zum Zeitpunkt t_0 gegeben. Die experimentellen Messungen für GlnK∼AMP sind durch schwarze Dreiecke und diejenigen für GlnK durch graue Quadrate dargestellt. Die Simulationen für GlnK∼AMP werden entsprechend durch eine schwarze, dünne Linie, die für GlnK durch eine graue, dicke Linie dargestellt.

in der ersten Minute zunächst sehr steil, anschließend etwas schwächer an und hat nach ungefähr 20 Minuten fast ihren Fixpunkt erreicht. Die GlnK-Konzentration dagegen fällt innerhalb kürzester Zeit ab und ist schon nach ca. einer Minute nahe dem Fixpunkt. Im Folgenden wird synthetisiertes GlnK direkt in GlnK∼AMP umgewandelt und schlägt sich nicht mehr in der GlnK-Konzentration nieder. Das qualitative Verhalten dieser Simulation ist erwartungsgemäß[2], jedoch ist anzumerken, dass x_1 länger benötigt, um seinen Fixpunkt zu erreichen, als in der Simulation des entgegengesetzten Experiments von Stickstoffmangel zu Stickstoffüberschuss.
Zwei weitere Simulationen für fiktive Bedingungen, die noch nicht in Experimenten umgesetzt wurden, finden sich in der Abbildung 5.9. Beide Simulationen beziehen sich auf den Wildtyp, der in stickstoffarmer Umgebung gezüchtet und dann in eine stickstoffreiche Umgebung verlagert wurde. Der Wildtyp ist in diesem Fall jedoch insofern eingeschränkt, dass die Proteasenaktivität verringert ist. Die Parameter $\gamma_{2,3}$ und $\lambda\gamma_{2,3}$ geben die Abbaurate von GlnK an, die durch AmtB und die Proteasen verursacht wird. In den Simulationen 5.9 wird nun eine eingeschränkte Proteasenaktivität angenommen und die oben genannten Parameter auf 10% bzw. 50% der ursprünglich geschätzten Werte gesetzt, also auf $\gamma_{2,3} = 0.053/min$ und $\lambda\gamma_{2,3} = 0.026/min$ in der ersten Simulation bzw. $\gamma_{2,3} = 0.27/min$ und $\lambda\gamma_{2,3} = 0.13/min$ in der zweiten Simulation der Abbildung 5.9. Die Anfangsbedingungen entsprechen in beiden Simulationen wieder dem

[2]Persönliche Mitteilung A. Burkovski, Erlangen

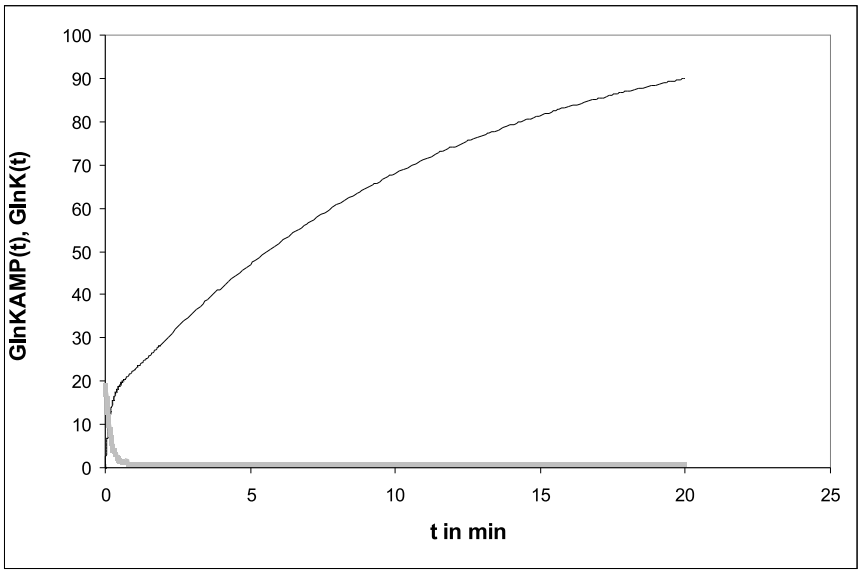


Abbildung 5.8: Wildtypsimulation. Das Bakterium befindet sich zunächst im Stickstoffüberschuss und wird zum Zeitpunkt t_0 in eine stickstoffarme Umgebung gebracht. Wieder wird die Simulation für GlnK∼AMP durch eine schwarze, diejenige für GlnK durch eine graue Linie dargestellt.

Fixpunkt der Differenzialgleichung für Stickstoffmangel, also $x_1(t_0) = 100$ und $x_2(t_0) = 1$. Ein ähnliches Verhalten wie in Abbildung 5.7 ist zu sehen, allerdings liegt der Fixpunkt von x_2 nun wesentlich höher, in der ersten Simulation nämlich bei $\bar{x}_2 = 58.2$. Auch das Maximum liegt höher als in Abbildung 5.7. Die GlnK-Konzentration erreicht nun ca. 84, geht also fast doppelt so hoch wie in der Simulation 5.7. Der Gleichgewichtspunkt von x_2 liegt in der zweiten Simulation bei $\bar{x}_2 = 30.7$, der maximale Wert bei ca. 59.
GlnK∼AMP zeigt in beiden Simulationen genau dasselbe Verhalten wie in der Simulation 5.7, da sich an den Differenzialgleichungen zur Beschreibung des Verhalten dieses Proteins nichts ändert.
Mithilfe der zweiten Differenzialgleichung aus Gleichung (5.26), die für GlnK im Wildtyp den Stickstoffüberschuss beschreibt, lässt sich das Verhältnis von reduzierter Proteasenaktivität zu ursprünglicher Proteasenaktivität folgendermaßen ausrechnen:

$$\frac{\gamma_{2,3,\text{reduziert}}}{\gamma_{2,3}} = \left(\frac{c_2}{\bar{x}_{2,\text{reduziert}}} - \gamma_2\right)\frac{1}{\lambda\gamma_{2,3}}. \tag{5.38}$$

Ebenso lässt sich bei gegebener Proteasenaktivität der Fixpunkt von GlnK berechnen:

$$\bar{x}_{2,\text{reduziert}} = \frac{c_2}{\lambda\gamma_{2,3,\text{reduziert}} - \gamma_2}. \tag{5.39}$$

In der folgenden Simulation, die in Abbildung 5.10 zu sehen ist, kann wieder ein Vergleich mit experimentellen Daten vorgenommen werden. Das Experiment besteht aus *amtB*-Mutanten,

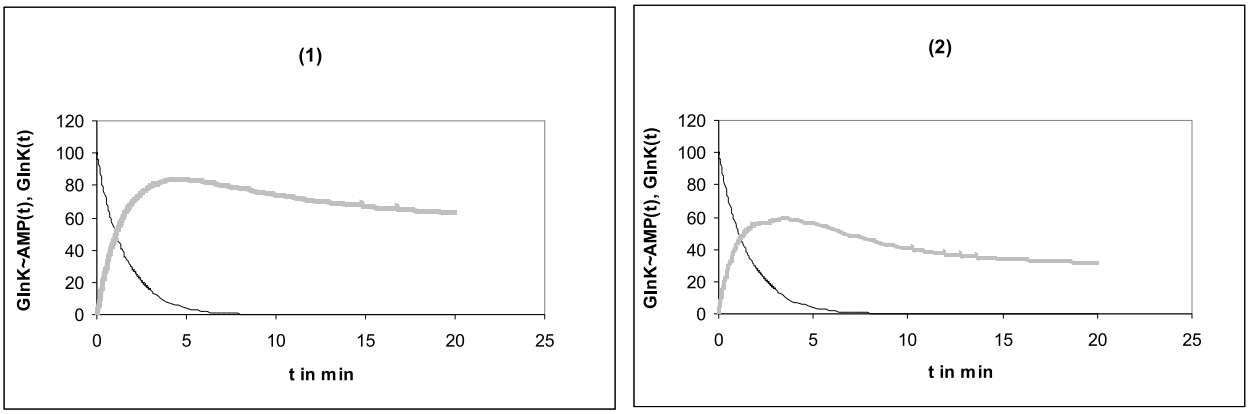


Abbildung 5.9: Wildtypsimulation. Das Bakterium befindet sich in stickstoffarmer Umgebung und zum Zeitpunkt t_0 wird ein Stickstoffpuls gegeben. Aufgrund verringerter Proteasenaktivität ist der Abbau von GlnK durch AmtB auf $\gamma_{2,3} = 0.053/min$ und $\lambda\gamma_{2,3} = 0.026/min$ in Simulation (1) bzw. $\gamma_{2,3} = 0.27/min$ und $\lambda\gamma_{2,3} = 0.13/min$ in Simulation (2) gesetzt worden, was einer Proteasenaktivität von 10% bzw. 50% der ursprünglichen Aktivität entspricht. Die schwarze Linie entspricht dem GlnK∼AMP-Verlauf, die graue Linie dem GlnK-Verlauf.

die zunächst in stickstoffarmer Umgebung gezüchtet und einem Stickstoffpuls zum Zeitpunkt t_0 ausgesetzt wurden. Bis auf den Zeitpunkt t_7 wurden diese Daten bisher nicht in die Parameterschätzung einbezogen. Für die Anfangsbedingungen der Simulation werden die Messungen der *amtB*-Mutante zugrunde gelegt, also $x_1(t_0) = 105.5$ und $x_2(t_0) = 1$. Da in der *amtB*-Mutante das Protein AmtB nicht existiert, wird der Abbau durch AmtB und Proteasen auf Null gesetzt, also $\gamma_{2,3} = \lambda\gamma_{2,3} = 0/min$. Übereinstimmend mit dem aus biologischer Sicht erwarteten Verhalten[3] fällt x_1 wie im Wildtyp innerhalb weniger Minuten auf Null und x_2 erreicht einen deutlich höheren Fixpunkt als im Wildtyp. Der Fixpunkt errechnet sich in dieser Simulation zu $\bar{x}_2 = 75$, der maximale Wert beträgt ungefähr 97. Verglichen mit den experimentellen Daten zeigt die GlnK-Simulation stark erhöhte Werte. Das Absinken von GlnK∼AMP in der Simulation ist dagegen im Vergleich zu den experimentellen Werten deutlich zu langsam. Diese Werte zeigen, dass schon nach einer Minute kein GlnK∼AMP mehr vorhanden ist.
Das zu langsame Absinken der GlnK∼AMP-Konzentration in der Simulation bedingt jedoch die zu hohe GlnK-Konzentration. Würden die Parameter derart verändert, dass GlnK∼AMP schneller abgebaut würde, so hätte man als Resultat auch einen niedrigeren GlnK-Verlauf. Solch eine Veränderung der Parameter beruhen auf der Vermutung[4], dass ein weiterer Abbaumechanismus von GlnK∼AMP in der *amtB*-Mutante involviert sein könnte. Die Gleichung von GlnK∼AMP

[3]Persönliche Mitteilung A. Burkovski, Erlangen
[4]Persönliche Mitteilung A. Burkovski, Erlangen

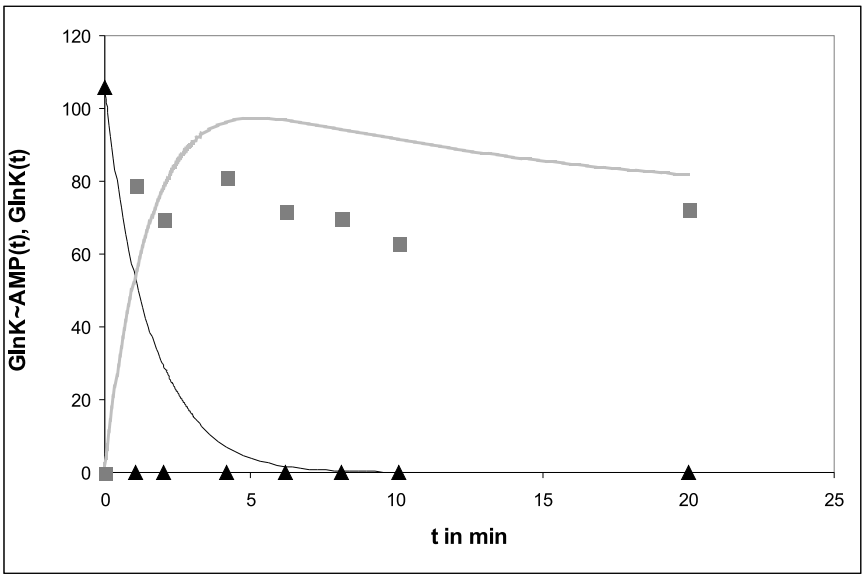


Abbildung 5.10: *amtB*-Mutantensimulation und -experiment. Im Experiment wird eine *amtB*-Mutante in stickstoffarmer Umgebung gezüchtet, die einen Stickstoffpuls zum Zeitpunkt t_0 erfährt. In der Simulation wird dementsprechend $\gamma_{2,3} = 0$ gesetzt. Die experimentellen Messungen für GlnK∼AMP sind durch schwarze Dreiecke und diejenigen für GlnK durch graue Quadrate dargestellt. Die Simulationen für GlnK∼AMP werden entsprechend durch eine schwarze Linie, die für GlnK durch eine graue Linie dargestellt.

im Stickstoffüberschuss,

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = -k_1 x_1(t) + rf_{1,1}(x_1(t), \theta_{1,1,1}, \theta_{1,1,2}) - \gamma_1 x_1(t),$$

kann durch Erhöhung der zwei Parameter k_1 und γ_1 zu einem deutlich höheren GlnK∼AMP-Abbau führen. Beide Parameter entsprechen biologischen Prozessen: Eine Erhöhung von k_1 entspricht einer höheren Deadenylylierungsrate von GlnK∼AMP in GlnK. Dies bedeutet, dass durch einen Signaltransduktionsweg das Protein GlnD, das für die Deadenylylierung verantwortlich ist, die Information erhält, mehr GlnK∼AMP zu demodifizieren. Eine Erhöhung von γ_1 entspricht einem weiteren Abbaumechanismus von GlnK∼AMP, der proportional zur Konzentration von GlnK∼AMP dieses Protein zersetzt.
Um herauszufinden, welcher der beiden Mechanismen verantwortlich für die schnellere Abnahme der GlnK∼AMP-Konzentration ist, werden im Folgenden die beiden Parameter k_1 und γ_1 variiert. Tabelle 5.7 listet die verwendeten Werte auf.
In Abbildung 5.11 sind vier verschiedene Simulationen zu sehen, in denen der Parameter k_1 durch die in Tabelle 5.7 aufgelisteten Werte von $\tilde{k}_1$ ersetzt wurde. In der ersten Simulation gilt $k_1 = \tilde{k}_1 = 0.61$, in der zweiten $\tilde{k}_1 = 0.97$, in der dritten $\tilde{k}_1 = 1.42$ und in der letzten $\tilde{k}_1 = 1.83$. Es lässt sich erkennen, dass die Abnahme der GlnK∼AMP-Konzentration umso schneller vor sich

Tabelle 5.7: Die aufgelisteten Parameter $\tilde{k}_1$ bzw. $\tilde{\gamma}_1$ ersetzen in den folgenden *amtB*-Mutanten-Simulationen die Parameter k_1 bzw. γ_1. Um einen vergleichbaren GlnK∼AMP-Abbau zu erhalten, wurde nach Wahl eines der beiden Parameter $\tilde{k}_1$ oder $\tilde{\gamma}_1$ der zweite Parameter derart gewählt, dass $\tilde{k}_1+\gamma_1 = k_1+\tilde{\gamma}_1$ gilt. In der zweiten Zeile wird beispielsweise ein fünffach erhöhtes γ_1 gewählt, also $\tilde{\gamma}_1 = 5\cdot\gamma_1$. Da $k_1+\tilde{\gamma}_1 = 1.06$ gilt, wird $\tilde{k}_1 = 1.06-0.09 = 0.97$ gesetzt.

	$\tilde{k}_1$	$\tilde{\gamma}_1$	$\tilde{k}_1+\gamma_1 = k_1+\tilde{\gamma}_1$
(1)	0.61 ($=1\cdot k_1$)	0.09 ($=1\cdot\gamma_1$)	0.7
(2)	0.97 ($=1.59\cdot k_1$)	0.45 ($=5\cdot\gamma_1$)	1.06
(3)	1.42 ($=2.33\cdot k_1$)	0.9 ($=10\cdot\gamma_1$)	1.51
(4)	1.83 ($=3\cdot k_1$)	1.31 ($=14.56\cdot\gamma_1$)	1.92

geht, je höher k_1 ist, jedoch dass dann auch der GlnK-Verlauf umso höher ist. Dies liegt an der Addition von k_1x_1 bei Stickstoffüberschuss in der Gleichung für GlnK:

$$\frac{\mathrm{d}x_2(t)}{\mathrm{d}t} = k_1x_1(t)+c_2-\gamma_2x_2(t)-s_{2,3}(x_2(t),x_1(t)).$$

Ebenfalls zu beobachten ist die Tatsache, dass das Maximum der GlnK-Kurve umso früher erreicht wird, je größer k_1 ist.
Durch alleinige Erhöhung des Parameters k_1 lassen sich die Simulationen also nicht an die Messungen anpassen.
In Abbildung 5.12 wird nun der Parameter γ_1 variiert. In der ersten Simulation gilt $\gamma_1 = \tilde{\gamma}_1 = 0.09$, in der zweiten $\tilde{\gamma}_1 = 0.45$, in der dritten $\tilde{\gamma}_1 = 0.9$ und in der vierten $\tilde{\gamma}_1 = 1.31$. Die Simulation mit fünffach erhöhter Abbaurate $\gamma_1 = 0.45$ ist zwar zufriedenstellend für GlnK, jedoch ist das Absinken der GlnK∼AMP-Konzentration zu langsam. Die Simulation mit zehnfach erhöhter Abbaurate $\gamma_1 = 0.9$ zeigt zwar ein besseres Verhalten für die GlnK∼AMP-Kurve, jedoch eine deutlich verschlechterte GlnK-Kurve. Je höher γ_1 wird, desto langsamer nähert sich GlnK seinem Fixpunkt an, desto schneller geht jedoch der GlnK∼AMP-Abbau vonstatten. Auch eine Erhöhung von γ_1 bringt demnach nicht vollends zufriedenstellende Resultate, auch wenn die Simulation mit fünffach erhöhtem γ_1 schon deutlich besser die Messdaten widerspiegelt.
Da eine Erhöhung von k_1 eine Erhöhung des GlnK-Verlaufs, eine Erhöhung von γ_1 jedoch eine Herabsetzung desselben nach sich zieht, ist eine kombinierte Veränderung beider Parameter zu betrachten.
In Abbildung 5.13 sieht man nun eine Simulation, in der sowohl k_1 als auch γ_1 verändert wurden. Beide Parameter sind wesentlich höher gewählt als in der Simulation des Wildtyps, um einen schnellen GlnK∼AMP-Abbau zu erreichen. Der Parameter $k_1 = 1.83$ entspricht seinem 3-fachen ursprünglichen Wert, der Parameter $\gamma_1 = 0.9$ seinem zehnfachen ursprünglichen Wert. In weiteren Simulationen ist festzustellen, dass γ_1 ungefähr 3n-fach erhöht werden muss, wenn k_1 n-fach erhöht wurde, um eine Simulation zu erhalten, die die Messdaten widerspiegelt. Je höher die beiden Werte k_1 und γ_1 werden, desto steiler ist der Abfall der GlnK∼AMP-Kurve und desto schneller nähert sich die GlnK-Kurve ihrem Fixpunkt an.
Die Simulationen der *amtB*-Mutante legen demnach folgende Schlussfolgerung nahe: In einer

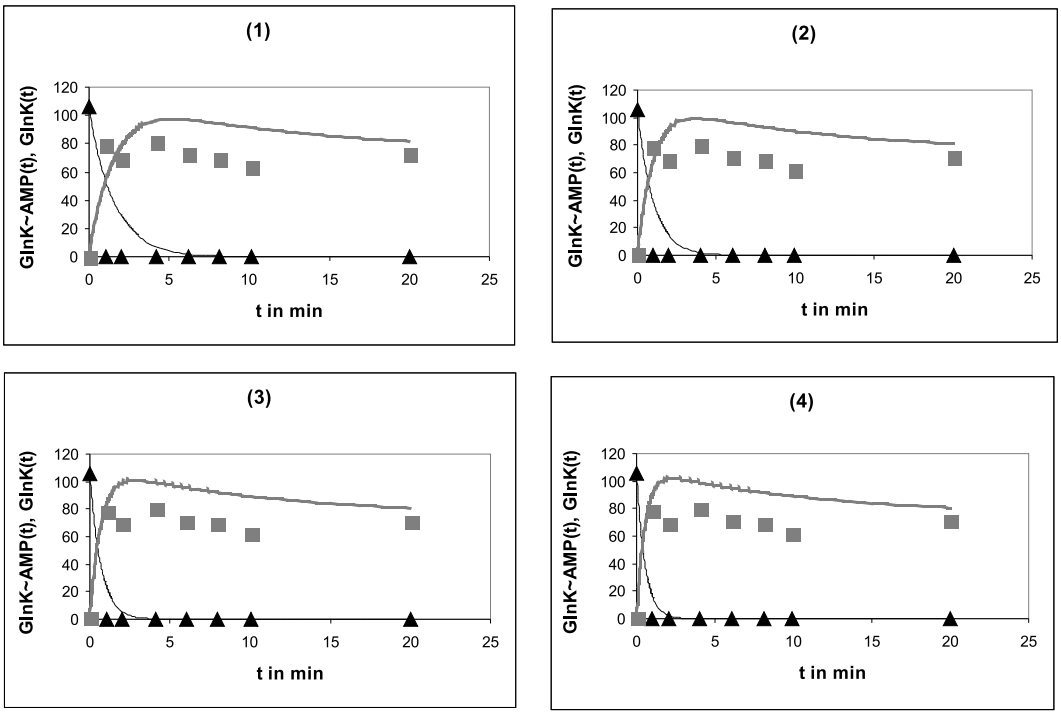


Abbildung 5.11: Verhalten der GlnK∼AMP- und GlnK-Kurve in einer *amtB*-Mutantensimulation bezüglich variiertem Parameter k_1. In (1) wird $k_1 = 0.61$, in (2) $k_1 = 0.97$, in (3) $k_1 = 1.42$ und in (4) $k_1 = 1.83$ verwendet. Je höher k_1 wird, desto schneller fällt die GlnK∼AMP-Kurve und desto schneller steigt die GlnK-Konzentration auf ihr Maximum an. Die schwarze Linie entspricht wieder dem GlnK∼AMP-Verlauf, die graue Linie dem GlnK-Verlauf. Schwarze Dreiecke kennzeichnen GlnK∼AMP-Messdaten, graue Rechtecke die GlnK-Messdaten.

amtB-Mutante ist eine alleinige Erhöhung der Deadenylylierungsrate eher nicht zu erwarten. Stattdessen können wir einen zusätzlichen Abbaumechanismus in Verbindung mit einer erhöhten Deadenylylierungsrate vermuten.

5.5 Bewertung des Modells und Ergebnisse

5.5.1 Bewertung des Modells

Die Diskussion des Modells teilt sich auf in eine Diskussion über die geschätzten Parameter und in eine Beurteilung der durchgeführten Simulationen.
Zunächst ist zu bemerken, dass alle Parameter wie erwartet größer als Null sind, also kein Wert dem zugrunde liegenden Modell der biologischen Prozesse widerspricht. Des Weiteren ist $\lambda\gamma_{2,3}<\gamma_{2,3}$, was ebenfalls notwendig für eine biologische Interpretation der Differenzialglei-

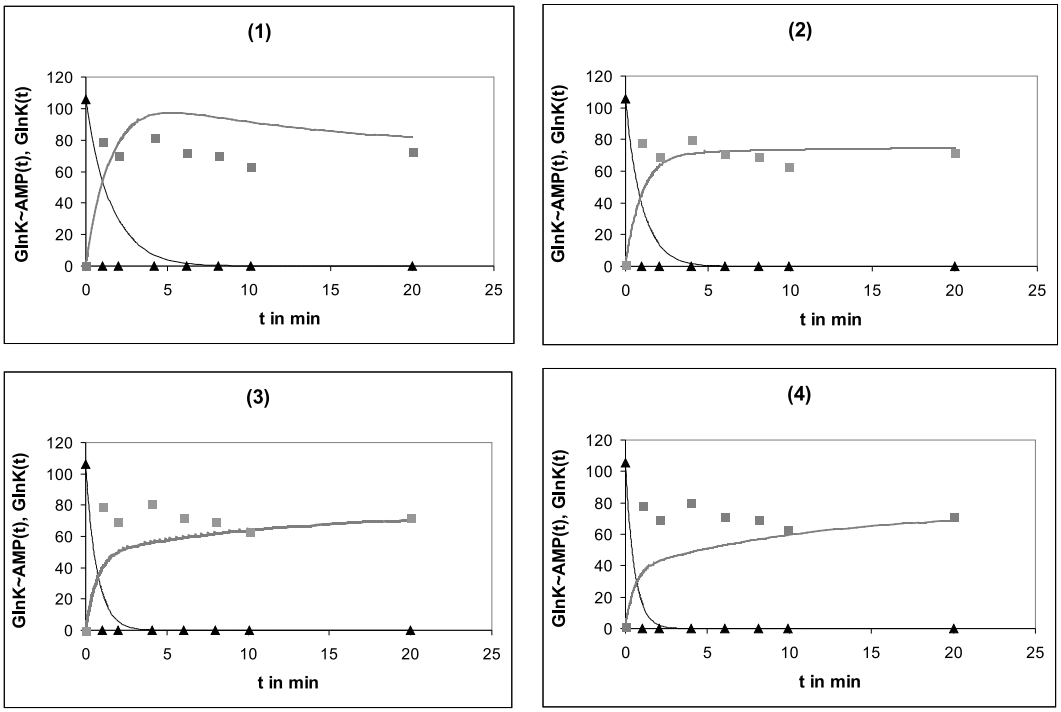


Abbildung 5.12: Verhalten der GlnK∼AMP- und GlnK-Kurve in einer *amtB*-Mutantensimulation bezüglich variiertem Parameter γ_1. In (1) wird $\gamma_1 = 0.09$, in (2) $\gamma_1 = 0.45$, in (3) $\gamma_1 = 0.9$ und in (4) $\gamma_1 = 1.31$ verwendet. Je höher γ_1 ist, desto schneller fällt die GlnK∼AMP-Kurve und desto langsamer nähert sich die GlnK-Kurve für $\gamma_1 > 0.45$ ihrem Fixpunkt an. Die schwarze Linie entspricht wieder dem GlnK∼AMP-Verlauf, die graue Linie dem GlnK-Verlauf. Schwarze Dreiecke kennzeichnen GlnK∼AMP-Messdaten, graue Rechtecke die GlnK-Messdaten.

chungen ist. Darüber hinaus erkennt man, dass $k_2 \gg k_1$ gilt, das heißt, die Rate, mit der die Modifizierung des Proteins GlnK im Fall eines Stickstoffmangels vorangetrieben wird, ist wesentlich größer als die Rate für die Demodifizierung im Fall eines Stickstoffüberschusses. Hierdurch wird es *C. glutamicum* möglich, auf Stickstoffmangel schnell zu reagieren, was notwendig für die Überlebensfähigkeit des Bakteriums ist. Außerdem sind wie erwartet hauptsächlich die Interaktionen mit AmtB und die Aktivität von Proteasen verantwortlich für den Abbau von GlnK, da die Grundabbaurate γ_2 nur etwa ein Fünftel des Abbaus durch die Interaktion mit AmtB und die nachfolgende Proteolyse, $\gamma_{2,3}$, beträgt. Das Ergebnis der Parameterschätzung liefert also zufriedenstellende Ergebnisse, die biologisch plausibel erscheinen. Hier erkennt man einen Vorteil des entwickelten Modells gegenüber allgemeineren Ansätzen. Die einzelnen Parameter haben jeweils eine biologische Bedeutung, die für die Entwicklung weiterer Experimente hilfreich sein kann.

Im Folgenden werden die durchgeführten Simulationen beurteilt. In der Simulation aus Abbil-

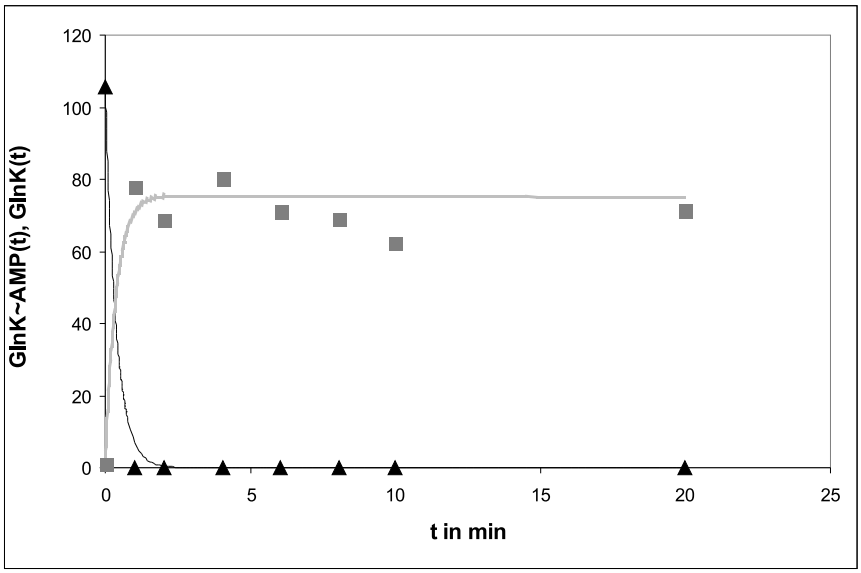


Abbildung 5.13: *amtB*-Mutantensimulation und -experiment. Die Parameter k_1 und γ_1 wurden durch $\tilde{k}_1 = 1.83$ und $\tilde{\gamma}_1 = 0.9$ ersetzt. Die schwarze Linie entspricht wieder dem GlnK∼AMP-Verlauf, die graue Linie dem GlnK-Verlauf. Schwarze Dreiecke kennzeichnen GlnK∼AMP-Messdaten, graue Rechtecke die GlnK-Messdaten.

dung (5.7) konnte gezeigt werden, dass das Modell mit den entsprechenden Bedingungen das für die Parameterschätzung verwendete Experiment zufriedenstellend simulieren konnte. Dies ist eine grundlegende Eigenschaft eines Modells, bevor es für weitere Simulationen verwendet werden kann. In den Simulationen, die in den Abbildungen (5.8) und (5.9) gezeigt sind, wurden veränderte Bedingungen festgelegt, so dass Prognosen für das Verhalten der GlnK- und GlnK∼AMP-Konzentrationen unter diesen Bedingungen gestellt werden konnten. Die Ergebnisse können zwar nur in Experimenten verifiziert werden, sind aus biologischer Sicht jedoch widerspruchsfrei und plausibel[5].

Die Bedingungen für die Simulation 5.10 entsprechen denen eines bereits durchgeführten Experiments. Das qualitative Verhalten der beiden Komponenten ist erwartungsgemäß, jedoch fällt x_1 langsamer als im Experiment. Dies unterstützt die Annahme, dass ein weiterer Abbaumechanismus von GlnK∼AMP existiert, der auch in anderen Experimenten bereits vermutet worden ist[6].

In den Simulationen, die in den Abbildungen 5.11, 5.12 und 5.13 gezeigt werden, kann durch Variieren der beiden Parameter, die für zwei Abbaumechanismen stehen, gezeigt werden, dass durch Erhöhung des Parameters, der die Umwandlung von GlnK∼AMP in GlnK beschreibt, keine zufriedenstellende Simulation erreicht werden kann, dass jedoch die kombinierte Erhöhung der zwei Parameter zu akzeptablen Simulationen führt.

[5] Persönliche Mitteilung A. Burkovski, Erlangen
[6] Persönliche Mitteilung A. Burkovski, Erlangen

5.5.2 Diskussion der Modellauswahl

Bei der Bewertung eines Modells wird üblicherweise ein Vergleich zwischen dem ausgewählten Modell und anderen Modellierungsansätzen gezogen, um die Vor- und Nachteile des ausgewählten Modells diskutieren zu können. Bei solch einer spezifischen Modellierung wie in diesem Kapitel, die die einzelnen Parameter entsprechend der zugrunde liegenden Prozesse wählt, muss jedoch auf einen Vergleich verzichtet werden. Das Modell wurde so weit vereinfacht, dass die Parameter schätzbar sind. Eine Erweiterung des Modells würde weitere Messdaten, beispielsweise für die ebenfalls an den Prozessen beteiligten Proteine AmtB und GlnD, erfordern. Eine Vereinfachung des Modells, beispielsweise durch Stufenfunktionen anstelle der verwendeten stückweise linearen Funktionen, könnte die im Experiment beobachteten zwei Maxima von GlnK nicht abbilden. Insofern ist das hier gezeigte Modell so detailliert wie möglich, aber auch so vereinfacht wie nötig.

5.5.3 Ergebnisse

Eines der Hauptresultate dieser Modellierung ist allgemein auf verschiedene Modellierungsansätze übertragbar: Trotz sehr kurzer Zeitreihen ist eine mathematische Modellierung mit Differenzialgleichungssystemen durchführbar, wenn detailliertes Wissen über die einzelnen Komponenten und die biologischen Prozesse im Bakterium vorliegt. Üblicherweise werden für quantitative Modelle mehr Parameter verwendet als in Modellen, die rein qualitative Aussagen machen können. Demzufolge werden für Modelle, denen Differenzialgleichungen zugrunde liegen, meist mehr Daten zur Parameterschätzung benötigt als für qualitative Modelle. In diesem Kapitel wurde jedoch gezeigt, dass Differenzialgleichungsmodelle trotz sehr geringer Datenlage möglich sind, wenn detaillierte biologische Informationen in das Modell mit eingebunden werden können. Beispielsweise kann man bei *C. glutamicum* davon ausgehen, dass die Proteinkonzentrationen im betrachteten Experiment nach 20 Minuten ihren Fixpunkt erreicht haben. Dieses Wissen muss in einen mathematischen Rahmen eingebettet werden, hier beispielsweise, dass die Differenzialgleichungen die Werte bei 20 Minuten exakt erfüllen müssen, wohingegen andere Messwerte nur zur Fehlerminimierung hinzugezogen werden.
Die Schlussfolgerung aus diesem Resultat ist also, dass bei geringer Datenlage so viele Quellen wie möglich in die Modellierung miteingebunden werden sollten. Dieser Trend lässt sich auch bei ausreichender Datenlage beobachten.
Speziell aus dem hier vorgestellten Modell können ebenfalls einige Schlussfolgerungen gezogen werden. Abhängig von den Anfangs- und Nebenbedingungen können die Zeitverläufe der Proteinkonzentrationen von GlnK und GlnK$\sim$AMP simuliert werden. So sind beispielsweise Halbwertszeiten und Maximalwerte berechenbar, genauso wie Zeitspannen bis zum Erreichen des Fixpunktes. Werte für Abbau- und Umwandlungsraten lassen sich direkt aus den Differenzialgleichungen ablesen, so dass spezielle Parameter miteinander verglichen werden können, um ein tieferes Verständnis des Systems zu erlangen. Beispielsweise können in dem hier vorgestellten Modell die beiden Abbauraten von GlnK verglichen werden, welche den Schluss zulassen, dass GlnK hauptsächlich über AmtB und die Proteasen abgebaut wird, im Zytoplasma dagegen nur ein geringer Abbau von GlnK existiert. Diese Information ist nicht mit in die Parameterschätzung eingeflossen, ist jedoch schon bekannt. Des Weiteren kann durch den Vergleich der

beiden Umwandlungsraten gefolgert werden, dass die Umwandlung von GlnK in GlnK∼AMP schneller ist als die entsprechende Rückumwandlung. Ebenfalls durch Umformen der Differenzialgleichungen erhält man zwei Gleichungen, die erlauben, aufgrund von Proteasenaktivitäten die GlnK-Konzentration im Fixpunkt auszurechnen bzw. umgekehrt die Proteasenaktivität aufgrund der GlnK-Konzentration im Fixpunkt zu bestimmen. In den *amtB*-Mutantensimulationen ist außerdem zu erkennen, dass beide vermuteten Mechanismen am schnelleren Abbau beteiligt sein könnten, da die Hinzunahme eines der Mechanismen alleine keine zufriedenstellende Simulation ermöglicht. Eine Kombination von schnellerer Umwandlung von GlnK∼AMP in GlnK und ein weiterer Abbaumechanismus könnten die Messdaten jedoch erklären. Hier konnten aus den Simulationen also Anregungen für weitere Experimente gewonnen werden.

Kapitel 6

Anwendung des Modells auf das DNA-Reparatursystem des *Mycobacterium tuberculosis*

Die Infektionskrankheit Tuberkulose wird durch den Erreger *Mycobacterium tuberculosis* (Mtb) ausgelöst. In den meisten Fällen wird die Lunge infiziert, jedoch können auch Niere, Wirbelsäule und Gehirn befallen sein. In der Regel bricht die Tuberkulose aber nur bei 5-10% der Infizierten aus. Allerdings führt eine Schwächung des Immunsystems, beispielsweise nach Organtransplantationen, bei Einnahme gewisser Rheumamedikamente oder bei HIV-Infektionen, zu einem starken Anstieg der Tuberkulose. Nach WHO-Schätzungen verursachte Tuberkulose im Jahr 2004 weltweit 1.7 Millionen Todesfälle. Das Bakterium wurde bereits 1882 von Koch isoliert. Der Prozentsatz der MDR-Erreger[1], bei denen es sich um Bakterienstämme handelt, die gegenüber mehreren der verfügbaren Medikamente resistent sind wächst an (World Health Organization). Seit den 90er Jahren beobachtet man einen leichten Zuwachs an Neuinfizierten, so dass die Entwicklung neuer Medikamente gegen den Erreger notwendig ist. Hierbei ist ein Verständnis seiner Funktionsweise auf molekularer Ebene hilfreich, um in Regulierungsprozesse auf dieser Ebene eingreifen zu können. Die vollständige Entschlüsselung der Genomsequenz im Jahr 2000 lieferte den Grundstein für solche Forschungen. Im Gegensatz zu vielen anderen Organismen sind die Regulierungsmechanismen jedoch noch teilweise unbekannt. Dies liegt unter anderem auch an der langsamen Fortpflanzung des Bakteriums, das sich nur alle 16-20 Stunden teilt, so dass Experimente im Labor wesentlich länger dauern als mit Modellorganismen wie z. B. *Escherichia coli*, der sich unter optimalen Bedingungen alle 20 Minuten teilt.
Einer dieser noch nicht vollständig aufgeklärten Regulierungsmechanismen ist das DNA-Reparatursystem, das die Reparatur beschädigter DNA reguliert. Wie auch in *E.coli* gibt es in diesem System sogenannte SOS-Gene, die von den beiden Proteinen RecA und LexA reguliert werden. Man hat in *recA*-Mutanten von Mtb jedoch beobachtet, dass im Gegensatz zu *E. coli* bei defekter DNA einige SOS-Gene trotzdem stärker exprimiert werden, so dass ein alternativer Mechanismus existieren muss, der nicht über RecA und LexA funktioniert.
In diesem Kapitel wird ein Teil des DNA-Reparatur-Systems mit Differenzialgleichungen modelliert. Hierzu wird zunächst die in Kapitel 3.1 beschriebene Methode zur Komponentenauswahl verwendet. Für die Schätzung der zugehörigen Modellparameter werden neben Expressionsdaten auch Wechselwirkungsinformationen verwendet. Aus den Simulationen geht hervor,

[1] MDR=Multi-drug-resistant

dass das Gen *Rv2719c* möglicherweise eine wichtige Rolle im DNA-Reparatursystem spielt und als Verbindungsgen zwischen den SOS-Genen, die von LexA und RecA reguliert werden, und dem in Mtb vermuteten alternativen Mechanismus agieren könnte.
Die Ergebnisse der Kapitel 6.1 bis 6.5 wurden in Zusammenarbeit mit Nicole Radde und Christian Forst in Radde u. a. (2006) veröffentlicht.
Anschließend wird in Kapitel 6.6 auch die zweite Methode der Komponentenauswahl auf die vorhandene Datengrundlage angewendet und die Ergebnisse bezüglich der ersten Komponentenauswahl diskutiert.

6.1 Qualitative Beschreibung des DNA-Reparatursystems

Das DNA-Reparatursystem wird eingeschaltet, wenn es zu einer Schädigung der DNA in einer Art und Weise kommt, dass die DNA anschließend zumindest teilweise einzelsträngig vorliegt. In *E. coli* wird das Reparatursystem bereits seit einigen Jahrzehnten untersucht (Little u. Mount, 1982; Walker, 1996), jedoch können nicht alle Erkenntnisse auf Mtb übertragen werden. Genauso wie in *E. coli* funktioniert jedoch das SOS-System von Mtb, deren Hauptkomponenten die Gene *recA* und *lexA* sowie deren Produkte RecA und LexA sind. Ungefähr 35-40 Gene mit SOS-Boxen werden im Zusammenspiel der beiden Komponenten reguliert. Dabei bindet das Protein RecA zunächst an einzelsträngige DNA und verändert dadurch die Struktur des Proteins LexA derart, dass es nicht mehr an SOS-Boxen binden kann. Diese SOS-Boxen sind spezifische Bindungsstellen in den Promotoren der SOS-Gene. Bindet LexA an solch eine Box, so wird die Expression des jeweiligen Gens inhibiert. Daraus folgt, dass eine Zerstörung der DNA eine stärkere Expression der SOS-Gene zur Folge hat. Auch *recA* und *lexA* besitzen SOS-Boxen. Somit liegen nach einer erfolgreichen Reparatur der DNA genügend RecA- und LexA-Moleküle in der Zelle vor, um die Expression der Gene ähnlich schnell abzuschalten wie der Prozess in Gang gesetzt wurde.

Rand u. a. (2003) fanden heraus, dass im Gegensatz zu *E. coli* in Mtb ein weiterer Mechanismus existieren muss, der die Transkriptmenge einiger der SOS-Gene erhöht (siehe dazu Abbildung 6.1). In der Abbildung ist zu erkennen, dass es einige Gene mit vorhergesagter SOS-Box wie z. B. *ruvC* gibt, die trotz des Ausschaltens von *recA* im Fall einer DNA-Schädigung ein höheres Transkriptniveau aufweisen als bei Vorliegen von unbeschädigter DNA. Andere Gene wie z. B. *linB* zeigen jedoch nur im Wildtyp eine erhöhte Transkriptmenge. Des Weiteren gibt es eine Reihe von Genen, die eine Funktion bei der DNA-Reparatur übernehmen und im Experiment in der *recA*-Mutante ein genauso hohes Transkriptniveau haben wie im Wildtyp. Ziel der Modellbildung ist es, zusätzliche Komponenten des DNA-Reparatursystems zu erkennen, die neben einer wichtigen Rolle bei der Regulierung von SOS-Genen möglicherweise auch im Zusammenhang mit dem alternativen Mechanismus stehen.

6.2 Datengrundlage und Datenvorverarbeitung

Boshoff u. a. (2004) führten zahlreiche Experimente durch, in denen Mtb verschiedenen Medikamenten ausgesetzt wurde. Zugang zu den Experimenten hat man über NCBI's Gene Ex-

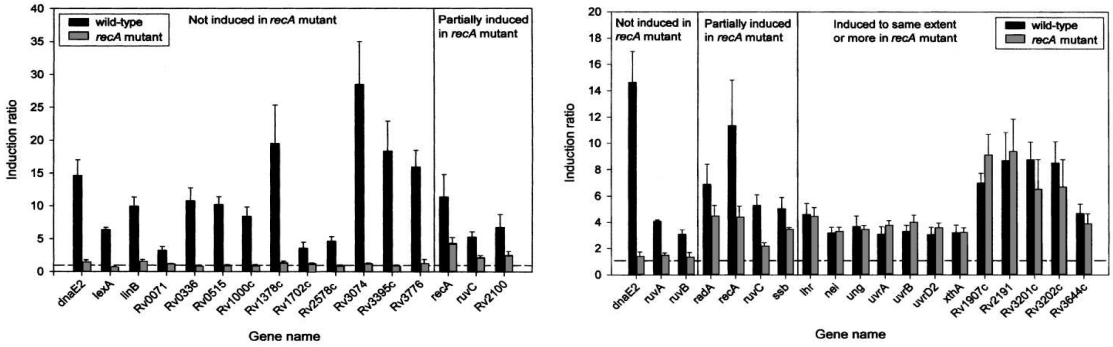


Abbildung 6.1: Diese beiden Abbildungen sowie die entsprechenden Experimente stammen aus Rand u. a. (2003). Induktion durch DNA-Zerstörung von Genen im Wildtyp (schwarze Balken) und in der *recA*-Mutante (graue Balken) von *M. tuberculosis*. Links zeigt Gene mit vorhergesagten SOS-Boxen, rechts Gene mit vorhergesagter Funktion im DNA-Reparatursystem. An der x-Achse sind die Gene aufgetragen, an der y-Achse die Induktionsverhältnisse für Wildtyp und *recA*-Mutante. Die gestrichelte Linie zeigt ein Induktionsverhältnis von 1, was bedeutet, dass keine Induktion stattgefunden hat. Die gezeigten Werte sind Mittelwerte von Messungen dreier unabhängiger Kulturen, denen jeweils drei Proben entnommen wurden. Ausreißer wurden vorher entfernt. Die Fehlerbalken zeigen die Standardabweichungen.

pression Omnibus[2]. Der Datensatz beinhaltet unter anderem 16 Experimente, in denen Mtb mit 0.2μg Mitomycin behandelt wird. Mitomycin führt zur Schädigung einzelner Basen der DNA und folglich auch zu einer stärkeren Expression der SOS-Gene. Nachdem Mtb Mitomycin zugesetzt wurde, folgten Transkriptanalysen sämtlicher Gene des Mtb-Genoms nach 0.33h, 0.75h, 1.5h, 2h, 4h, 6h, 8h und 12h. Die auf der GEO-Plattform verwendeten Namen der Experimente finden sich im Anhang A.1. Für die ersten drei Zeitpunkte existiert jeweils nur eine Messung, zum Zeitpunkt 2h wurden vier Experimente gemacht, zum Zeitpunkt 4h fünf, für 6h und 12 h nur eine, und für 8h zwei. Im Folgenden werden die acht Zeitpunkte mit $t_0, ..., t_7$ bezeichnet.
Für jedes Experiment liegen die Rohdaten wie in Tabelle 6.1 skizziert vor. Die Tabellen zu jedem Experiment beinhalten jeweils 4608 Messungen, wobei jedoch nicht nur Gene gemessen wurden, sondern z. B. auch ORFs[3] oder Kontrolloligomere, also auch eventuell nicht genkodierende Sequenzen. Ohne weitere Normalisierungen oder Filter zu verwenden, kann die Tabelle zu einer ersten Abschätzung verwendet werden, indem man für beide Farben jeweils den Hintergrund von dem Signal subtrahiert und das Verhältnis der erhaltenen Intensitäten betrachtet. Dies reicht für einen ersten Eindruck der Daten, jedoch werden die Daten vor der weiteren Analyse noch Normalisierungs- und Filtermethoden unterzogen. Für die Weiterverarbeitung der Tabellen wurden die BRB Array Tools verwendet. Diese Software wird dazu verwendet, statistische Analysen vorzunehmen, insbesondere für Normalisierungen von Mikroarrays, und ist

[2]http://www.ncbi.nlm.nih.gov/geo, GEO platform accession number ist GPL1396
[3]ORF=Open Reading Frame

Tabelle 6.1: Datengrundlage. Hier als Beispiel die ersten drei Zeilen eines Experiments, in dem Mtb mit Mitomycin behandelt wurde. Zeitpunkt der Messung ist vier Stunden nach Zugabe von Mitomycin. Rotes Signal und roter Hintergrund (Bkg) gehören zum behandelten Mtb, grünes Signal und grüner Hintergrund zum unbehandelten Mtb (Kontrollexperiment), so dass sich also berechnen lässt, um wieviel höher oder niedriger die Genexpression im behandelten Bakterium im Vergleich zum unbehandelten ist. 'Size' bezeichnet die Größe entsprechend der Anzahl der Pixel des Signals.

Red Signal	Green Signal	Red Bkg	Green Bkg	Size
214	1381	129	66	99
270	1202	125	67	100
316	1325	132	69	100
⋮	⋮	⋮	⋮	⋮

frei verfügbar[4]. Zunächst wird der Intensitätsfilter verwendet, der alle Gene, deren Signal nach Abzug des Hintergrundes ≤ 10 ist, herausfiltert. Die Tabellenspalte 'Size' wird in unserem Fall nicht als Filtermerkmal verwendet. Anschließend wird das Verhältnis $\frac{s_r - h_r}{s_g - h_g}$ berechnet, wobei s_r bzw. s_g für rotes bzw. grünes Signal und h_r bzw. h_g für roten bzw. grünen Hintergrund stehen. Die Normalisierung geschieht ebenfalls für jedes Array einzeln, indem zunächst eine Transformation der Daten über den Logarithmus zur Basis 2 erfolgt und anschließend der Median des jeweiligen Arrays von jedem Wert subtrahiert wird. Dadurch erreicht man, dass die logarithmierten Quotienten jedes normalisierten Arrays einen Median von Null besitzen. Im nächsten Schritt werden nur noch diejenigen Zeilen weiter verwendet, deren 'WellID' einer Rv-Nummer zugeordnet werden kann. Zuletzt werden all diejenigen Rv-Nummern aus den Tabellen entfernt, für die über alle 16 Experimente gesehen über 60% der Daten fehlen. Von Susanne Motameny wurde mithilfe eines Quade-Tests für diejenigen Arrays eine Auswahl getroffen, die jeweils den gleichen Zeitpunkt nach Zugabe von Mitomycin gemessen haben. Des Weiteren wurde von ihr ein Nächster-Nachbar-Ansatz verwendet, um die fehlenden Daten zu ersetzen. Dies führt dazu, dass in der weiteren Analyse auf 3658 Gene zurückgegriffen werden kann. Für die Zeitpunkte t_3, t_4 und t_6 wurde anschließend der Mittelwert gebildet, siehe auch Tabelle 6.2.
Wie in Abbildung 6.2 zu sehen, reagiert das *recA-lexA*-System sehr langsam auf das Medikament. Erst nach zwei Stunden lässt sich ein Anwachsen der Expressionswerte erkennen, die nach sechs bzw. acht Stunden den höchsten gemessenen Wert erreichen. Auch nach 12 Stunden ist die Antwort des *recA-lexA*-Systems noch nicht abgeklungen. Das Gen *recA* ist nach 12 Stunden immer noch über zehnfach stärker exprimiert als im Kontrollexperiment und auch das Gen *lexA* ist zu diesem Zeitpunkt immerhin noch fast dreieinhalbfach höher als in der Kontrolle. Im Anhang findet sich die Datenmatrix der logarithmierten, normalisierten und gemittelten Daten für die Gene *recA, lexA, ruvC* und *linB*.

[4] http://linus.nci.nih.gov/brb/. Aus rechtlichen Gründen sei an dieser Stelle auch auf die Verfasser der Software verwiesen: Analysen wurden mit den BRB Array Tools durchgeführt, welche von Dr. Richard Simon und Amy Peng Lam entwickelt wurden.

Tabelle 6.2: Verwendete zurücktransformierte Datenmatrix. Hier als Beispiel die ersten drei Rv-Nummern der insgesamt 3658 Gene.

Genname	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
Rv0001	0.8631	1.7537	1.5328	0.9767	0.9644	1.3659	1.3161	1.1241
Rv0002	0.6697	0.5413	0.6614	0.7923	0.5170	0.8298	0.8321	0.8472
Rv0003	0.7806	1.3328	0.9753	1.3305	0.6946	0	1.0860	0.7595
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

6.3 Auswahl der Komponenten

Die Auswahl der Komponenten wird mithilfe k-kürzester Wege getroffen, siehe dazu Kapitel 3.1. Zunächst wird die Menge $\mathscr{Q}$ der Quellgene bestimmt, zu der die Gene *recA* und *lexA* als Hauptkomponenten des betrachteten Systems gehören. Des Weiteren werden Repräsentanten der von *recA*-*lexA*-regulierten Gene in die Menge $\mathscr{Q}$ aufgenommen. Dies sind die beiden Gene *ruvC* und *linB*. Das Gen *ruvC* fungiert als Repräsentant für die Menge der Gene, die im Fall von DNA-Zerstörung neben der betrachteten Regulierung auch durch einen alternativen Mechanismus exprimiert werden. Das Gen *linB* repräsentiert diejenigen Gene, die in den Experimenten in *recA*-Mutanten keine erhöhte Expression zeigen. Somit setzen wir

$$\mathscr{Q} = \{recA, lexA, ruvC, linB\}.$$

Die symmetrische Relation $\mathscr{R} \subseteq X \times X$ mit X als Menge aller betrachteten Gene von Mtb ist durch eine Arbeit von Christian Forst gegeben, der über Datenbanken, Literaturdurchsicht und Vorhersagen die Beeinflussungen der Gene untereinander auflistete, siehe Tabelle 6.3. Die Ex-

Tabelle 6.3: Relation $\mathscr{R}$. Die ersten zwei Spalten dieser Matrix mit 68295 Zeilen definieren die Relation.

Genname	Genname	Art der Interaktion
accA1	accA2	ATP
accA1	accA3	ATP
accA1	accA3	BCCP-BIOTIN
⋮	⋮	⋮

pressionsmatrix $E = (e(x_1), ..., e(x_{3658}))^t$ mit $e(x_i) = (e_0(x_i), ..., e_7(x_i))$ entspricht der Matrix aus Tabelle 6.2. Jede Zeile $e(x_i)$ der Matrix E beinhaltet also die Expressionswerte der Zeitpunkte $t_0, ..., t_7$ des Gens $x_i \in X$. Mithilfe von E und $\mathscr{R}$ wird nun der Interaktionsgraph (siehe Kapitel 3.2) mit gewichteten Kanten aufgestellt, der ohne den Gewichten in Abbildung 6.3 zu sehen ist. Der Interaktionsgraph besitzt etwa 1000 Gene und 10.000 Kanten, die für die Interaktionen zwischen den Genen stehen. Auch hier wird ersichtlich, dass ein Großteil der Genfunktionen in Mtb unerforscht ist, da nur knapp ein Viertel aller Gene im Ausgangsgraphen vorhanden sind.

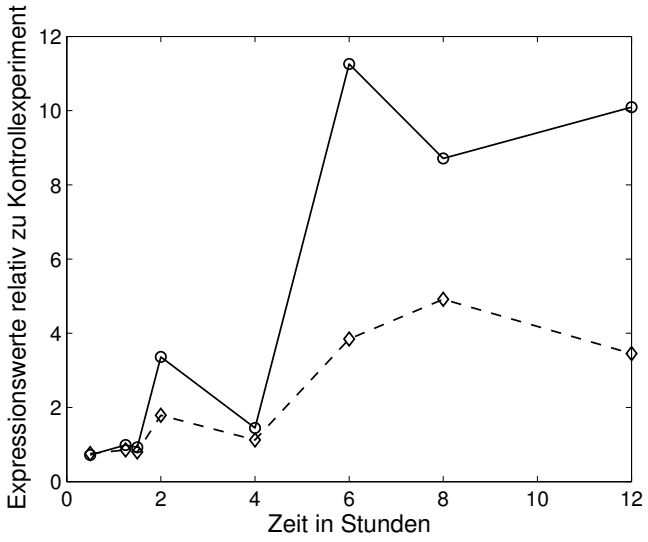


Abbildung 6.2: Nicht logarithmierte, normalisierte und gemittelte Expressionswerte von *recA* und *lexA* für die Zeitpunkte 0.33h, 0.75h, 1.5h, 2h, 4h, 6h, 8h und 12h nach Zugabe von 0.2μ Mitomycin. Die Messungen für *recA* sind durch Kreise, die für *lexA* durch Rauten gekennzeichnet.

Werden die Informationen nach und nach detaillierter, so wird auch die nachfolgende Methode der Komponentenauswahl bessere Ergebnisse erzielen können. Die Gewichtung $c(x_i, x_j)$ der Kante zwischen Gen x_i und Gen x_j erfolgt entsprechend der in Kapitel 3.2 beschriebenen Weise über den Kendallschen Korrelationskoeffizienten $\tau(e(x_i), e(x_j))$. Das Ergebnis der 4-kürzesten-Wege-Suche ist in Abbildung 6.4 zu finden.

Zusätzlich zu den Quellgenen sind die Gene *Rv2719c, dnaE2* und *infB* dem Subnetzwerk hinzugefügt worden. Das Gen *Rv2719c* wurde bereits von Dullaghan u. a. (2002) als ein durch DNA-Zerstörung induziertes Gen entdeckt. Das Gen *infB* ist ein möglicher Translationsinitiationsfaktor und *dnaE2* kodiert vermutlich für eine DNA Polymerase. All diese Gene könnten möglicherweise einen großen Einfluss auf das Netzwerk haben. Es lässt sich erkennen, dass *infB* im Vergleich zu *Rv2719c* und *dnaE2* schwächere Korrelationen zu den Quellgenen aufweist. Um jedoch darauf zu schließen, welche der Gene mit in die Modellierung aufgenommen werden müssen, wird im Folgenden die statistische Signifikanz der Kendallkorrelationskoeffizienten bestimmt, die in Tabelle 6.4 aufgelistet sind. Es gibt Genpaare, die sehr starke Korrelationen aufweisen, wie beispielsweise *recA* mit *linB*. Um nun die Signifikanz des Korrelationskoeffizientens einschätzen zu können, wird die Verteilung $\mathscr{D}$ aller Korrelationskoeffizienten zwischen Genen aus $\mathscr{Q}$ und allen anderen gemessenen Genen von Mtb berechnet. Abbildung 6.5 zeigt die Verteilung $\mathscr{D}$. Mittelwert m und Standardabweichung σ von $\mathscr{D}$ sind $m = -0.004$ und $\sigma = 0.415$. Für ein Signifikanzniveau von $\alpha = 5\%$ gilt, dass Werte oberhalb von $\tau_{\max} = 0.714$ bzw. Werte unterhalb von $\tau_{\min} = -0.714$ als signifikant angesehen werden können. Wie in Tabelle 6.5 zu erkennen, sind sämtliche Korrelationen von *infB* mit den anderen Genen nicht signifikant,

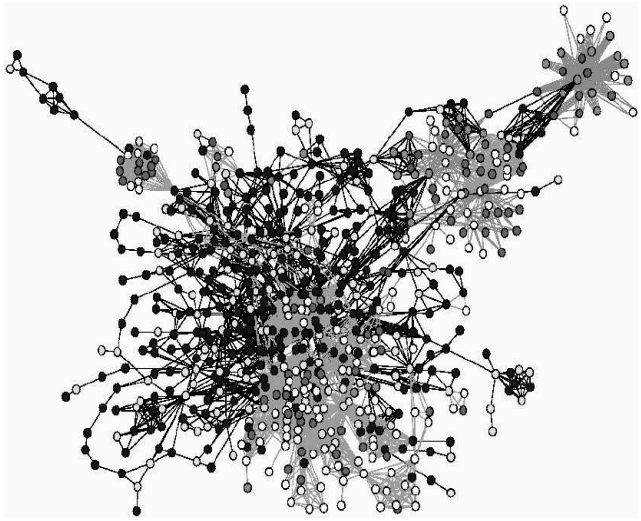


Abbildung 6.3: Die Abbildung erfolgt mit freundlicher Genehmigung von Christian Forst. Sie zeigt den Interaktionsgraphen für die Berechnung des Subnetzwerks. Die verschiedenen Helligkeiten der Kanten stehen für verschiedene Arten der Wechselwirkung zwischen den Genen.

so dass dieses Gen von den weiteren Untersuchungen ausgeschlossen wird. Die beiden Gene *Rv2719c* und *dnaE2* zeigen jedoch beide signifikante Korrelationen zu den Quellgenen. Sehr stark sind die Korrelationen z. B. zwischen *Rv2719c* und *recA* bzw. zwischen *dnaE2* und *lexA*. Die Expressionswerte von *dnaE2* und *Rv2719c* nehmen den in Tabelle 6.6 dargestellten Verlauf. Hierbei ist zu erkennen, dass das Gen *dnaE2* unterhalb einer zweifach verstärkten Expression bleibt. *Rv2719c* dagegen zeigt eine bis zu 12.5fach verstärkte Expression gegenüber dem Kontrollexperiment. Die relativ schwache Antwort von *dnaE2* ist der Grund, als einziges Gen *Rv2719c* mit in das Modell aufzunehmen. Im folgenden Unterkapitel wird nun überprüft, ob die Aufnahme dieses Gens in das Modell Verbesserungen der Simulation gegenüber dem Modell ohne *Rv2719c* zeigt. Das Modell ohne *Rv2719c* wird im Folgenden Grundmodell, jenes mit *Rv2719c* erweitertes Modell genannt. Die bisher bekannten bzw. angenommenen Regulierungen zwischen den Genen sind in Abbildung 6.6 skizziert. Das Protein LexA kann in aktiviertem und nicht aktiviertem Zustand existieren. Aktiviert heißt das Protein, wenn es an die SOS-Boxen in der DNA binden kann. Daher werden hier zwei Variablen benutzt. LexA bezieht sich auf beide Zustände, also auf die Gesamtmenge von LexA in der Zelle, wohingegen LexASOS den Anteil der aktiven Proteinkonzentration meint, den Anteil also, der an SOS-Boxen binden kann. In Abbildung 6.6 sieht man eine positive Kante von einem Signal zu *recA*. Das Signal ist einzelsträngige DNA und bewirkt, dass das Protein RecA aktiviert wird, was wiederum eine Verminderung der LexASOS-Konzentration bewirkt, also derjenigen LexA-Konzentration, die fähig ist, an die DNA zu binden. Die Kante von *lexA* auf *lexASOS* in Abbildung 6.6 soll verdeutlichen, dass LexASOS eine Teilmenge von LexA ist. Bei hoher LexASOS-Konzentration haben wir eine negative Regulierung auf sämtliche Gene, die eine SOS-Box besitzen. Das sind in der Abbildung

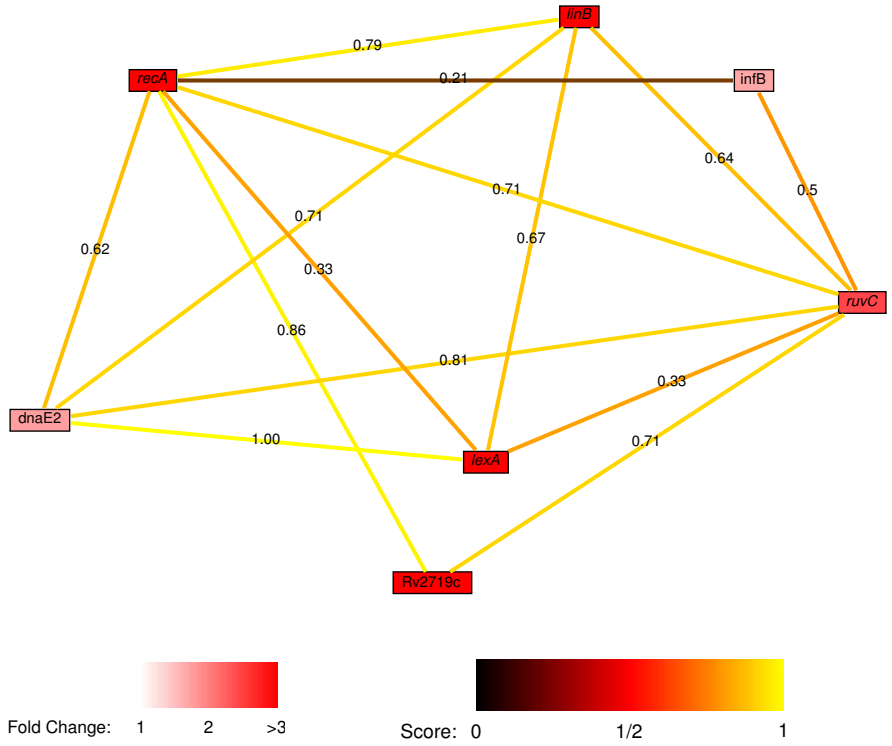


Abbildung 6.4: Die Abbildung erfolgt mit freundlicher Genehmigung von Christian Forst. Sie zeigt das Ergebnis aus der Berechnung des Subnetzwerks mit den Quellgenen $\mathscr{Q} = \{recA, lexA, ruvC, linB\}$. Die Parameter für die Anzahl der kürzesten Wege und für die maximale Pfadlänge sind auf $K = 4$ und $L = 10$ gesetzt. Helle Kanten entsprechen starker Korrelation und dunkle Kanten schwacher Korrelation zwischen den jeweiligen Genen. Je dunkler die Knoten sind, desto höheres Transkriptniveau haben die Gene.

6.6 die Gene *recA, lexA, linB, ruvC* und *Rv2719c*. Diese negative Regulierungen sind durch negativ markierte, gerichtete Kanten dargestellt. Im erweiterten Modell sind außerdem noch zwei positiv markierte, gerichtete Kanten von *Rv2719c* zu *ruvC* und zu *recA* zu sehen. Diese Regulierungen entsprechen der Vermutung, dass *Rv2719c* einem alternativen Mechanismus bezüglich des DNA-Reparatursystems zugeordnet werden könnte und die Gene *recA* und *ruvC* in *lexA*-Mutantenexperimenten bei DNA-Schäden ein hohes Transkriptniveau haben (Dullaghan u. a., 2002).

6.4 Modellierung des Systems, Parameterschätzung und Simulationen

Die Differenzialgleichungen bezüglich des Grundmodells und des erweiterten Modells werden entsprechend der in Kapitel 2 eingeführten Gleichung (2.34) gebildet. Gen *recA* wird im Fol-

Tabelle 6.4: Kendall-Korrelationskoeffizienten für jedes Paar von Genen im System.

Gene	*lexA*	*recA*	*ruvC*	*linB*	*infB*	*dnaE2*	*Rv2719c*
lexA	1.00	0.33	0.33	0.67	-0.67	1.00	0.00
recA	0.33	1.00	0.71	0.79	0.21	0.62	0.86
ruvC	0.33	0.71	1.00	0.64	0.5	0.81	0.71
linB	0.67	0.79	0.64	1.00	0.29	0.71	0.79
infB	-0.67	0.21	0.50	0.29	1.00	0.62	0.36
dnaE2	1.00	0.62	0.81	0.71	0.62	1.00	0.52
Rv2719c	0.00	0.86	0.71	0.79	0.36	0.52	1.00

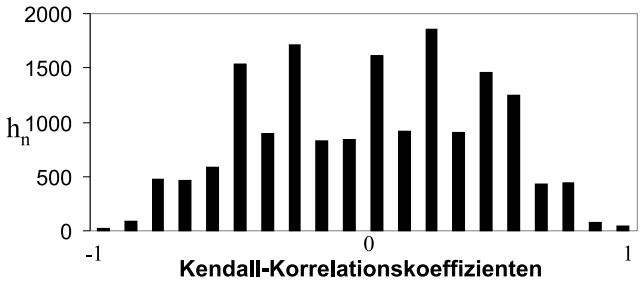


Abbildung 6.5: Verteilung der Kendall-Korrelationskoeffizienten für die Netzwerkgene mit allen anderen gemessenen Genen des Mtbs. Auf der y-Achse ist h_n angegeben, die absolute Häufigkeit der jeweiligen Korrelationskoeffizienten.

genden dem Index 1 entsprechen, *lexA* Index 2, *lexASOS* Index 3, *ruvC* Index 4, *linB* Index 5 und *Rv2719c* Index 6. Die Variable x_i bezeichnet wie zuvor die mRNA-Konzentration von i, $i = 1,...,6$. Die Variable x_6 wird als externe Variable behandelt, da die regulatorischen Einflüsse auf dieses Gen noch unbekannt sind. Im Folgenden werden daher die Werte für *Rv2719c* direkt aus den Messungen genommen. Auch können nur Annahmen über die Variable x_3 getroffen werden, da in den Experimenten nur die mRNA-Konzentration von *lexA* gemessen wurde, nicht jedoch die Bindungsfähigkeit des Proteins LexA nachgewiesen werden konnte. Es wird eine externe Variable *signal* verwendet, die für eine Zerstörung der DNA steht und in dem Experiment bis sechs Stunden ansteigt und anschließend auf Null fällt.
Zunächst werden die Differenzialgleichungen für x_1, x_2, x_4 und x_5 in dem Grundmodell ohne

Tabelle 6.5: Wahrscheinlichkeiten, den Kendall-Korrelationskoeffizienten für ein Genpaar oder eine noch höhere Abweichung vom Mittelwert m der Verteilung $\mathscr{D}$ zu sehen. Signifikante Werte sind fett gedruckt.

Gene	*lexA*	*recA*	*ruvC*	*linB*	*infB*	*dnaE2*	*Rv2719c*
lexA	0.00	0.51	0.51	0.07	0.07	**<0.01**	1.00
recA	0.51	0.00	**0.04**	**0.03**	0.62	0.13	**0.01**
ruvC	0.51	**0.04**	0.00	0.07	0.25	**0.01**	**0.04**
linB	0.07	**0.03**	0.07	0.00	0.57	**0.04**	**0.03**
infB	0.07	0.62	0.25	0.57	0.00	0.13	0.45
dnaE2	**<0.01**	0.13	**0.01**	**0.04**	0.13	0.00	0.20
Rv2719c	1.00	**0.01**	**0.04**	**0.03**	0.45	0.20	0.00

Tabelle 6.6: Expressionswerte der Gene *dnaE2* und *Rv2719c*.

	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
dnaE2	0.8504	0.4023	0.7930	1.4070	1.2419	1.8933	1.9249	1.7318
Rv2719c	0.8421	1.3797	0.8575	1.1351	1.7993	8.7817	5.8426	12.4909

Rv2719c aufgestellt:

$$\begin{aligned}
\dot{x}_1(t) &= c_1 - \gamma_1 \cdot x_1(t) + \alpha_1 \cdot signal + k^-_{1,3} \cdot b(x_3(t), \theta_{1,3}) \\
\dot{x}_2(t) &= c_2 - \gamma_2 \cdot x_2(t) + k^-_{2,3} \cdot b(x_3(t), \theta_{2,3}) \\
\dot{x}_4(t) &= c_4 - \gamma_4 \cdot x_4(t) + k^-_{4,3} \cdot b(x_3(t), \theta_{4,3}) \\
\dot{x}_5(t) &= c_5 - \gamma_5 \cdot x_5(t) + k^-_{5,3} \cdot b(x_3(t), \theta_{5,3})
\end{aligned}$$

mit $\alpha_1 \in \mathbf{R}^+$, $c_i, \gamma_i \in \mathbf{R}^+$ für $i = 1,2,4,5$ und $k^-_{i,j} \in \mathbf{R}^-$ für die Paare $(1,3),(2,3),(4,3)$ und $(5,3)$.

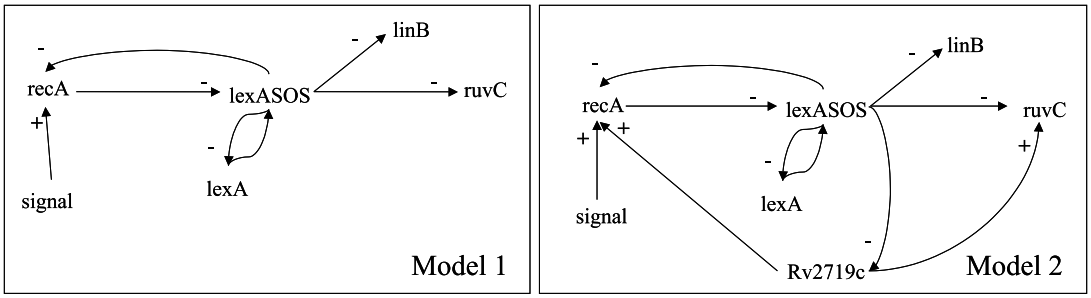


Abbildung 6.6: Grundmodell und erweitertes Modell des DNA Reparatursystems.

Es werden Boolesche Funktionen

$$k \cdot b(x(t), \theta) \quad = \quad k \cdot \begin{cases} 0 & \text{für } x(t) \leq \theta \\ 1 & \text{für } x(t) > \theta \end{cases}$$

mit $k \in \mathbf{R}^+$ für k^+ und $k \in \mathbf{R}^-$ für k^- sowie $\theta \in \mathbf{R}^+$ zugrunde gelegt. Dies ist eine Approximation der stückweise linearen Regulierungsfunktionen rf^+, rf^- aus Gleichung (2.34), da die Anzahl der Datenpunkte keine Aussage über detailliertere Regulierungsfunktionen zulässt. Anstelle von zwei Schwellwerten pro Funktion wird also nur ein Schwellwert θ gewählt. Je nach Lage von x bezüglich des Schwellwertes θ ist somit der Regulierungseinfluss nicht oder direkt maximal vorhanden.
Für das um *Rv2719c* erweiterte Modell werden zwei zusätzliche Funktionen eingeführt:

$$\begin{aligned}
\dot{x}_1(t) &= c_1 - \gamma_1 \cdot x_1(t) + \alpha_1 \cdot signal + k^-_{1,3} \cdot b(x_3(t), \theta_{1,3}) \\
&\quad + k^+_{1,6} \cdot b(x_6(t), \theta_{1,6}) && (6.1) \\
\dot{x}_2(t) &= c_2 - \gamma_2 \cdot x_2(t) + k^-_{2,3} \cdot b(x_3(t), \theta_{2,3}) && (6.2) \\
\dot{x}_4(t) &= c_4 - \gamma_4 \cdot x_4(t) + k^-_{4,3} \cdot b(x_3(t), \theta_{4,3}) + k^+_{4,6} \cdot b(x_6(t), \theta_{4,6}) && (6.3) \\
\dot{x}_5(t) &= c_5 - \gamma_5 \cdot x_5(t) + k^-_{5,3} \cdot b(x_3(t), \theta_{5,3}) && (6.4)
\end{aligned}$$

mit $k^+_{1,6}, k^+_{4,6} \in \mathbf{R}^+$.
Für die Parameterschätzung müssen daher eine Reihe von Annahmen getroffen werden, um die Schätzung durchführen zu können. Die Parameterschätzung wird wie in Kapitel 4.2 beschrieben durchgeführt. Für die Variablen x_3 und x_6 können aufgrund der Datenlage keine Differenzialgleichungen aufgestellt werden. Für die Parameterschätzungen werden die jeweils gemessenen Daten für *Rv2719c* verwendet, also die in Tabelle 6.6 aufgeführten Werte für *Rv2719c*. Für x_3 sind keine Messungen vorhanden, so dass angenommen wird, dass LexA während der Zeitpunkte t_4 und t_5 nicht an die SOS-Boxen binden kann, zu allen übrigen Zeitpunkten jedoch bindungsfähig ist:

$$x_3(t) \quad = \quad \begin{cases} 1 & \text{für die Zeitpunkte } t_0, ..., t_3 \\ 0 & \text{für die Zeitpunkte } t_4, t_5 \\ 1 & \text{für die Zeitpunkte } t_6, t_7. \end{cases}$$

Die Zerstörung der DNA geht über *signal* in die Differenzialgleichung für x_1 ein. Die Funktion $signal(t)$ wird als linear ansteigend bis zum Zeitpunkt t_5 angenommen, danach als konstant gleich Null festgesetzt. Als Nebenbedingung setzen wir, dass alle Parameter größer oder gleich Null sein müssen bis auf die Parameter $k^-_{1,3}, k^-_{2,3}, k^-_{4,3}$ und $k^-_{5,3}$, die kleiner oder gleich Null sein müssen. Zur Reduktion der Parameter wird außerdem eine Abbaurate von $\gamma_i = 0.1$ für alle $i = 1,2,4,5$ gesetzt. Die Grundsyntheseraten c_i, $i = 1,2,4,5$, werden entsprechend so berechnet, dass die Fixpunkte für alle Variablen bei 1 liegen. Das Ergebnis der Parameterschätzung für das Grundmodell lautet dann

$$\alpha_1 = 0.548h^{-1}, k^-_{1,3} = 0h^{-1}, k^-_{2,3} = -0.898h^{-1},$$
$$k^-_{4,3} = -0.168h^{-1}, k^-_{5,3} = -0.82h^{-1}$$

und für das erweiterte Modell

$$\alpha_1 = 0.511h^{-1}, k^-_{1,3} = 0h^{-1}, k^-_{2,3} = -0.898h^{-1},$$
$$k^-_{4,3} = -0.013h^{-1}, k^-_{5,3} = -0.82h^{-1},$$
$$k^+_{4,6} = 0.233h^{-1}, k^+_{1,6} = 0.464h^{-1}.$$

Man erkennt, dass *ruvC* nun stärker durch *Rv2719c* beeinflusst wird als durch LexASOS, da sich der Parameter $k^-_{4,3} = -0.168h^{-1}$ in dem Grundmodell auf $k^-_{4,3} = -0.013h^{-1}$ im erweiterten Modell abschwächt.

Simulationen für das Grund- bzw. das erweiterte Modell werden mit den Anfangsbedingungen durchgeführt, die dem Experiment entsprechen, aus dem die Daten gewonnen wurden. Die Simulationen sind in Abbildungen 6.7 und 6.8 zu sehen. Die experimentellen Daten sind ebenfalls eingetragen, bei mehreren Messungen zu einem Zeitpunkt ist der Mittelwert gebildet.

6.5 Diskussion der Ergebnisse

Beide Simulationen zeigen ein ähnliches Verhalten zwischen t_0 und t_5. Die Simulation für die mRNA-Konzentration von *recA*, x_1, fängt direkt von Zeitpunkt t_0 an zu steigen. Die Simulationen von x_2, x_4 und x_5, also der mRNA-Konzentrationen von *lexA, ruvC* und *linB*, bleiben zunächst unverändert bei 1, und fangen dann an, ab Zeitpunkt t_4 zu steigen. Nach dem Zeitpunkt t_5 unterscheiden sich die Simulationen, da im erweiterten Modell ein positiver Einfluss von *Rv2719c* auf *ruvC* und *recA* modelliert ist. Im Grundmodell fällt x_1 ab Zeitpunkt t_5 wieder ab, die drei übrigen folgen ab Zeitpunkt t_6. Im erweiterten Modell fällt x_1 ebenfalls ab Zeitpunkt t_5, jedoch nicht so stark wie im Grundmodell, was den experimentellen Daten näher kommt. x_2 und x_5 zeigen unverändertes Verhalten, da auch die Differenzialgleichungen im erweiterten Modell im Vergleich zu dem Grundmodell unverändert sind. Variable x_4 dagegen zeigt ein verbessertes Verhalten in dem Sinne, dass die Simulation den experimentellen Daten besser entspricht. Die Konzentration steigt auch nach dem Zeitpunkt t_6 weiterhin an. Es haben sich also nur zwei Simulationen verändert, diejenige von x_1 und diejenige von x_4. Beide fallen im Grundmodell schneller auf ihr Ausgangsniveau zurück, so dass die Antwort des Systems auf das Signal wieder schnell abfällt. Im erweiterten Modell dagegen weisen beide Gene auch im späteren Verlauf nach ca. 10 Stunden eine höhere Expression auf, erhalten die Antwort auf das Signal also länger aufrecht. Dieses Verhalten entspricht den experimentell erhobenen Daten besser, so dass gefolgert werden kann, dass das hypothetische Gen *Rv2719c* eine möglicherweise wichtige Rolle im DNA Reparatursystem spielt. Dieses Ergebnis stimmt auch überein mit den experimentellen Ergebnissen von Dullaghan u. a. (2002).

6.6 Alternative Komponentenauswahl

Wie bereits im dritten Kapitel erwähnt, kann die Komponentenauswahl nicht nur mit dem in Kapitel 3.1 beschriebenen graphentheoretischen Ansatz erfolgen, sondern auch mithilfe des in Kapitel 3.2 beschriebene Ansatzes, der auf der formalen Begriffsanalyse beruht. Im Folgenden

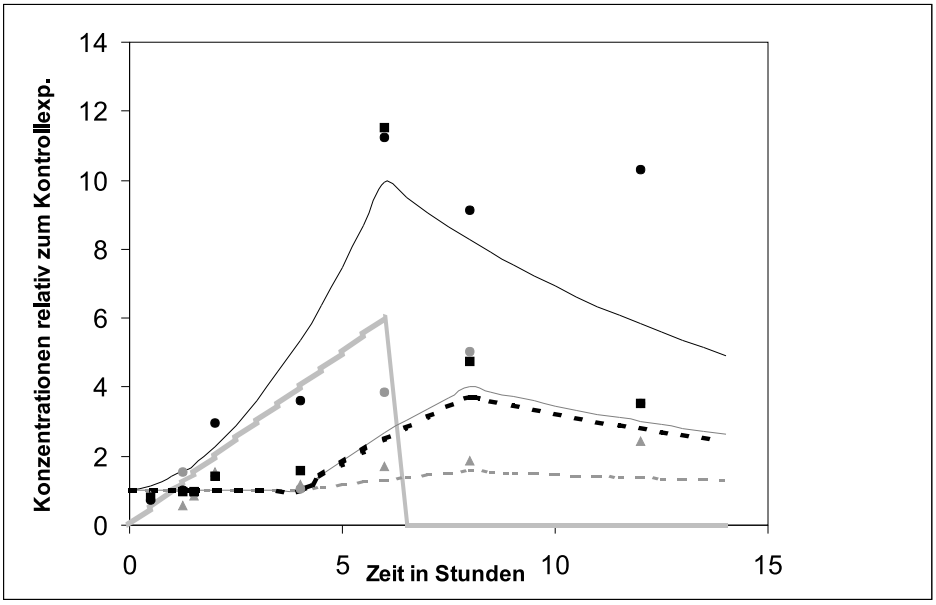


Abbildung 6.7: In dieser Simulation werden die Differenzialgleichungen des Grundmodells mit Anfangsbedingungen $x_i = 1, i = 1,2,4,5$, verwendet. Die externe Variable *signal*, die die Zerstörung der DNA durch Mitomycin beschreibt, wächst linear an bis zum Zeitpunkt t_5, danach ist das Signal ausgeschaltet. Das Signal ist durch eine graue dicke Linie gekennzeichnet. Die experimentellen Messungen sind durch schwarze Kreise für *recA*, graue Kreise für *lexA*, graue Dreiecke für *ruvC* und schwarze Quadrate für *linB* gekennzeichnet. Die Simulationen sind durch eine schwarze Linie für *recA*, eine graue dünne Linie für *lexA*, eine gestrichelte graue Linie für *ruvC* und eine gestrichelte schwarze Linie für *linB* dargestellt.

soll dieser zweite Ansatz dem ersten gegenüber gestellt werden. Der erste und dritte Schritt wurde von mir, der zweite Schritt wurde von Susanne Motameny durchgeführt. Um das Ergebnis der beiden Auswahlmethoden miteinander vergleichen zu können, werden wieder die Quellgene

$$\mathscr{Q} = \{recA, lexA, ruvC, linB\}$$

zugrunde gelegt und sowohl die Interaktionsdaten über die Relation $\mathscr{R}$ als auch die Genexpressionsdaten $E = (e_1(t), ..., e_{3658}(t))^t$ verwendet.

Erster Schritt: Vorauswahl der Gene

Die Genlisten werden wie in Kapitel 3.2.2 beschrieben für $K = 4$ berechnet. Als Ergebnis erhält man zwölf Gene für die Genliste S_1,

$$S_1 = \{recA, lexA, ruvC, linB, Rv2719c, dnaE2, dinX, infB, lprJ, Rv1691, Rv2840c, Rv3055\},$$

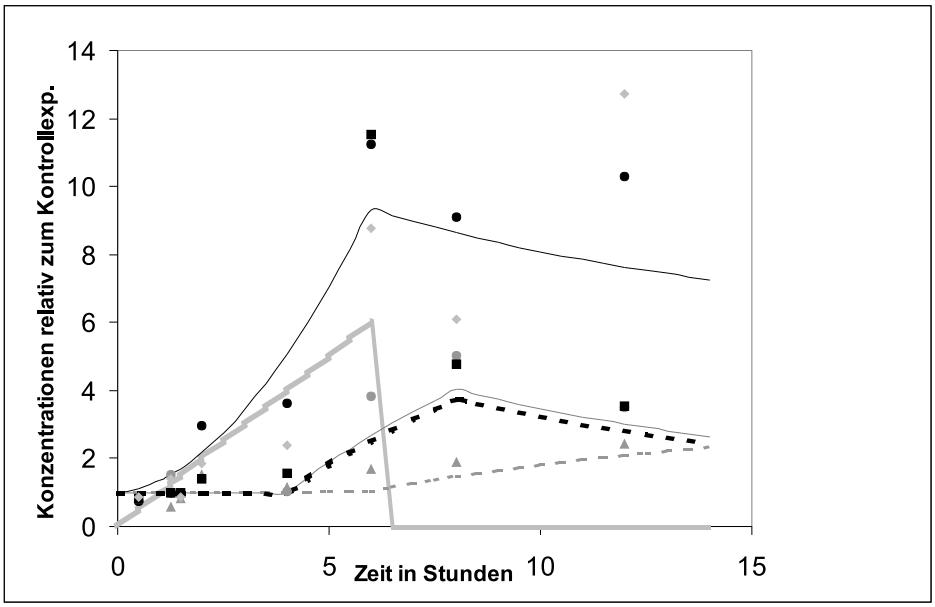


Abbildung 6.8: In dieser Simulation werden die Differenzialgleichungen des erweiterten Modells mit Anfangsbedingungen $x_i = 1, i = 1,2,4,5$, verwendet sowie das gleiche Signal wie in Abbildung 6.7. Wieder ist das Signal durch die graue dicke Linie gekennzeichnet. Die experimentellen Messungen sind durch schwarze Kreise für *recA*, graue Kreise für *lexA*, graue Dreiecke für *ruvC*, schwarze Quadrate für *linB* und graue Rauten für *Rv2719c* gekennzeichnet. Die Simulationen sind durch eine schwarze Linie für *recA*, eine graue dünne Linie für *lexA*, eine gestrichelte graue Linie für *ruvC* und eine gestrichelte schwarze Linie für *linB* dargestellt.

196 Gene für die Genliste S_2, 808 Gene für die Genliste S_3 und 925 Gene für die Genliste S_4. Genliste S_2 ist in Anhang A.1 zu finden. Anschließend werden sämtliche Gene aus diesen Ausgangslisten entfernt, für die gilt, dass sie über sämtliche Zeitpunkte gesehen nicht mindestens einmal zweifach hoch- oder herunterreguliert sind, über 60 % der Zeitpunkte nicht gemessen wurden oder die Differenz zwischen Signal und Hintergrundrauschen für Experiment oder Kontrolle kleiner als 10 ist. Somit reduziert sich die Menge S_1 auf 8, S_2 auf 68, S_3 auf 258 und S_4 auf 339 Gene.

Zweiter Schritt

Es wird eine Menge $\mathscr{M} = \{\mathscr{M}_1, ..., \mathscr{M}_6\}$ von sechs Mustern über die acht Zeitpunkte definiert, und zwar durch

$$\begin{aligned}
\text{Temp1} = \mathscr{M}_1 &= (0,0,0,1,2,3,2,2),\\
\text{Temp2} = \mathscr{M}_2 &= (0,0,0,0,1,2,1,1),\\
\text{Temp3} = \mathscr{M}_3 &= (0,0,0,1,2,3,2,1),\\
\text{Temp4} = \mathscr{M}_4 &= (0,0,0,0,1,2,1,0),\\
\text{Temp5} = \mathscr{M}_5 &= (0,0,0,1,2,3,3,3),\\
\text{Temp6} = \mathscr{M}_6 &= (0,0,0,1,1,2,1,1).
\end{aligned}$$

Hierbei wurden die Quellgene als Vorbild genommen. Wir sind interessiert an dem Verhalten, dass zu den ersten drei bis vier Zeitpunkten keine Änderung erfolgt und dass bei Zeitpunkt vier oder fünf die Expression ansteigt. Anschließend bleibt die Expression auf hohem Niveau oder fällt wieder ab.

Nun werden für $i = 1,...,4$ sämtliche erweiterte Kendallsche Korrelationskoeffizienten (siehe Definition 3.2) $\tau_{\tilde{K}}(e_k, \mathscr{M}_j)$ für $k = 1,...,|S_i|$, $j = 1,...,6$ berechnet. Für $i = 1,...,4$ ist $\mathbb{K}_i = (S_i, \mathscr{M}, I_i)$ ein Kontext mit

$$(g, \mathscr{M}_j) \in I_i \quad \Leftrightarrow \quad |\tau_{\tilde{K}}(e, \mathscr{M}_j)| > 0.6 \text{ für } g \in S_i \text{ und } \mathscr{M}_j \in \mathscr{M},$$

wobei e die Genexpressionszeitreihe von Gen $g \in S_i$ ist.
Die zugehörigen Begriffsverbände $\underline{\mathscr{B}}(\mathbb{K}_1)$ und $\underline{\mathscr{B}}(\mathbb{K}_2)$ sind in Abbildungen 6.9 und 6.10 zu sehen.

Dritter Schritt

Die Menge S' der weiter zu analysierenden Gene wird über Zwischenbegriffe der Gegenstandsbegriffe μg für $g \in \mathscr{Q}$ bestimmt, siehe dazu Kapitel 3.2.2. Das Ergebnis für $i = 1$ und 2 ist S'_1={*recA, lexA, ruvC, linB, Rv2719c, dnaE2, lprJ* } und S'_2={*recA, lexA, ruvC, linB, Rv1781c, Rv2719c, dnaB, dnaE2, fadD21, fadD23, fadD25, idsA, lprJ, mtrA, phyA, uvrA, uvrD*}. Für $i = 3$ und $i = 4$ siehe Anhang A.1. Die Genmenge S'_3 enthält 59 Gene, S'_4 enthält 69 Gene.
In Tabelle 6.7 sind die erweiterten Kendallschen Korrelationskoeffizienten für die Gene aus S_2 mit den Quellgenen sowie die entsprechenden Wahrscheinlichkeiten aufgelistet (siehe Kapitel 3.1.2).
Setzt man das Signifikanzniveau auf $\alpha = 0.05$, so erhält man die Mengen $\tilde{S}_i$, $i = 1,...,4$, derjenigen Gene, die möglicherweise stark mit den Quellgenen interagieren. Diese Menge besteht für $\tilde{S}_1$ aus

$$\tilde{S}_1 \quad = \quad \{Rv2719c, dnaE2\}$$

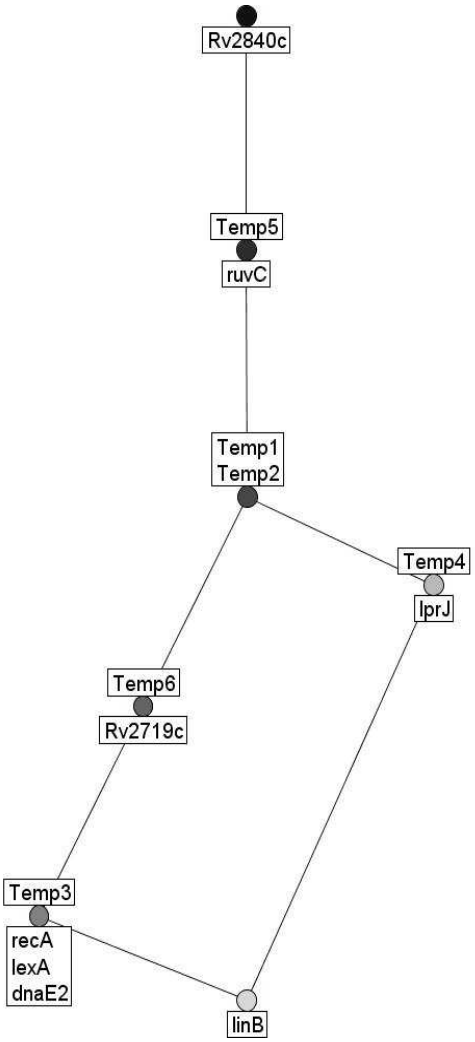


Abbildung 6.9: Der Begriffsverband $\underline{\mathscr{B}}(\mathbb{K}_1)$. Die Erstellung der Begriffsverbände erfolgte mit der Software ToscanaJ (Becker u. a., 2002).

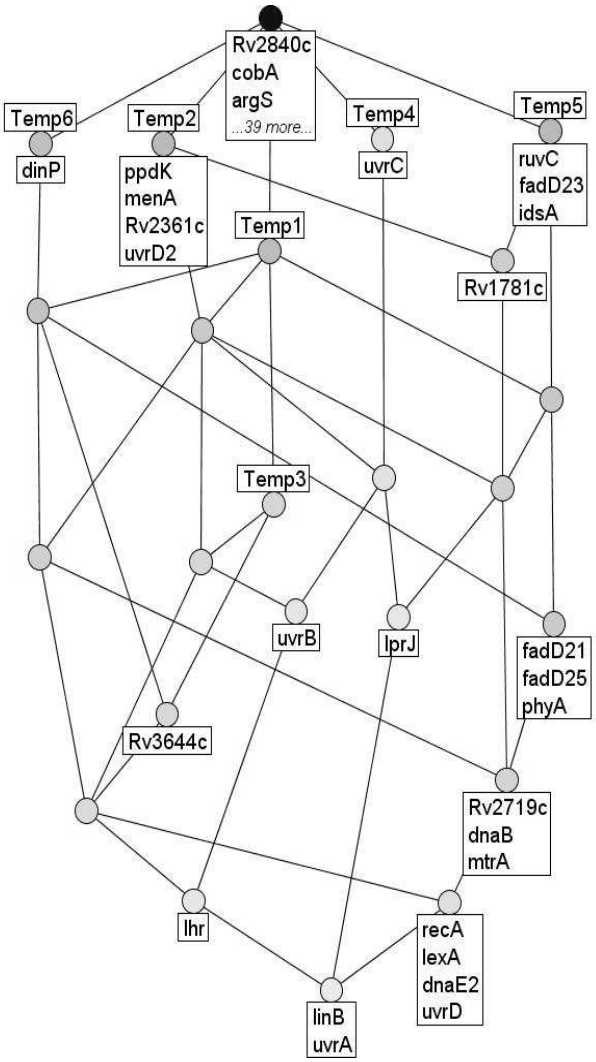


Abbildung 6.10: Der Begriffsverband $\underline{\mathscr{B}}(\mathbb{K}_2)$, dargestellt mit ToscanaJ (Becker u. a., 2002).

Tabelle 6.7: Erweiterte Kendallsche Korrelationskoeffizienten und zugehörige Wahrscheinlichkeiten in Klammern. Signifikante Werte bezüglich eines Signifikanzniveaus $\alpha = 5\%$ sind fett gedruckt.

Gene	*lexA*	*recA*	*ruvC*	*linB*
Rv1781c	0.294 (0.498)	0.309 (0.476)	0.192 (0.643)	0.347 (0.428)
Rv2719c	0.584 (0.118)	**0.667 (0.051)**	0.512 (0.204)	**0.720 (0.026)**
dnaB	**0.681 (0.043)**	**0.707 (0.031)**	**0.681 (0.043)**	**0.732 (0.022)**
dnaE2	**0.668 (0.051)**	**0.732 (0.023)**	**0.784 (0.011)**	**0.732 (0.023)**
fadD21	-0.626 (0.164)	**-0.786 (0.033)**	**-0.768 (0.039)**	**-0.725 (0.065)**
fadD23	**-0.783 (0.033)**	**-0.756 (0.046)**	**-0.756 (0.046)**	-0.554 (0.260)
fadD25	-0.642 (0.139)	**-0.720 (0.067)**	-0.618 (0.170)	-0.658 (0.129)
idsA	-0.634 (0.152)	-0.668 (0.115)	**-0.713 (0.080)**	-0.520 (0.303)
lexA	1.000 (0.000)	**0.784 (0.010)**	**0.725 (0.025)**	**0.706 (0.032)**
linB	**0.706 (0.032)**	**0.756 (0.018)**	**0.642 (0.065)**	1.000 (0.000)
lprJ	-0.524 (0.297)	-0.577 (0.225)	-0.477 (0.363)	**-0.709 (0.081)**
mtrA	-0.401 (0.462)	-0.550 (0.266)	-0.505 (0.320)	-0.642 (0.141)
phyA	0.544 (0.164)	0.607 (0.097)	0.618 (0.087)	0.550 (0.156)
recA	**0.784 (0.010)**	1.000 (0.000)	**0.780 (0.011)**	**0.756 (0.018)**
ruvC	**0.725 (0.025)**	**0.780 (0.011)**	1.000 (0.000)	**0.642 (0.065)**
uvrA	0.521 (0.191)	0.577 (0.129)	0.472 (0.261)	**0.637 (0.068)**
uvrD	0.404 (0.350)	0.456 (0.278)	0.342 (0.436)	0.524 (0.188)

und für $\tilde{S}_2$ aus

$$\tilde{S}_2 \quad = \quad \{Rv2719c, dnaB, dnaE2, fadD21, fadD23\}.$$

Für $\tilde{S}_3$ erhalten wir zusätzlich zu den Genen aus $\tilde{S}_2$ die Gene {*Rv0059, Rv0881, Rv2165c, accD2, ald, alkA, fadE21, fbpC2, fmt, ltp1, sdhA, thiL*}, so dass $|\tilde{S}_3| = 17$ gilt. $\tilde{S}_4$ erweitert sich gegenüber $\tilde{S}_3$ um das Gen {*rpe*} und wir erhalten $|\tilde{S}_4| = 18$. Wird K immer höher gesetzt, so wächst die Menge $\tilde{S}$. Wählt man anstelle S_k direkt die gesamte Genmenge G, so enthält $\tilde{S}$ all diejenigen Gene, deren Genexpressionen den Genexpressionen der Quellgene im Sinne der Begriffsanalyse nahe kommen (siehe Kapitel 3.2.2) und die zumindest zu einem der Quellgene eine sehr hohe Korrelation besitzen.

Aufgrund der Werte in Tabelle 6.7 werden also für S_1 die Gene *Rv2719c* und *dnaE2* sowie für S_2 die Gene *Rv2719c, dnaB, dnaE2, fadD21* und *fadD23* in das genregulatorische Netzwerk aufgenommen, siehe Abbildung 6.11.
Die Interaktionen zwischen den Quellgenen sind bekannt, so dass diese Interaktionen über gerichtete Kanten dargestellt werden können. Eine ungerichtete Kante zwischen einem Quellgen und einem anderen Gen existiert, falls diese Kante signifikant ist (siehe Kapitel 3.2.2).

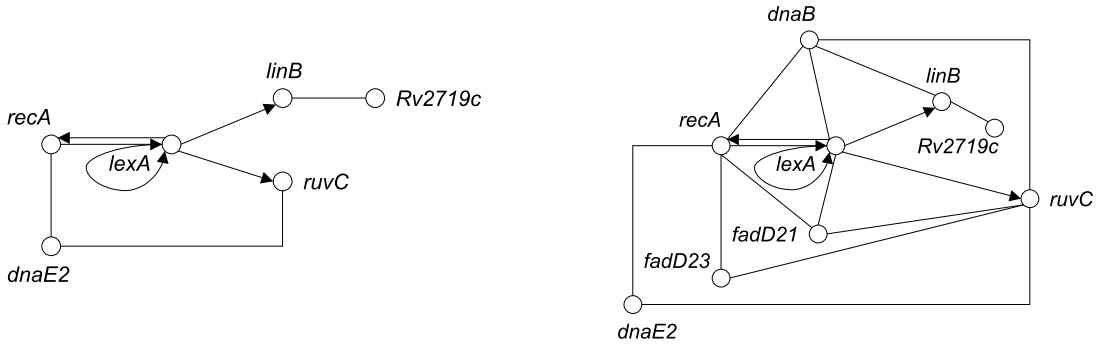


Abbildung 6.11: Ergebnis des genregulatorischen Netzwerkes für $\tilde{S}_1$ (links) und $\tilde{S}_2$ (rechts). Gerichtete Kanten werden aufgrund der Kenntnis der biologischen Prozesse während der *recA-lexA*-Regulierung eingefügt. Ungerichtete Kanten entsprechen den signifikanten Werten aus Tabelle 6.7.

Ergebnisse und Vergleich mit der ersten Methode

Ausgehend von S_1 werden die Gene *Rv2719c* und *dnaE2* in das genregulatorische Netzwerk aufgenommen, ausgehend von S_2 sind dies zusätzlich *dnaB, fadD21* und *fadD23*. Die folgenden Informationen über die Funktionen der Gene sind der Datenbank TubercuList[5] des Pasteur Institutes Paris zu entnehmen:

- *Rv2719c*: Dieses Gen wird im Fall einer DNA-Zerstörung stärker exprimiert und es wird angenommen, dass es von LexA reguliert wird (Dullaghan u. a., 2002).
- *dnaE2*: Dieses Gen kodiert vermutlich für eine DNA Polymerase, welche ein Enzym ist, das bei der Replikation der DNA in Bakterien eine Rolle spielt.
- *dnaB*: Dieses Gen kodiert vermutlich für eine replikative DNA Helikase.
- *fadD21*: Dieses Gen kodiert vermutlich für eine Fettsäure-CoA-Ligase.
- *fadD23*: Dieses Gen kodiert vermutlich ebenfalls für eine Fettsäure-CoA-Ligase.

Das zuerst aufgeführte Gen *Rv2719c* wurde auch in der ersten Methode als möglicherweise wichtiges Gen im DNA-Reparatur-Netzwerk identifiziert. In dieser zweiten Methode erhalten wir je nach Wahl von S_k noch weitere Gene. Die drei erstgenannten Gene *Rv2719c, dnaE2* und *dnaB* haben alle etwas mit dem SOS-System oder - allgemeiner - einer Veränderung der DNA zu tun. Daher werden diese Ergebnisse als plausibel bewertet. Die beiden letztgenannten Gene dagegen, *fadD21* und *fadD23*, scheinen keine Verbindung zur DNA-Reparatur zu haben und sind eventuell nur zufällig so stark mit den Quellgenen korreliert.

[5] http://genolist.pasteur.fr/TubercuList/

Die Interaktionsdaten über *M. tuberculosis* sind in diesem Fall aufgrund der Arbeit von Christian Forst im Gegensatz zu vielen anderen Organismen relativ umfassend, wenn auch nicht vollständig. Daher lassen sich bei beiden Methoden diese Daten verwenden. Die erste Methode nimmt Einfluss auf die Anzahl der Gene haben, welche in die Ergebnisliste aufgenommen werden, über die zwei Parameter für die Anzahl der kürzesten Wege und für die maximale Pfadlänge. Bei der zweiten Methode haben die selbst zu definierenden Muster $\mathscr{M}$, wiederum ein Parameter zur Bestimmung der aufgrund der Interaktionsdaten zu verwendenden Gene und ein Schwellwert für die Korrelationskoeffizienten einen Einfluss auf das Ergebnis. Insofern ist die zweite Methode vorzuziehen, falls - wie im hier vorliegenden Fall - die Form der Muster $\mathscr{M}$ leicht zu bestimmen ist. In unserem Fall ist bekannt, dass eine Schädigung der DNA zu einer verstärkten Expression einiger Gene führt, welche nach einiger Zeit wieder schwächer wird. Alternativ könnten ursprünglich stark exprimierte Gene schwächer exprimiert werden und ebenfalls nach einiger Zeit ihr Ursprungsniveau wieder erreichen. Dies lässt sich in den ausgewählten Mustern widerspiegeln. Hat man solch ein Wissen jedoch nicht, so ist die erste Methode vorzuziehen, da die Auswahl der Muster in der zweiten Methode einen starken Einfluss hat.

Kapitel 7

Verbandstheoretische Visualisierungen biochemischer Netzwerke

Die Dynamik von Reaktionssystemen kann mithilfe von Differenzialgleichungen beschrieben werden. Um das Verhalten des Systems zu visualisieren, ist es zum Beispiel möglich, zu gegebenen Anfangszuständen Lösungskurven oder Trajektorien graphisch darzustellen, wie dies zum Beispiel in den Kapiteln 2, 5 und 6 erfolgte. In diesem Kapitel sollen zwei andere Möglichkeiten der Visualisierung für die beiden Systeme aus Kapitel 5 und 6 erweitert und diskutiert werden. Beide Ansätze verwenden Verbände, im ersten werden sämtliche Anfangszustände betrachtet, wobei jedoch Informationsverluste hingenommen werden müssen, im zweiten Ansatz dagegen wird nur ein einzelner Anfangszustand im Verband weiter verfolgt. Die chemische Organisationstheorie wurde bisher für Reaktionen zwischen Molekülen verwendet, wird hier jedoch auf Genregulierungen erweitert, indem Aktivierungen und Inhibierungen zugelassen werden.

7.1 Chemische Organisationstheorie

Diese Art der Visualisierung von biochemischen Netzwerken wurde durch Dittrich u. di Fenizio (2005); di Fenizio u. Dittrich (2002) bekannt gemacht und basiert auf Überlegungen von Fontana u. Buss (1994). Im Zentrum stehen algebraische Chemien. Eine algebraische Chemie ist definiert als Tupel $< M,R >$ mit einer Menge M von Molekülen und einer Relation R auf $\mathscr{P}_M(M) \times \mathscr{P}_M(M)$. Dabei ist $\mathscr{P}_M(M)$ wie üblich die Menge aller Multimengen auf der Potenzmenge von M. Beispielsweise seien $H_2, O_2, H_2O \in M$. Für die chemische Reaktion $2H_2 + O_2 \rightarrow 2H_2O$ gilt dann $(\{H_2, H_2, O_2\}, \{H_2O, H_2O\}) \in R$. Die Relation gibt also an, welche Reaktionen stattfinden können.
Diese Betrachtungsweise kann auch auf Aktivierungen und Inhibierungen in den vorher modellierten biochemischen Netzwerken von *C. glutamicum* und *M. tuberculosis* etwas modifiziert angewendet werden. Gen a aktiviere Gen b, und es müssen mindestens k_1 Moleküle des Proteins von Gen a vorliegen, bevor eine schwache Aktivierung möglich ist bzw. mindestens k_2 Moleküle, um die volle Aktivierung zu erhalten. Entspricht P_i einem Molekül des Proteins von Gen i für $i = a$ oder $i = b$, so lässt sich diese Situation mit den folgenden Reaktionen beschreiben:

$$\begin{aligned} &kP_a + P_b \rightarrow kP_a + P_b \text{ für } 1 \leq k < k_1, \\ &kP_a + P_b \rightarrow kP_a + lP_b \text{ für } k_1 \leq k < k_2 \text{ und} \\ &kP_a + P_b \rightarrow kP_a + mP_b \text{ für } k_2 \leq k. \end{aligned} \tag{7.1}$$

Hierbei sind $k,l,m \in \mathbb{N}^+$, mit $l < m$ und $k_1,k_2 \in \mathbb{N}^+$ zwei feste Schwellwerte. In der ersten Reaktion hat man keine Veränderung der Proteinmenge P_b, in der zweiten eine Erhöhung dieser Menge um $l-1$ Proteine, und in der dritten Reaktion eine Erhöhung um $m-1$ Proteine. Weitere Abstufungen sind möglich, indem Zwischenschritte zwischen k_1 und k_2 eingefügt werden.
Inhibierungen können analog definiert werden:

$$\begin{aligned} &kP_a + P_b \rightarrow kP_a + P_b \text{ für } 1 \leq k < k_1, \\ &kP_a + lP_b \rightarrow kP_a + P_b \text{ für } k_1 \leq k < k_2 \text{ und} \\ &kP_a + mP_b \rightarrow kP_a + P_b \text{ für } k_2 \leq k, \end{aligned} \tag{7.2}$$

ebenfalls mit $k,l,m \in \mathbb{N}^+$, $l < m$ und $k_1,k_2 \in \mathbb{N}^+$.
Im Folgenden soll erläutert werden, wie die Verhaltensweisen des Modells der Stickstoffaufnahme, das in Kapitel 5 entwickelt wurde, über die chemische Organisationstheorie visualisiert werden kann. In diesem speziellen Modell hat GlnD zwei verschiedene Funktionen, die separat betrachtet werden müssen. Man erhält also zwei algebraische Chemien $< M_C, R_{C1} >$ und $< M_C, R_{C2} >$. Diese Aufteilung bezüglich der Funktion von GlnD vereinfacht jedoch die Aktivierungsreaktionen, da diese in Abhängigkeit der Funktion von GlnD aufgestellt werden können. Die Aktivierung sei in unserem Fall maximal, so dass keine weitere Fallunterscheidung bezüglich der Anzahl der regulierenden Moleküle nötig ist.
Die Molekülmenge M_C hat die Form

$$M_C = \{\text{GlnD},\ \text{GlnK},\ \text{GlnK} \sim \text{AMP},\ \text{AmtB}\}.$$

Die Relationen R_{C1} und R_{C2} sehen je nach Funktion von GlnD wie folgt aus:

- R_{C1} (GlnD wandelt GlnK in GlnK∼AMP um):
  1. GlnD + GlnK → GlnK∼AMP + GlnD,
  2. GlnK∼AMP → 2 GlnK∼AMP,
  3. GlnK∼AMP + AmtB → GlnK∼AMP + 2 AmtB,
  4. GlnK + AmtB → AmtB,
  5. ∅ → GlnK,
  6. ∅ → AmtB,
  7. GlnK → ∅,
  8. AmtB → ∅,
  9. GlnK∼AMP → ∅.
- R_{C2} (GlnD wandelt GlnK∼AMP in GlnK um):
  1. GlnD + GlnK∼AMP → GlnK + GlnD,
  2. GlnK + AmtB → AmtB,
  3. ∅ → GlnK,

4. $\emptyset \rightarrow$ AmtB,
5. GlnK $\rightarrow \emptyset$,
6. AmtB $\rightarrow \emptyset$,
7. GlnK$\sim$AMP $\rightarrow \emptyset$.

Diese Gleichungen gelten jedoch nur für den Wildtyp. In der *amtB*-Mutante würde beispielsweise Reaktion 3 der Relation R_{C1} bzw. Reaktion 2 der Relation R_{C2} wegfallen.
Um den Verband zur Visualisierung zu bilden, müssen zunächst verschiedene Eigenschaften von algebraischen Chemien $< M,R >$ definiert werden.

Definition 7.1. *(abgeschlossen, (Dittrich u. di Fenizio, 2005))*
Eine Menge $S \subseteq M$ ist abgeschlossen genau dann, wenn für alle $(A,B) \in R$ mit $A \in \mathscr{P}_M(S)$ auch $B \in \mathscr{P}_M(S)$ gilt.

Das bedeutet, eine abgeschlossene Menge S enthält nur Moleküle, die wiederum Moleküle produzieren, welche schon in S enthalten sind.
Für den Fall $< M_C, R_{C1} >$ gibt es insgesamt 14 abgeschlossene Mengen, und zwar $\{S \subseteq M\} \setminus \{\{\text{GlnD, GlnK}\}, \{\text{GlnD, GlnK, AmtB}\}\}$. Nur wenn GlnD und GlnK zusammen in einer Menge vorkommen, kann das neue Molekül GlnK$\sim$AMP produziert werden, so dass diese Mengen nicht abgeschlossen sind, falls GlnK$\sim$AMP nicht auch bereits in dieser Menge enthalten ist.

Definition 7.2. *(massenerhaltend, (Dittrich u. di Fenizio, 2005))*
Eine Menge $S \subseteq M$ heißt massenerhaltend genau dann, wenn alle Moleküle mit nichtnegativer Rate produziert werden können.

Mathematisch betrachtet bedeutet es folgendes: Es sei $\mathscr{M}$ die stöchiometrische Matrix, wobei die Zeilen den Molekülen entsprechen und die Spalten den einzelnen Reaktionen. Der Eintrag m_{ij} gibt an, wieviele Moleküle vom Typ i in Reaktion j verbraucht bzw. produziert werden. Als Beispiel betrachten wir $< M_C, R_{C1} >$ wie oben definiert. Die Matrix $\mathscr{M}$ hat dann die Form

$$\mathscr{M} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & -1 & 0 \end{pmatrix}.$$

Es sei n die Anzahl der Reaktionen. $C \subseteq M$ ist genau dann massenerhaltend, wenn ein Vektor $v \in \mathbf{R}^n$ existiert, so dass gilt:

1. $\forall (A,B) \in R$ mit $A \in \mathscr{P}_M(C) \Rightarrow v_{(A,B)} > 0$,
2. $\forall (A,B) \in R$ mit $A \notin \mathscr{P}_M(C) \Rightarrow v_{(A,B)} = 0$,
3. $\mathscr{M} v \geq 0$.

Der Index (A,B) lässt sich wie folgt lesen. Jeder Reaktion $A \rightarrow B$ entspricht eine Spalte der stöchiometrischen Matrix $\mathscr{M}$. Ist $A \rightarrow B$ der Spalte k zuzuordnen, so ist $v_{(A,B)}$ der k-te Eintrag des

Vektors v. Die massenerhaltenden Mengen der algebraischen Chemie $< M_C, R_{C1} >$ sind alle 16 möglichen Teilmengen von M_C. Man betrachte zum Beispiel die Molekülmenge {GlnD, GlnK}. Die Spalten der Matrix $\mathscr{M}$, die die Reaktionen beschreiben, bei denen {GlnD}, {GlnK}, {GlnD, GlnK} oder $\emptyset$ auf der linken Seite der Reaktion stehen, sind die Spalten 1, 5, 6 und 7. Demnach muss nun ein Vektor $v \in \mathbf{R}^9_+$ gefunden werden, für den gilt:
i) $v_i > 0$ für $i = 1,5,6,7$,
ii) $v_i = 0$ für $i = 2,3,4,8,9$,
iii) $\mathscr{M} v \geq 0$.
Dies schreiben wir um in folgende Form: Gesucht ist ein Vektor $y \in \mathbf{R}^4_+$, so dass $y_i > 0$ für alle i und $Ay \geq 0$ für die Matrix $A = \begin{pmatrix} -1 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. Sofort sieht man, dass die Bedingungen z. B. durch den Vektor $y = (1,3,1,1)^t$ erfüllt werden.
Somit erfüllt der Vektor $v = (1,0,0,0,3,1,1,0,0)^t$ die Bedingungen i) bis iii).
Für alle weiteren 15 Mengen ist der Beweis der Massenerhaltung analog führbar, siehe Anhang A.2.
Für den Begriff der Abgeschlossenheit benötigt man keine Erweiterung der Theorie für Aktivierungen und Inhibierungen. Bei dem Begriff der Massenerhaltung muss jedoch eine Fallunterscheidung getroffen werden, in welchem Bereich jeweils der Wert k für die Reaktionen (7.1) und (7.2) liegt. Im Fall des betrachteten *C. glutamicums* kann man nach der Aufteilung bezüglich der Funktion von GlnD ohne Zwischenstufen auskommen, so dass hier keine Fallunterscheidungen getroffen werden müssen.

Definition 7.3. *(Organisation, (Dittrich u. di Fenizio, 2005))*
Eine Organisation wird definiert als massenerhaltende und abgeschlossene Menge von Molekülen.

Organisationen sind die einzigen Kombinationen von Molekülen, die in einem Reaktionsbehälter für einige Zeit verbleiben können. Andererseits bedeutet es nicht, dass sämtliche Moleküle der Organisation für immer in dem Reaktionssystem verbleiben, sondern es bedeutet nur, dass man zumindest für eine kurze Zeit jene Molekülmenge in dem System messen kann. Diese Molekülmenge muss jedoch nicht stabil[1] sein. Bisher fließen keine kinetischen Parameter mit ein, die beschreiben, wie schnell die Reaktionen unter den gegebenen Umständen ablaufen. So kann es sein, dass ein Molekül mit der Zeit abgebaut wird, auch wenn man einen Flussvektor definieren kann, der eine nichtnegative Produktionsrate möglich macht.

Im Beispiel von $< M_C, R_{C1} >$ bilden alle abgeschlossenen Mengen auch schon eine Organisation, da jede der hier vorkommenden Mengen massenerhaltend ist.

Betrachtet man die Menge aller möglichen Organisationen, so bilden sie für Reaktionssysteme mit gewissen Voraussetzungen einen Verband. Solch ein Reaktionssystem muss konsistent sein:

[1] Stabilität bedeutet in diesem Zusammenhang, dass selbst nach Ablauf unendlich vieler Reaktionen immer noch dieselben Molekülarten vorhanden sind.

Definition 7.4. *(konsistent, (Dittrich u. di Fenizio, 2005))*
Ein Reaktionssystem $< M,R >$ heißt konsistent, wenn Mengen $P \subseteq M$ und $\bar{P} \subseteq M$ mit $M = P \cup \bar{P}$ und $P \cap \bar{P} = \emptyset$ existieren, so dass gilt

- $\forall i \in \bar{P} : \exists(\{i\} \to \emptyset) \in R$ *oder* $\exists(A \to B) \in R$ *mit* $\#(i \in A) - 1 = \#(i \in B)$*, wobei die Reaktion $A \to B$ erster Ordnung sein muss. (Die Moleküle werden also nur in Reaktionen erster Ordnung aufgebraucht.)*
- $\forall i \in P : \nexists(A \to B) \in R$ *mit* $\#(i \in A) > \#(i \in B)$*. (Diese Moleküle werden in keiner Reaktion aufgebraucht, d.h. sie sind Katalysatoren.)*

Als Veranschaulichung soll folgendes Gegenbeispiel dienen:

Beispiel 7.5. *$M = \{a,b,c\}$, $R = \{a+b \to c, \emptyset \to b, c \to \emptyset\}$. Hier ist das Molekül b dafür verantwortlich, dass $< M,R >$ kein konsistentes Reaktionssystem ist. Es ist kein Katalysator, wird nämlich in einer Reaktion umgewandelt ($a+b \to c$), ist aber auch nicht ein Molekül der ersten Menge $\bar{P}$, da es keine Reaktion in R gibt, in der das Molekül aufgebraucht wird. Somit ist dieses Reaktionssystem nicht konsistent.*

Sei $S \subseteq M$. Im Folgenden bezeichnet $G_{ABG}(S)$ die kleinste abgeschlossene Menge, die S enthält und $G_{ME}(S)$ die größte massenerhaltende Menge, die in S enthalten ist.

Definition 7.6. *(Schnitt und Vereinigung von Organisationen, (di Fenizio u. Dittrich, 2002))*
Es sei $\mathcal{O}$ die Menge aller Organisationen einer konsistenten algebraischen Chemie $< M,R >$ und $G(S) := G_{ME}(G_{ABG}(S))$ für alle $S \subseteq M$. Für $U,V \in \mathcal{O}$ definieren wir
i) $U \sqcup V := G(U \cup V)$,
ii) $U \sqcap V := G(U \cap V)$.

In di Fenizio u. Dittrich (2002) wird gezeigt, dass für $U,V \in \mathcal{O}$ dann auch $U \sqcup V$ und $U \sqcap V$ ebenfalls wieder Organisationen sind. Damit zeigen sie dann folgendes Theorem:

Satz 7.7. *(di Fenizio u. Dittrich, 2002)*
Für konsistente Reaktionssysteme $< M,R >$ bildet $(\mathcal{O}, \sqcup, \sqcap)$ einen Verband.

Für den Beweis siehe wieder di Fenizio u. Dittrich (2002). Die umgekehrte Richtung des Theorems gilt jedoch nicht. Auch für nicht konsistente Reaktionssysteme kann in manchen Fällen ein Organisationsverband gebildet werden. Beispielsweise lässt sich aus den Organisationen der algebraischen Chemie des oben beschriebenen Beispiels eines nicht konsistenten Reaktionssystems dennoch ein Verband bilden.

In Abbildung 7.1 ist der Verband von $\mathcal{O}$ bezüglich $< M_C, R_{C1} >$ zu sehen.

Um Dynamiken in dem Verband veranschaulichen zu können, müssen die Differenzialgleichungen betrachtet werden, mit denen man das Verhalten des Systems beschreibt. Es sei $\dot{x} = f(x)$ dieses System von Differenzialgleichungen. Dabei sei $x \in X$ der Zustand des Systems, das heißt,

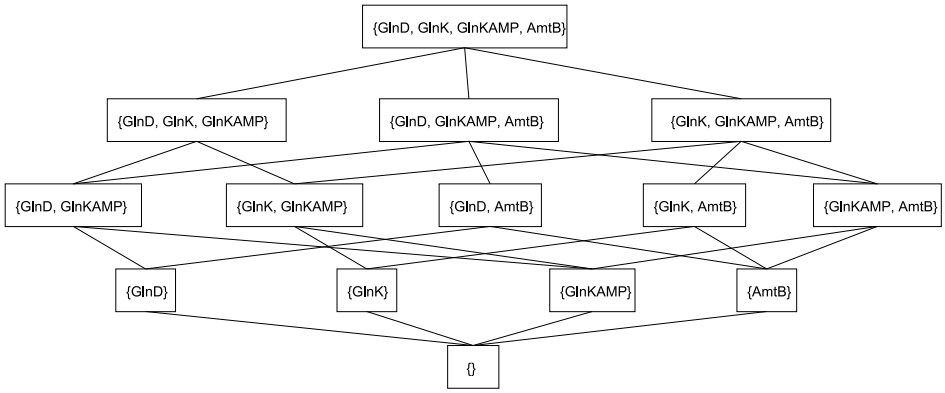


Abbildung 7.1: *Der Verband der Menge aller Organisationen*

wir haben im Vektor $x(t)$ Konzentrationen jedes Proteins zum Zeitpunkt t vorliegen.

Definition 7.8. *(Abstraktion)*
Eine Funktion von den möglichen Zuständen X des Systems in die Molekülmenge M, $\phi : X \to M$, die den Zustand des Systems auf die Menge von Molekülen abbildet, die während dieses Zustands vorhanden sind, heißt Abstraktion.

Vorhanden bedeutet hier, dass die Konzentration des jeweiligen Moleküls eine zu definierende Schranke θ überschreitet, wobei $\theta = 0$ erlaubt ist.

Definition 7.9. *(Instanz, (Dittrich u. di Fenizio, 2005))*
Ein Zustand $x \in X$ heißt Instanz von $A \subseteq M$, wenn $\phi(x) = A$ ist.

Es kann viele verschiedene Instanzen von A geben.

Satz 7.10. *(Dittrich u. di Fenizio, 2005)*
Jeder Fixpunkt ist Instanz einer Organisation.

Zum Beweis siehe Dittrich u. di Fenizio (2005). Anschaulich ist es aber klar, dass jeder Fixpunkt Instanz einer Organisation sein muss. Es werden weder neue Moleküle generiert noch verschwinden bereits vorhandene Moleküle, so dass die zugehörige Molekülmenge abgeschlossen und massenerhaltend sein muss.
Betrachtet man Systeme von Molekülen, so kann man deren Verhalten zwar mit gewöhnlichen Differenzialgleichungen beschreiben, doch solch eine Beschreibung entspricht in manchen Fällen nicht der Realität, siehe auch Dittrich u. di Fenizio (2005). Als Beispiel nennen sie den Fall, dass ein Molekül durch ein anderes aufgebraucht werden kann. Im Molekülbehälter ist dann von dieser Molekülart keines dieser Moleküle mehr vorhanden. Die Differenzialgleichung sagt jedoch nur, dass die Molekülmenge gegen Null geht, wenn die Zeit gegen unendlich geht. Dittrich und Speroni schlagen daher vor, dass man eine Schranke θ für die Differenzialgleichung definieren soll, ab der man sagt, dass das Molekül nicht mehr vorhanden sei. Somit bewegt sich das

System von einem Zustand, in dem einige Moleküle vorhanden sind, zu einem anderen Zustand, in dem eine andere Menge von Molekülen vorhanden ist. Solch eine Bewegung soll in dem Organisationsverband eingezeichnet werden, um die Dynamik eines Systems zu veranschaulichen. Es werden zwei verschiedene Bewegungen eingezeichnet:

1. Abwärtsbewegung: Diese Bewegung ist zu sehen, wenn eine Molekülart unter den vorgegebenen Schwellwert θ fällt.

2. Aufwärtsbewegung: Diese Bewegung ist zu sehen, wenn neue Molekülarten im System erscheinen. Hier muss festgelegt werden, auf welche Weise die Moleküle in das System hineinkommen (da sie ja aufgrund der Abgeschlossenheit der Organisationen nicht durch bereits bestehende Moleküle produziert werden). Man nimmt hier zufällige Störungen des Systems oder Eingriffe durch außen an.

In der folgenden Abbildung 7.2 sind die Auf- und Abwärtsbewegungen in dem Verband eingezeichnet, wenn man unter Aufwärtsbewegung versteht, dass nur *ein* neuer Molekültyp auftauchen kann.
Wie man sieht, ist die Visualisierung der Dynamik für größere Netze eher unübersichtlich. Da

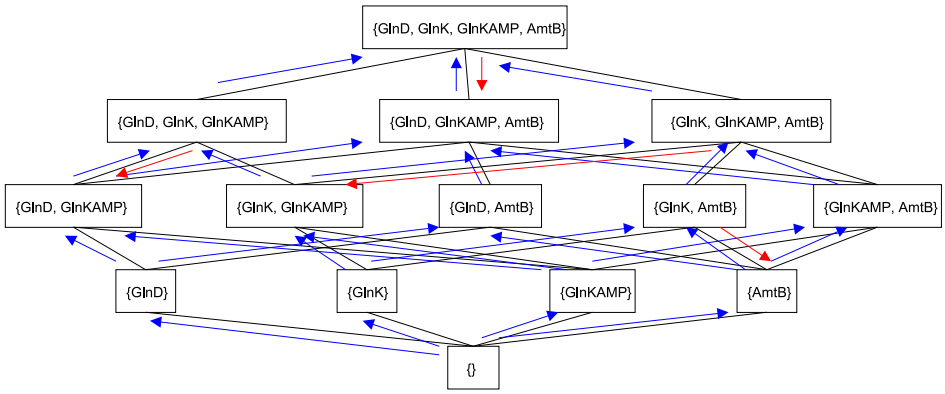


Abbildung 7.2: *Der Verband der Menge aller Organisationen, in dem die Auf- und Abwärtsbewegungen eingezeichnet sind.*

die Aufwärtsbewegungen willkürlich definiert werden, wird in der folgenden Abbildung 7.3 dasselbe Bild nur mit Abwärtsbewegungen gezeigt.

7.2 Formale Begriffsanalyse

Anstelle der chemischen Organisationstheorie kann jedoch auch ein Ansatz zur Visualisierung verwendet werden, der wiederum auf der formalen Begriffsanalyse basiert. Diese weitere Möglichkeit der Visualisierung von sich zeitlich verändernden Systemen ist von Wolff erarbeitet worden (Wolff, 2000, 2001, 2002). Zusätzlich zu den in Kapitel 3.2 beschriebenen Grundbegriffen der Begriffsanalyse benötigen wir hier noch mehrwertige Kontexte und Skalierungen derselben.

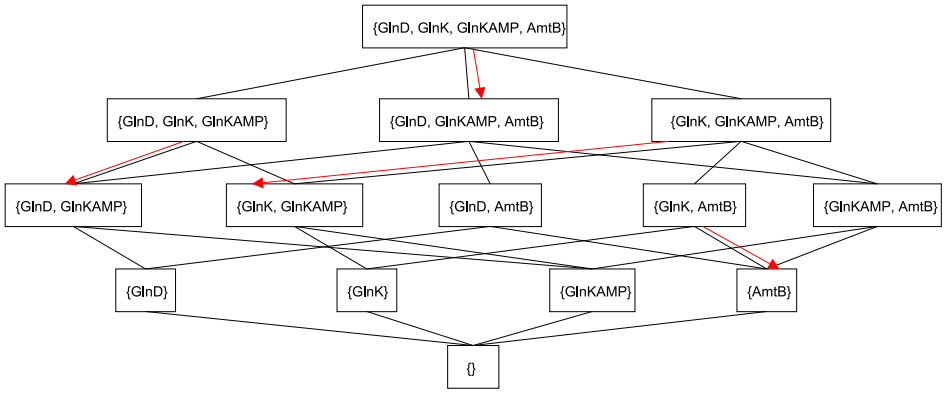


Abbildung 7.3: *Der Verband der Menge aller Organisationen mit Abwärtsbewegungen.*

Definition 7.11. *(mehrwertige Kontexte, (Wolff, 2001))*
Das Quadrupel (G,M,W,I) *besteht aus einer Menge G von Gegenständen, einer Menge M von Merkmalen, einer Menge W von Werten und einer ternären Relation* $I \subseteq G \times M \times W$*, wobei für alle* $g \in G$ *und für alle* $m \in M$ *höchstens ein* $w \in W$ *existiert mit* $(g,m,w) \in I$*. Das Quadrupel* $\mathbb{K} := (G,M,W,I)$ *nennen wir mehrwertigen Kontext.*

Dabei bedeutet $(g,m,w) \in I$, dass Gegenstand g das Merkmal m mit Wert w hat. Über diese Relation lässt sich eine partielle Funktion $m : G \rightarrow W$ für jedes Merkmal $m \in M$ bilden, so dass m(g)=w gilt, falls $(g,m,w) \in I$.

Definition 7.12. *(begriffliche Skala, (Wolff, 2001))*
$S_m = (W_m, M_m, I_m)$ *heißt begriffliche Skala für das Merkmal* $m \in M$*, wenn* $W_m \supseteq m(G) := \{m(g) : g \in G\}$ *gilt.*

Im folgenden Beispiel wird ein mehrwertiger Kontext von Genexpressionsdaten skaliert, welcher in Tabelle 7.1 beschrieben ist. In der Mikroarrayanalyse werden üblicherweise nur Gene betrachtet, die entweder über zweifach verstärkt bzw. weniger als halb so stark exprimiert wurden. Daher bietet sich die Skalierung mit < 0.5 und > 2.0 an. Da keines der Gene unter 0.5 fällt, wählen wir alternativ die Skalierung < 1.0 und > 2.0. Daher bietet sich für alle Merkmale dieselbe Skala S an, die in Tabelle 7.2 dargestellt ist. Aufgrund der Skalierung erhält man den in Tabelle 7.3 aufgeführten Kontext. Ein skalierter mehrwertiger Kontext wird als Paar $((G,M,W,I),(S_m : m \in M))$ beschrieben.

Ebenso wie für die Expressionswerte kann auch eine Skalierung für die Zeitpunkte erfolgen, siehe Tabelle 7.4. Die Einteilung der Zeitpunkte sollte dabei das Verhalten der Expressionszeitreihe berücksichtigen. In unserem Beispiel ist während der ersten drei Zeitpunkte keine starke Veränderung zu der Expression zu sehen. Zu den Zeitpunkten t_3 und t_4 erfolgt eine verstärkte Expression einiger Gene und zu den Zeitpunkten t_5 bis t_7 sind alle Gene bis auf *ruvC* durchgängig stark exprimiert. Beide abgeleiteten Kontexte können nun als Begriffsverband dargestellt werden, siehe Abbildungen 7.4 und 7.5.

Tabelle 7.1: Mehrwertiger Kontext. Die Gegenstände G bestehen aus einer Menge von Zeitpunkten, die Merkmale M aus Genen, die Werte W sind numerische Werte aus $\mathbf{R}_0^+$, und die Relation $(g,m,w) \in I$ bedeutet, dass die Genexpression von Gen m zum Zeitpunkt g die Höhe w hat.

Zeit \ Gen	*recA*	*lexA*	*ruvC*	*linB*
t_0	0.717	0.756	0.815	0.827
t_1	0.988	0.853	0.960	0.558
t_2	0.925	0.801	0.962	0.828
t_3	3.361	1.787	1.458	2.038
t_4	1.448	1.122	2.068	1.192
t_5	11.259	3.845	11.547	1.686
t_6	8.710	4.914	4.571	1.844
t_7	10.094	3.450	3.466	2.396

Tabelle 7.2: Skalierung für die Genexpressionswerte aus Tabelle 7.1.

Wert	<1.0	>2.0	Wert	<1.0	>2.0
0.717	x		0.756	x	
0.815	x		0.827	x	
0.988	x		0.853	x	
0.960	x		0.558	x	
0.925	x		0.801	x	
0.962	x		0.828	x	
3.361		x	1.787		
1.458			2.038		x
1.448			1.122		
2.068		x	1.192		
11.259		x	3.845		x
11.547		x	1.686		
8.710		x	4.914		x
4.571		x	1.844		
10.094		x	3.450		x
3.466		x	2.396		x

Tabelle 7.3: Aufgrund der Skala in Tabelle 7.2 abgeleiteter Kontext.

Zeitpunkt \ Gen	*recA*		*lexA*		*ruvC*		*linB*	
	<1.0	>2.0	<1.0	>2.0	<1.0	>2.0	<1.0	>2.0
t_0	x		x		x		x	
t_1	x		x		x		x	
t_2	x		x		x		x	
t_3		x				x		
t_4								x
t_5		x		x				x
t_6		x		x				x
t_7		x		x		x		x

Tabelle 7.4: Abgeleiteter Kontext bezüglich einer Skalierung der Zeitpunkte t_i, $i = 0,...,7$, über $i \geq 3$ und $i \geq 5$.

Zeitpunkt	t_i mit $i \geq 3$	t_i mit $i \geq 5$	Zeitpunkt	t_i mit $i \geq 3$	t_i mit $i \geq 5$
t_0			t_4	x	
t_1			t_5	x	x
t_2			t_6	x	x
t_3	x		t_7	x	x

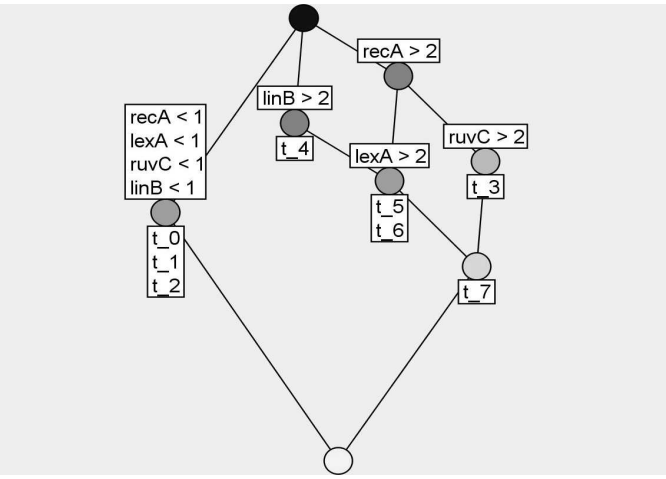


Abbildung 7.4: Begriffsverband des abgeleiteten Kontextes aus Tabelle 7.3.

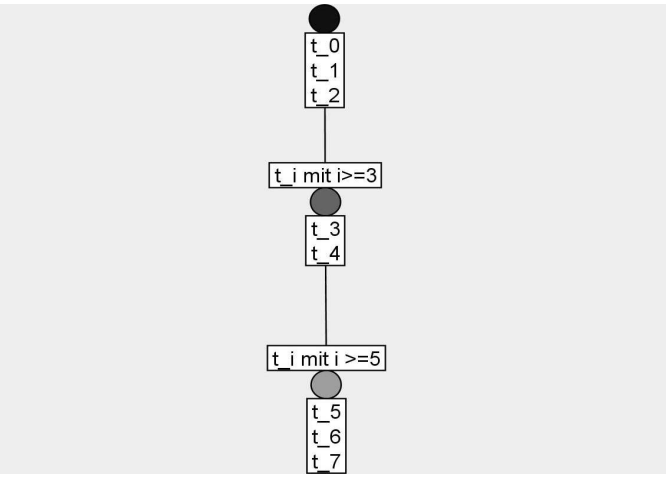


Abbildung 7.5: Begriffsverband des abgeleiteten Kontextes aus Tabelle 7.4.

Ein begriffliches Zeitsystem nach Wolff (2000) erhält man, indem ein skalierter mehrwertiger Kontext T zur Beschreibung der Zeit und ein skalierter mehrwertiger Kontext C zur Beschreibung der Ereignisse zusammengeführt werden.

Definition 7.13. *(begriffliches Zeitsystem, (Wolff, 2000))*
Es sei G eine Menge. $T := ((G,M,W,I_T),(S_m : m \in M))$ und $C := ((G,E,V,I_C),(S_e : e \in E))$ seien zwei skalierte mehrwertige Kontexte. (T,C) heißt dann begriffliches Zeitsystem auf G mit T als Zeitteil und C als Ereignisteil.

In unserem Fall von Genexpressionszeitreihen ist der bezüglich der Zeit skalierte Kontext der Zeitteil und der bezüglich der Genexpressionswerte skalierte Kontext der Ereignisteil.

Definition 7.14. *(Zustand, Zustandsraum eines begrifflichen Zeitsystems, (Wolff, 2000))*
Es sei (T,C) ein begriffliches Zeitsystem und K_T und K_C seien die abgeleiteten Kontexte von T bzw. C. Dann wird für jedes $g \in G$ der Zustand $s(g)$ durch den Gegenstandsbegriff von g in K_C definiert, also $s(g) = \gamma_C(g)$.
Der Zustandsraum von (T,C) definiert sich dann als $S(T,C) = \{s(g) : g \in G\}$. Den Begriffsverband $\mathscr{B}(\mathbb{K}_C)$ nennen wir verallgemeinerten Zustandsraum .

$\mathscr{B}(\mathbb{K}_C)$ kann neben sämtlichen Zuständen $s(g)$, die ja Gegenstandsbegriffe sind, noch weitere Begriffe enthalten, was die Bezeichnung ‘verallgemeinerter Zustandsraum’ rechtfertigt.

Bemerkung 7.15. *Der Zustandsraum des Beispiels besteht aus allen Begriffen des Begriffsverbands aus Abbildung 7.4 ausgenommen der drei Begriffe $(G,\emptyset)$, $(\emptyset,M)$ und $(\{t_3,t_5,t_6,t_7\},\{recA > 2\})$.*

Der verallgemeinerte Zustandsraum besteht aus allen Begriffen des Begriffsverbandes, der in Abbildung 7.4 dargestellt ist.

Definition 7.16. *(Zeitgranulie, Zeitraum, (Wolff, 2000))*
Es sei (T,C) ein begriffliches Zeitsystem und K_T und K_C seien die abgeleiteten Kontexte von T bzw. C. Dann wird für jedes $g \in G$ die Zeitgranulie $t(g)$ durch den Gegenstandsbegriff von g in $\mathbb{K}_T$ definiert, also $t(g) = \gamma_T(g)$.
Der Zeitraum von (T,C) definiert sich dann als $G(T,C) = \{t(g) : g \in G\}$. Den Begriffsverband $\mathscr{B}(\mathbb{K}_T)$ nennen wir verallgemeinerten Zeitraum.

Auch $\mathscr{B}(\mathbb{K}_T)$ kann neben den Zeitgranulien weitere Begriffe enthalten, was zu der Bezeichnung 'verallgemeinerter Zeitraum' führt.

Bemerkung 7.17. *Der Zeitraum des Beispiels besteht nur aus den drei Begriffen $(\{t_0,...,t_7\},\emptyset)$, $(\{t_3,...,t_7\},\{t_i \text{ mit } i \geq 3\})$ und $(\{t_5,t_6,t_7\},\{t_i \text{ mit } i \geq 3, t_i \text{ mit } i \geq 5\})$.*

In diesem Fall stimmt der Zeitraum mit dem verallgemeinerten Zeitraum überein.

Definition 7.18. *(Phase, Phasenraum eines begrifflichen Zeitsystems, (Wolff, 2000))*
Es sei (T,C) ein begriffliches Zeitsystem, und K_T, K_C seien die abgeleiteten Kontexte von T bzw. C. Für jedes $g \in G$ wird das Paar $(t(g),s(g)) \in \mathscr{B}(\mathbb{K}_T) \times \mathscr{B}(K_C)$ eine Phase von (T,C) genannt. Den Phasenraum bilden dann sämtliche solcher Paare, also $P(T,C) = \{(t(g),s(g)) : g \in G\}$. Das direkte Produkt $\mathscr{B}(K_T) \times \mathscr{B}(K_C)$ heißt verallgemeinerter Phasenraum.

Der verallgemeinerte Phasenraum kann auch folgendermaßen dargestellt werden: $\mathscr{B}(K_T)$ wird als Begriffsverband wie in Abbildung 7.5 gezeichnet, jedoch mit vergrößerten Knoten. In diese vergrößerten Knoten wird jeweils der Begriffsverband $\mathscr{B}(K_C)$ eingezeichnet, siehe Abbildung 7.6. Die grobe Struktur enthält also den Zeitraum, die feine Struktur den Zustandsraum. Die gestrichelten Pfeile in der Abbildung geben die Dynamik des Systems wieder. Es befindet sich zu den Zeitpunkten t_0, t_1 und t_2 in dem Zustand, dass die Expression aller vier Gene <1 ist (oberer Knoten der groben Struktur und linker Knoten der feinen Struktur). Im mittleren Knoten der groben Struktur lassen sich die Bewegungen zu den Zeitpunkten t_3 und t_4 ablesen. Zunächst gilt ruvC>2 und recA>2, anschließend zum Zeitpunkt t_4 linB>2. Hier befinden wir uns in dem Zustandsraum also am Begriff $(\{t_4,t_5,t_6,t_7\},\{linB > 2\})$. Von diesem Begriff aus startet man im unteren Knoten der groben Struktur, der die Dynamik während der Zeitpunkte t_5, t_6 und t_7 beschreibt. Hier bewegt sich das System über einen Zwischenzustand zu dem Endzustand, in dem die Expression sämtlicher vier Gene >2 ist.
Solch eine Darstellung wird begrifflicher Film genannt (Wolff, 2001). Jeder Knoten der groben Struktur entspricht einem Foto. Das erste Foto ist 'scharf', die letzten beiden 'unscharf'. Mit 'unscharf' ist gemeint, dass man mehrere Phasen pro Foto sehen kann. Die 'unscharfen' Fotos können verschärft werden, indem die Granularität für die Zeit verfeinert wird. Einen 'scharfen' begrifflichen Film erhält man, wenn jedes Foto nur genau einer Phase entspricht. Allerdings würde dies zu Lasten der Übersichtlichkeit gehen.

Ein Vorteil dieser Art von Visualisierung ist, dass beliebig geordnete Zeiten erlaubt sind. Bisher sind wir von einer linearen Struktur $t_0 \leq t_1 \leq t_2$... ausgegangen. Hier sind jedoch auch

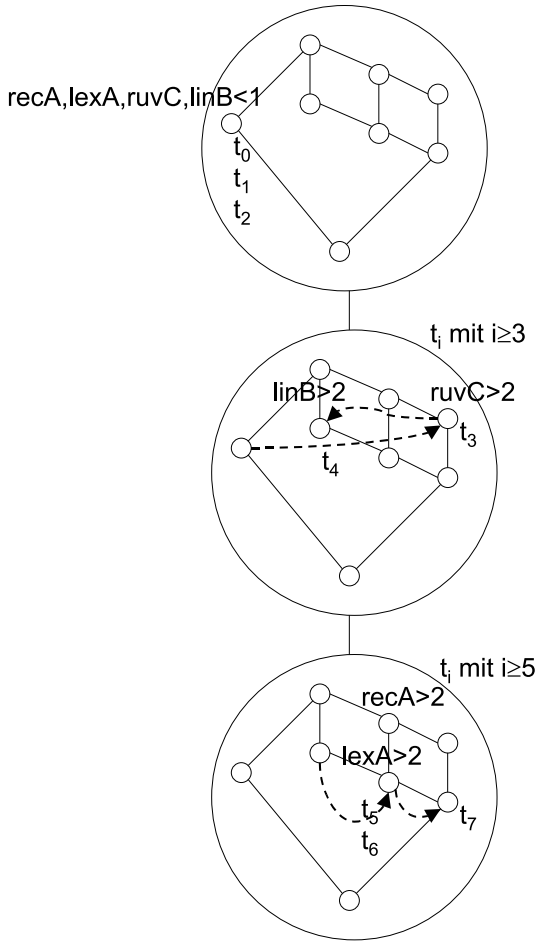


Abbildung 7.6: Direktes Produkt der Begriffsverbände aus Abbildungen 7.4 und 7.5, welches den verallgemeinerten Phasenraum darstellt. Die Zeitpunkte t_0, t_1 und t_2 sind im oberen Knoten zu finden, die Zeitpunkte t_3 und t_4 im mittleren Knoten und die Zeitpunkte t_5, t_6 und t_7 im unteren Knoten. Die jeweiligen Bewegungen von einem Begriff zum nächsten sind durch gestrichelte Pfeile eingetragen.

Strukturen erlaubt wie t_0 hat Merkmal 1, t_1 hat Merkmal 2 etc. Ein Beispiel hierfür wäre eine Zeiteinteilung des Tages durch morgens, mittags, abends, nachts und tagsüber. Einen weiteren Vorteil bietet diese Visualisierung, wenn sich Zustände in einem gewissen zeitlichen Abstand wiederholen. Diese Zustände fallen dann in einen gemeinsamen Begriff und werden in dieser Art der Visualisierung zusammengefasst dargestellt.
Beiden hier vorgestellten Visualisierungsmethoden ist gemein, dass qualitative Verhaltensweisen stärker herausgestellt werden können als beispielsweise bei der Visualisierung, in der die verschiedenen Variablen $x_i(t)$ für $i = 1,...,n$ über t aufgetragen werden. Dies kann insbesondere für größere als die hier vorgestellten Systeme einen guten Überblick über die möglichen Verhaltensweisen geben.

Kapitel 8

Diskussion und Ausblick

Die vorliegende Arbeit lässt sich in den Rahmen der Systembiologie einordnen, deren Zielsetzung unter anderem die Konstruktion einer virtuellen Zelle bzw. virtueller Organe ist. Wird dieses Ziel erreicht, so eröffnet es die Möglichkeit, Medikamentenwirkungen für jeden Menschen spezifisch vorherzusagen. Durch solch eine spezifische Medikamentierung ist als positives Resultat eine Verringerung der Nebenwirkungen zu erwarten. Das Verständnis der genregulatorischen Netzwerke kann ebenfalls neue Impulse zum Verständnis der Entwicklungsbiologie im Rahmen der Zelldifferenzierungen geben. Hier muss man sich jedoch auch der Gefahren bewusst sein, die solch ein Fortschritt in sich birgt. Neben Organen oder Gewebe könnten beispielsweise auch neuartige Viren mit gezielt eingebauten Wirkungsmechanismen gezüchtet werden. Wir möchten an dieser Stelle daher nur kurz darauf verweisen, dass wissenschaftliche Entwicklungen und ethische Diskussionen Hand in Hand gehen müssen.

In dieser Arbeit wurde ein Modell eingeführt, das genregulatorische und allgemeinere biochemische Netzwerke mit Systemen von stückweise linearen Differenzialgleichungen modelliert. Hierbei wurde versucht, eine Balance zwischen Detailliertheit und Einfachheit des Modells zu finden. Es abstrahiert zum einen die biochemischen Vorgänge auf einem mathematisch handhabbaren Niveau, zum anderen ist es stark an der Realität orientiert, das heißt, es ist im Gegensatz zu anderen einfachen Modellen biologisch motiviert. Hierzu wurden die zugrunde liegenden nichtlinearen Regulierungsfunktionen auf einfache, nämlich stückweise lineare Weise vereinfacht, was wiederum biologisch sinnvolle Ergebnisse liefert. Durch die stückweise lineare Modellierung erhält man also ein nichtlineares Modell, welches jedoch in einer Umgebung um den jeweiligen Zustand linear ist (ausgenommen der Schwellwerte). Dadurch lassen sich in biologischen Systemen beobachtbare Verhaltensweisen wie beispielsweise Multistationarität modellieren und das Modell erlaubt eine einfache Parameterschätzung, die Nebenbedingungen berücksichtigen kann. Den Parametern in diesem Modell sind außerdem biologische Bedeutungen zugeordnet. Sie bieten über die Begriffe Syntheserate, Abbaurate oder maximale Regulierungsstärke in interdisziplinären Projekten eine gute Erklärungsmöglichkeit, so dass Simulationen mit veränderten Parametern in Absprache mit allen Projektteilnehmern erstellt werden können.

Das Modell wurde auf drei Systeme angewendet. Zwei hiervon sind in dieser Arbeit aufgeführt. Diese zeigen, dass die Konstruktion eines Modells mithilfe biologischen Wissens bereits aufgrund weniger Messwerte möglich ist. Eine der Anwendungen beschäftigte sich mit der Repression des *bgl*-Operons durch das Protein H-NS. Eine dreifache Erhöhung der Transkriptions-

rate zeigt in Experimenten eine 100fach stärkere Expression des entsprechenden Gens. Diese Beobachtung konnte durch die Modellierung mit zwei Feedbackschleifen erklärt werden. Die zweite und dritte Anwendung ist in Kapitel 5 und Kapitel 6 beschrieben. Bei der Stickstoffaufnahme in *C. glutamicum* wurde ein kleines Netzwerk modelliert, in dem auch posttranslationale Prozesse eine wichtige Rolle spielen. In den anschließenden Simulationen wurde der Schwerpunkt auf eine Mutante und verschiedene Proteasenaktivitäten gelegt, so dass Aussagen über das Verhalten des Systems unter verschiedenen Anfangsbedingungen gemacht werden konnten. Ein Vergleich der Simulationen mit weiteren Daten zeigte, dass vermutet werden kann, der Abbau eines der beteiligten Proteine könnte auf zwei verschiedenen Wegen vor sich gehen. Bei der dritten Anwendung, der Modellierung des DNA-Reparatursystems in *M. tuberculosis*, wurde der Schwerpunkt auf die Komponentenauswahl gelegt. Hier erfolgten zwei verschiedene Modellierungen, die auf Grundlage verschiedener Variablenmengen zustande kamen. Eine der Modellierungen zeigte eine bessere Simulation des Experiments, aus dem die Daten gewonnen wurden, und ließ die Schlussfolgerung zu, dass Gen *Rv2719c* einen möglicherweise starken Einfluss auf das DNA-Reparatursystem hat.

In dieser letzten Anwendung wird deutlich, dass die Variablenauswahl der Modelle eine wichtige Rolle bei der Vorhersagequalität eines Modells spielt. Zwei verschiedene Auswahlkriterien sind in dieser Arbeit vorgestellt worden. Beide Methoden verwenden eine Liste von Genen, deren Korrelationskoeffizienten zu den Quellgenen berechnet werden, und nur statistisch signifikante Gene werden weiter analysiert. Um diese Liste von Genen zu erstellen, sucht die graphentheoretische Methode kürzeste Wege in einem Interaktionsgraphen, wohingegen die begriffsanalytische Methode Zwischenbegriffe in Begriffsverbänden verwendet. Der Vorteil der zweiten Methode liegt darin, dass biologisches Wissen über die Erstellung von Mustern mit einfließen kann. Beiden Methoden ist gemein, dass Interaktionsdaten nicht zur Berechnung der Ausgangsliste von Genen notwendig sind, jedoch mit berücksichtigt werden können. Allerdings müssen Parameter gesetzt werden, die die Größe der Genliste bestimmen.
Der Fragestellung nach Auswahlkriterien der Variablen eines Modells ist bisher wenig Beachtung geschenkt worden, jedoch ist eine möglichst gute Auswahl notwendig für ein vorhersagekräftiges Modell.

Um schlussendlich verschiedene Schwerpunkte in der Visualisierung setzen zu können, wurden zwei Visualisierungsmethoden auf die bereits betrachteten Subsysteme von *C. glutamicum* und *M. tuberculosis* angewandt, die jeweils auf Verbänden basieren. Um die Dynamik für sämtliche Anfangszustände sichtbar zu machen, wurde die chemische Organisationstheorie auf das Stickstoffaufnahmesystem angewandt, und um Zustände als Begriffe aufzufassen, wurde die Begriffsanalyse auf das DNA-Reparatursystem angewandt. Solche Visualisierungsmethoden können insbesondere qualitative Verhaltensweisen gut widerspiegeln.

Modellierungsarbeiten wie in Kapitel 5 oder 6 lassen sich oft nicht als beendet einstufen, da die Simulationen neue Fragen aufwerfen, die über neue Experimente geklärt werden müssen. Dies lässt sich sehr gut am Modell der Stickstoffaufnahme des *C. glutamicum* erkennen. Hier wird aufgrund der Simulationen vermutet, dass ein weiterer Abbauweg von GlnK∼AMP sowie eine schnellere Umwandlung von GlnK∼AMP zu GlnK in der *amtB*-Mutante wirken könnten.

Um die Verringerung der GlnK∼AMP-Konzentration detaillierter beschreiben zu können, sind daher weitere Experimente notwendig, die den Abbau von GlnK∼AMP klären. Auch bei der Modellierung des DNA-Reparatursystems in *M. tuberculosis* sind neue Experimente notwendig, die überprüfen, welche Rolle das Gen *Rv2719c* in dem DNA-Reparatursystem spielt. Insofern kann eine Modellierung und die daraus resultierenden Simulationen als Grundlage für weitere Experimente dienen. Ein Modell kann jedoch selten als vollständig in dem Sinn angesehen werden, dass das Modell weder Erweiterungen noch Verfeinerungen nötig hat.

Die vorliegende Arbeit kann in verschiedene Richtungen weiterentwickelt werden, die im Folgenden kurz skizziert werden:

Als additives Modell ist das hier vorgestellte Modell weiterhin eine starke Vereinfachung der Wirklichkeit, die in Anbetracht der jetzigen Datenlage sinnvoll scheint. Dennoch kann davon ausgegangen werden, dass längere Zeitreihen, die detailliertere Modelle zulassen, in nächster Zukunft vorhanden sein werden. Ein multiplikatives Modell könnte in zwei Schritten entwickelt werden. Im ersten Schritt könnten mögliche Komplexbildungen von Proteinen vorhergesagt werden, um im zweiten Schritt die Parameter eines Modells zu schätzen, das neben den hier erwähnten Regulierungsfunktionen auch Funktionen, die die Komplexbildung beschreiben, beinhaltet.

Neben der Verwendung von Zeitreihen ist die Einbindung verschiedener Datenbanken zur Schätzung der Parameter sinnvoll. Das Wissen über Bindestellen zusammen mit den entsprechenden Transkriptionsfaktoren kann verwendet werden, um die Struktur des Netzwerkes im Vorfeld zu bestimmen. Datenbanken, in denen Wissen über bekannte Proteinkomplexe zu finden ist, können verwendet werden, um im multiplikativen Modell die Regulierungsfunktionen im Vorfeld der Modellierung zu parametrisieren.

In der vorliegenden Arbeit wurden gewöhnliche Differenzialgleichungen als Modellklasse verwendet. Um jedoch auch räumliche Inhomogenitäten beschreiben zu können, ist die Klasse der partiellen Differenzialgleichungen heranzuziehen. Diffusion ist ein Prozess, in dem Moleküle sich in Bereiche der Zelle bewegen, in denen eine niedrigere Konzentration dieses Moleküls herrscht (Bormann u. a., 2001). Im bisherigen Modell wurde angenommen, dass die Moleküle in der Zelle gleichmäßig verteilt sind. Insbesondere bei Eukaryonten ist diese Annahme stark vereinfachend. Dennoch ist auch bei Prokaryonten eine gleichmäßige Verteilung der im Modell einbezogenen Moleküle meist nicht gegeben. Laktose, also eine Zuckerart, wird beispielsweise durch ein Enzym, das Permease genannt wird, durch die Zellwand transportiert. Je nach Konzentration oder Aktivität der Permease ändert sich auch die Konzentration von Laktose innerhalb der Zelle. Allerdings ist die Laktose zunächst an der Zellmembran und verteilt sich dann erst in der Zelle. Eine Erhöhung der Permeasenkonzentration hat also zunächst eine Erhöhung der Laktosekonzentration in dem Bereich nahe der Zellmembran zur Folge. Aufgrund der Diffusion der Moleküle ist anschließend auch eine Erhöhung der Laktosekonzentration in der gesamten Zelle zu erwarten. Der Diffusionskoeffizient gibt an, wie schnell ein Molekül diffundiert. Er ist unter anderem abhängig von der Flüssigkeit, durch die das Molekül sich bewegt und von der Form des Moleküls. Auch räumlich getrennte Regulierungsprozesse können damit modelliert werden.

Das System, das modelliert werden soll, muss jedoch sehr detailliert bekannt sein. Selbst wenn Kenntnisse über Regulierungsfunktionen sowie kinetische Parameter in den einzelnen Bereichen vorhanden sind, müssen dennoch sehr viele Annahmen - beispielsweise bei der Schätzung der Diffusionskoeffizienten - gemacht werden. Trotzdem ist auch dies ein Schritt in die Richtung, ganze Zellen oder sogar ganze Organe oder Organismen auf molekularer Ebene zu modellieren. Insbesondere in der Entwicklungsbiologie ist eine verstärkte Anwendung von Modellen, die räumliche Aspekte beinhalten, zu erkennen.

Verschiedene Methoden zur Komponentenauswahl, wie sie in dieser Arbeit zu finden sind, werden in Zukunft vermutlich eine größere Rolle spielen, da selbst bei großen Netzwerken das Modell zur besseren Handhabbarkeit oft in kleinere Module zerlegt wird und somit - selbst wenn in Zukunft gesamte Genome modelliert werden - auch hier eine Methode zur Variablenauswahl für die kleineren Module notwendig ist. Eine mögliche Erweiterung betrifft die Bestimmung der zeitlichen Verschiebung des zeitlich verschobenen, erweiterten Kendallschen Korrelationskoeffizienten. Als weitere Richtung wäre eine Zusammenfassung der bisher als zwei separat erfolgten Schritte denkbar.

Die in dieser Arbeit beschriebenen spezifischen Modelle für *C. glutamicum* und *M. tuberculosis* können in verschiedene Richtungen weiterentwickelt werden.
Um das Modell für *C. glutamicum* zu verbessern, kann beispielsweise die Vereinfachung, die zur Stufenfunktion $s_{2,3}(x_2, x_1)$ (siehe Gleichung (5.14)) führte, zurückgenommen werden, wenn in Zukunft in den Experimenten neben GlnK und GlnK∼AMP auch AmtB gemessen werden kann. Dann nämlich lässt sich die Funktion auch abhängig von AmtB schätzen, so dass die Funktion $s^*_{2,3}(x_2, x_3)$ (siehe Gleichung (5.10)) in dem Modell verwendet werden kann. Des Weiteren kann auch GlnD als Variable in das Modell aufgenommen werden, sobald Messungen hierüber vorliegen. Dadurch könnte die Rückkopplung integriert werden, die aufgrund der indirekten Regulierung von GlnK∼AMP auf GlnD und die bereits modellierte Modifizierung von GlnK in GlnK∼AMP über GlnD besteht. Ein weiteres Ziel ist die Integration des vermuteten, unbekannten Abbaumechanismus von GlnK∼AMP sowie die Modellierung des unbekannten Signalweges von der Stickstoffkonzentration zur GlnD-Konzentration, sobald die Kenntnisse über das System und die Datenlage dies zulassen.
In ähnlicher Weise ist auch das zweite modellierte System erweiterbar. Nicht nur das *recA-lexA*-System reagiert auf zerstörte DNA, sondern auch das *uvrABCD*-System. In *E. coli* ist der Zusammenhang der Systeme bekannt, der sich jedoch nicht auf *M. tuberculosis* übertragen lässt. Da auf diesem Gebiet weitere Forschung betrieben wird, kann das bisherige Modell entsprechend um das *uvrABCD*-System erweitert werden, sobald wiederum die biologischen Kenntnisse und die Datenlage dies erlauben.

Allgemein lässt sich bemerken, dass bei relativ kurzen Zeitreihen, wie sie in dieser Arbeit vorlagen, weiteres Wissen über die Vorgänge innerhalb der Zelle für eine Modellierung existieren muss. Hier kann entweder die Integration verschiedener Datenquellen Abhilfe schaffen oder das Modell kann im interdisziplinären Wissensaustausch erstellt werden, da durch Expertenwissen beispielsweise die Parameter direkt beschränkt oder Netzwerkstrukturen ausgeschlossen oder vorgegeben werden können.

Anhang A

Datengrundlage und ergänzende Ergebnisse

A.1 Daten für das *Mycobacterium tuberculosis*

Verwendete Experimente

Auf der GEO-Plattform unter Accession Number GPL1396 sind die folgenden, für die Modellierung verwendeten 16 Experimente zugänglich:

- 3_10 0_2μg_mL Mtm_ctrl _4h_ 24076,
- 3_15 0_2μg_mL Mtm_ctrl _0_33h_ 24066,
- 3_17 0_2μg_mL Mtm_ctrl _0_75h_ 24068,
- 3_18 0_2μg_mL Mtm_ctrl _1_5h_ 24069,
- 3_58 0_2μg_mL Mtm_ctrl _2h_ 24070,
- 3_62 0_2μg_mL Mtm_ctrl _4h_ 24074,
- 4_27 0_2μg_mL Mtm_ctrl _2h_ 24071,
- 4_29 0_2μg_mL Mtm_ctrl _2h_ 24072,
- 4_71 0_2μg_mL Mtm_ctrl _4h_ 24075,
- 4-61 0_2μg_mL Mtm_ctrl _8h__ H37Rv 25444,
- 4-62 0_2μg_mL Mtm_ctrl _8h__ H37Rv 25451,
- 4-67 0_2μg_mL Mtm_ctrl _12h__ H37Rv 25497,
- 5_23 0_2μg_mL Mtm_ctrl _2h_ 24073,
- 5_25 0_2μg_mL Mtm_ctrl _4h_ 24077,
- 5_58 0_2μg_mL Mtm_ctrl _6h_ 24079,
- 5_67 0_2μg_mL Mtm_ctrl _4h_ 24078.

Datenmatrix mit Expressionswerten für die Gene *recA, lexA, ruvC* und *linB*

Tabelle A.1: Logarithmierte, normalisierte und gemittelte Daten für *recA, lexA, ruvC* und *linB*.

	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7
recA	-0,48086	-0,017203	-0,11173	1,749	0,5341	3,493	3,1227	3,3354
lexA	-0,404	-0,22901	-0,32045	0,83787	0,16567	1,943	2,297	1,7866
ruvC	-0,27346	-0,8427	-0,27299	1,027	0,25326	0,75343	0,88304	1,2605
linB	-0,29476	-0,058448	-0,055374	0,5439	1,0479	3,5294	2,1926	1,7931

Weitere Genlisten und Begriffsverbände

Genliste S_2

Die Genliste S_2 enthält die folgenden 196 Gene:
$S_2 = \{$*recA, lexA, ruvC, linB, Rv2719c, dnaE2, dinX, infB, lprJ, Rv1691, Rv2840c, Rv3055, uvrA, uvrB, uvrC, uvrD, apt, cobA, ctaB, ctpH, dinP, dnaB, dnaE1, acs, alaS, argG, argS, asnB, aspS, birA, cdsA, cysD, cysN, cysS2, cysS, dnaQ, dnaZX, fadD11, fadD14, fadD15, fadD21, fadD23, fadD24, fadD25, fadD26, fadD30, fadD32, fadD35, fadD3, fadD4, galT, galTp, galU, glgC, glmU, glnD, glnE, gltS, glyS, guaA, hisS, ileS, kdtB, leuS, ligB, lysS, lysX, mbtA, menE, metS, miaA, panC, pcnA, pheS, pheT, ppdK, proS, ribF, rmlA2, rmlA, rpoA, rpoB, rpoC, Rv1390, Rv2421c, Rv2438c, Rv3644c, serS, thrS, trpS, tyrS, valS, dnaN, dut, embA, embB, embC, folP2, folP, grcC1, grcC2, helY, hisG, hisH, hisI, hpt, idsA, idsB, lhr, lipB, menA, metK, mutT2, nadC, phyA, pknA, pknB, pknD, pknE, pknF, pknH, pknI, pknJ, pknK, pknL, polA, ppa, purF, pyrR, recG, ribA2, ribA, Rv1086, Rv1318c, Rv1319c, Rv1320c, Rv2191, Rv2361c, Rv2413c, Rv3201c, thiE, trpD, umpA, upp, uvrD2, ilvB, alr, ansA, cdd, fabG3, glpK, ilvA, lipT, lpdA, mmpL6, mtrA, phoP, pncA, regX3, rmlB2, Rv0110, Rv0415, Rv0458, Rv0476, Rv0903c, Rv0981, Rv1033c, Rv1234, Rv1343c, Rv1903, Rv2000, Rv2041c, Rv2044c, Rv2250c, Rv2358, Rv3329, Rv3330, Rv3764c, Rv3765c, sahH, sugI, tcrA, tmk, ugpA, pstC, rpmH, Rv0938, Rv1678, Rv1781c, Rv3675, Rv3676, Rv3697c, Rv3843c, Rv3920c, uspA, glmS.*$\}$

Die Genlisten S' für $k = 3$ und $k = 4$

Die Genliste S'_3 enthält die folgenden 59 Gene:
$S'_3 = \{$*recA, lexA, ruvC, linB, Rv0059, Rv0807, Rv0822c, Rv0881, Rv1075c, Rv1222, Rv1781c, Rv1786, Rv1814, Rv1945, Rv2165c, Rv2719c, Rv3059, accD2, ald, alkA, appC, carA, dapB, dfrA, dgt, dnaB, dnaE2, echA8, fadD21, fadD23, fadD25, fadE21, fbpC2, fmt, gcvB, gltD, hsdM, idsA, ilvD, lprJ, ltp1, metA, moaA, mtrA, narK1, nhoA, nuoD, nuoK, phyA, pks6, pssA, purD, sdhA, sigI, thiL, thyA, truB, uvrA, uvrD.*$\}$

Die Genliste S'_4 enthält die folgenden 69 Gene:
$S'_4 = \{$*recA, lexA, ruvC, linB, Rv0059, Rv0807, Rv0822c, Rv0881, Rv1075c, Rv1136, Rv1222, Rv1781c, Rv1786, Rv1814, Rv1945, Rv2111c, Rv2165c, Rv2719c, Rv2795c, Rv3059, accD2, ald, alkA, appC, carA, clpX, dapB, dfrA, dgt, dnaB, dnaE2, echA13, echA18, echA8, entC,*

fadD21, fadD23, fadD25, fadE21, fbpC2, fmt, furA, gcvB, gltD, hsdM, idsA, ilvD, lprJ, ltp1, metA, moaA, mtrA, narK1, nhoA, nuoD, nuoK, phyA, pks4, pks6, pssA, purD, rpe, sdhA, sigI, thiL, thyA, truB, uvrA, uvrD.}

Begriffsverbände für $k = 3$ und $k = 4$

Die Begriffsverbände für $k = 3$ und $k = 4$ sind in Abbildungen A.1 und A.2 dargestellt.

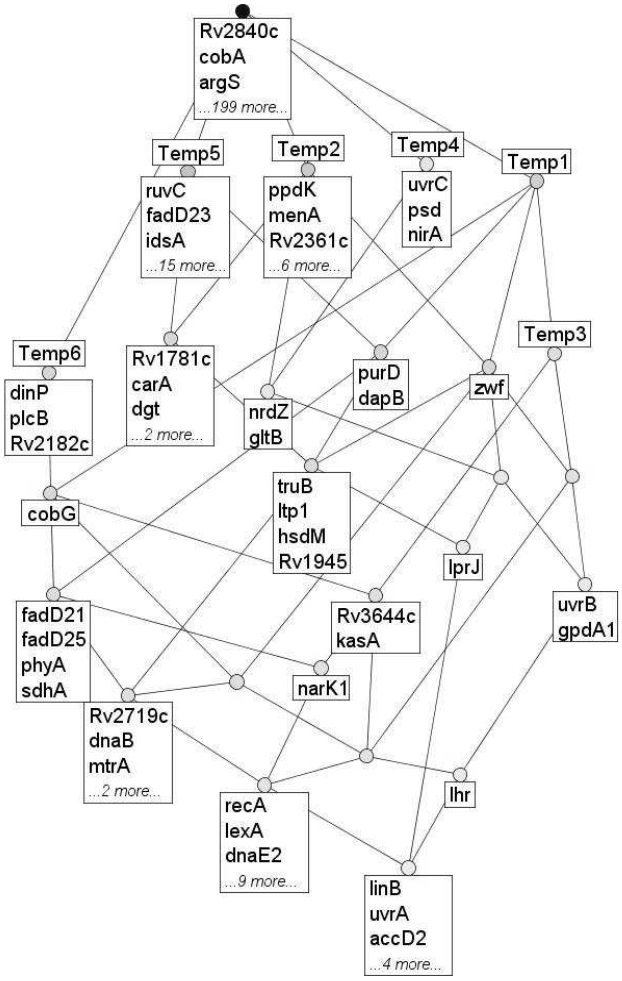


Abbildung A.1: Der Begriffsverband $\underline{\mathscr{B}}(\mathbb{K}_3)$, dargestellt mit ToscanaJ (Becker u. a., 2002).

A.2 Massenerhaltende Mengen bei der Visualisierung des *C. glutamicum*

Für die algebraische Chemie $< M_C, R_{C1} >$, die in Kapitel 7.1 definiert wurde, wird im Folgenden bewiesen, dass sämtliche Mengen $S \subseteq M_C$ massenerhaltend sind.

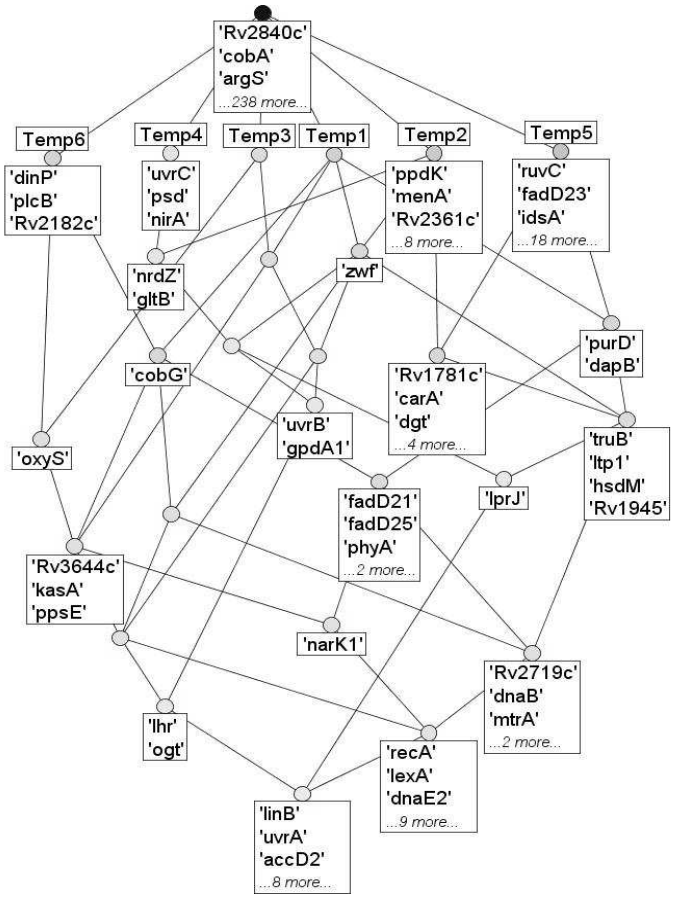


Abbildung A.2: Der Begriffsverband $\underline{\mathscr{B}}(\mathbb{K}_4)$, dargestellt mit ToscanaJ (Becker u. a., 2002).

Als Vektor v kann jeweils folgender Vektor gewählt werden:

- $v = (0,0,0,0,1,1,0,0,0)^t$ für die Menge $S = \emptyset$,
- $v = (0,0,0,0,1,1,0,0,0)^t$ für die Menge $S = \{\text{GlnD}\}$,
- $v = (0,0,0,0,2,1,1,0,0)^t$ für die Menge $S = \{\text{GlnK}\}$,
- $v = (0,2,0,0,1,1,0,0,1)^t$ für die Menge $S = \{\text{GlnK}\sim\text{AMP}\}$,
- $v = (0,0,0,0,1,2,0,1,0)^t$ für die Menge $S = \{\text{AmtB}\}$,
- $v = (1,0,0,0,3,1,1,0,0)^t$ für die Menge $S = \{\text{GlnD, GlnK}\}$,
- $v = (0,2,0,0,1,1,0,0,1)^t$ für die Menge $S = \{\text{GlnD, GlnK}\sim\text{AMP}\}$,
- $v = (0,0,0,0,1,2,0,1,0)^t$ für die Menge $S = \{\text{GlnD, AmtB}\}$,
- $v = (0,2,0,0,2,1,1,0,1)^t$ für die Menge $S = \{\text{GlnK, GlnK}\sim\text{AMP}\}$,
- $v = (0,0,0,1,3,2,1,1,0)^t$ für die Menge $S = \{\text{GlnK, AmtB}\}$,
- $v = (0,2,1,0,1,1,0,1,1)^t$ für die Menge $S = \{\text{GlnK}\sim\text{AMP, AmtB}\}$,
- $v = (1,1,0,0,3,1,1,0,1)^t$ für die Menge $S = \{\text{GlnD, GlnK, GlnK}\sim\text{AMP}\}$,
- $v = (1,0,0,1,4,2,1,1,0)^t$ für die Menge $S = \{\text{GlnD, GlnK, AmtB}\}$,
- $v = (0,2,1,0,1,1,0,1,1)^t$ für die Menge $S = \{\text{GlnD, GlnK}\sim\text{AMP, AmtB}\}$,
- $v = (0,2,1,1,3,1,1,1,1)^t$ für die Menge $S = \{\text{GlnK, GlnK}\sim\text{AMP, AmtB}\}$,
- $v = (1,1,1,1,4,1,1,1,1)^t$ für die Menge $S = \{\text{GlnD, GlnK, GlnK}\sim\text{AMP, AmtB}\}$.

Für diese Vektoren v mit der zugehörigen Menge S gilt jeweils

1. $\forall (A,B) \in R$ mit $A \in \mathscr{P}_M(S) \Rightarrow v_{(A,B)} > 0$,
2. $\forall (A,B) \in R$ mit $A \notin \mathscr{P}_M(S) \Rightarrow v_{(A,B)} = 0$,
3. $\mathscr{M} v \geq 0$.

Somit sind alle Teilmengen von M_C massenerhaltend.

Literaturverzeichnis

[Akhmet u. a. 2005] AKHMET, M.U. ; GEBERT, J. ; ÖKTEM, H. ; PICKL, S.W. ; WEBER, G.-W.: An improved algorithm for analytical modeling and anticipation of gene expression patterns. In: *J. Comp. Tech.* 10(4) (2005), S. 3–20

[Akutsu u. a. 1999] AKUTSU, T. ; MIYANO, S. ; KUHARA, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: *Proc. Pac. Symp. Biocomp.* (1999), S. 17–28

[Bansal u. a. 2006] BANSAL, M. ; GATTA, G.D. ; BERNADO, D. di: Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. In: *Bioinformatics* 22(7) (2006), S. 815–822

[Becker u. a. 2002] BECKER, P. ; HERETH, J. ; STUMME, G.: ToscanaJ: An open source tool for qualitative data analysis. In: *Advances in Formal Concept Analysis for Knowledge Discovery in Databases* (2002), S. 1–2

[Beckers u. a. 2005] BECKERS, G. ; STRÖSSER, J. ; HILDEBRANDT, U. ; KALINOWSKI, J. ; FARWICK, M. ; KRÄMER, R. ; BURKOVSKI, A.: Regulation of AmtR-controlled gene expression in *Corynebacterium glutamicum*: mechanism and characterization of the AmtR regulon. In: *Mol. Microbiol.* 58(2) (2005), S. 580–595

[Behnke 2003] BEHNKE, H.: *Einführung in die Theorie der Differentialgleichungen*. Osnabrücker Schriften zur Mathematik, Reihe V, Heft 38, 2003

[Birkhoff 1938] BIRKHOFF, G.: Lattices and their applications. In: *Bull. Amer. Math. Soc.* 44 (1938), S. 793–800

[Bormann u. a. 2001] BORMANN, G. ; BROSENS, F. ; SCHUTTER, E. D.: Diffusion. In: *Comp. Mod. Gen. Biochem. Netw.* (2001), S. 189–224

[Boshoff u. a. 2004] BOSHOFF, H. ; MYERS, T.G. ; COPP, B.R. ; MCNEILL, M.R. ; WILSON, M.A. ; BARRY, C.E.: The transcriptional response of *Mycobacterium tuberculosis* to inhibitors of metabolism: novel insights into drug mechanisms of action. In: *Biol. Chem.* 279(38) (2004), S. 40174–40184

[Breiman 1993] BREIMAN, L.: Hinging hyperplanes for regression, classification, and function approximation. In: *IEEE Trans. Inf. Theory* 39(3) (1993), S. 999–1013

[Bronstein u. Semendjajew 1991] BRONSTEIN, I.N. ; SEMENDJAJEW, K.A.: *Taschenbuch der Mathematik*. 25. Auflage. B.G. Teubner Verlagsgesellschaft, Stuttgart, Leipzig und Verlag Nauka, Moskau, 1991

[Burkovski 2003a] BURKOVSKI, A.: Ammonium assimilation and nitrogen control in *Corynebacterium glutamicum* and its relatives: an example for new regulatory mechanisms in actinomycetes. In: *FEMS Microbiol. Rev.* 27 (2003), S. 617–628

[Burkovski 2003b] BURKOVSKI, A.: I do it my way: Regulation of ammonium uptake and ammonium assimilation in *Corynebacterium glutamicum*. In: *Arch. Microbiol.* 179(2) (2003), S. 83–88

[Burkovski 2005] BURKOVSKI, A.: Nitrogen metabolism and its regulation. In: BOTT, M. (Hrsg.) ; EGGELING, L. (Hrsg.): *Handbook of Corynebacterium glutamicum*. Boca Raton : CRC Press LLC, 2005, S. 333–349

[Butte 2002] BUTTE, A.: The use and analysis of microarray data. In: *Nature Rev.* 1 (2002), S. 951–959

[Cabusora u. a. 2005] CABUSORA, L. ; SUTTON, E. ; FULMER, A. ; FORST, C.V.: Differential network expression during drug and stress response. In: *Bioinformatics* 21 (2005), S. 2898–2905

[Casey u. a. 2006] CASEY, R. ; JONG, H. de ; GOUZÉ, J.-L.: Piecewise-linear models of genetic regulatory networks: equilibria and their stability. In: *J. Math. Biol.* 52 (2006), S. 27–56

[Chen u. a. 1999] CHEN, T. ; HE, H.L. ; CHURCH, G.M.: Modeling gene expression with differential equations. In: *Proc. Pac. Symp. on Biocomp.* (1999), S. 29–40

[Cui u. Churchill 2003] CUI, X. ; CHURCHILL, G.A.: Statistical tests for differential expression in cDNA microarray experiments. In: *Genome Biol.* 4(4) (2003), S. 210

[Darnell u. a. 1994] DARNELL, J. ; LODISH, H. ; BALTIMORE, D.: *Molekulare Zellbiologie*. 2. Auflage. Walter de Gruyter, Berlin, New York, 1994

[d'Haeseleer u. a. 1999] D'HAESELEER, P. ; WEN, X. ; FUHRMAN, S. ; SOMOGYI, R.: Linear modeling of mRNA expression levels during CNS development and injury. In: *Proc. Pac. Symp. Biocomp.* (1999), S. 41–52

[Dittrich u. di Fenizio 2005] DITTRICH, P. ; FENIZIO, P. S.: Chemical organization theory: Towards a theory of constructive dynamical systems. In: *arXiv:q-bio.MN/0501016* (2005)

[Duda u. a. 2001] DUDA, R.O. ; HART, P.E. ; STORK, D.G.: *Pattern classification*. John Wiley & Sons, 2001

[Dullaghan u. a. 2002] DULLAGHAN, E.M. ; BROOKS, P.C. ; DAVIS, E.O.: The role of multiple SOS boxes upstream of the *Mycobacterium tuberculosis* lexA gene - identification of a novel DNA-damage-inducible gene. In: *Microbiol.* 148 (2002), S. 3609–3615

[Faigle 1983] FAIGLE, U.: Linear growth: a unifying approach to linear systems of difference and differential equations. In: *Int. J. Math. Educ. Sci. Technol.* 14 (1983), Nr. 2, S. 137–143

[Faigle u. a. 2002] FAIGLE, U. ; KERN, W. ; STILL, G.: *Algorithmic principles of mathematical programming*. Kluwer Academic Publishers, 2002

[Fakler 2006] FAKLER, P.: *Ein verbandstheoretisches Modell zur Prognose von Kreditausfallwahrscheinlichkeiten*, Universität zu Köln, Diss., 2006

[di Fenizio u. Dittrich 2002] FENIZIO, P. S. ; DITTRICH, P.: Artificial chemistry's global dynamics. Movement in the lattice of organisation. In: *J. Three Dim. Im.* 16(4) (2002), S. 160–163

[Fontana u. Buss 1994] FONTANA, W. ; BUSS, L.W.: The arrival of the fittest: Toward a theory of biological organization. In: *Bull. Math. Biol.* 56 (1994), S. 1–64

[Forster 1992] FORSTER, O.: *Analysis 1*. 4. Auflage. Vieweg Studium, 1992

[Friedman u. a. 2000] FRIEDMAN, N. ; LINIAL, M. ; NACHMAN, I. ; PE´ER, D.: Using bayesian networks to analyze expression data. In: *J. Comp. Biol. 7* (2000), S. 601–620

[Ganter u. Wille 1996] GANTER, B. ; WILLE, R.: *Formale Begriffsanalyse, Mathematische Grundlagen*. Springer-Verlag, Berlin, Heidelberg, 1996

[Gebert u. a. 2003a] GEBERT, J. ; LÄTSCH, M. ; PICKL, S. W. ; RADDE, N. ; WEBER, G.-W. ; WÜNSCHIERS, R.: Genetic networks and anticipation of gene expression patterns. In: *AIP Conf. Proc.* 718 (2003), S. 474–485

[Gebert u. a. 2006] GEBERT, J. ; LÄTSCH, M. ; PICKL, S. W. ; WEBER, G.-W. ; WÜNSCHIERS, R.: An algorithm to analyze stability of gene-expression pattern. In: *Discr. Appl. Math.* 154(7) (2006), S. 1140–1156

[Gebert u. a. 2004a] GEBERT, J. ; LÄTSCH, M. ; QUEK, E.M.P. ; WEBER, G.-W.: Analyzing and optimizing genetic network structure via path-finding. In: *J. Comp. Techn.* 9(3) (2004), S. 3–12

[Gebert u. a. 2004b] GEBERT, J. ; ÖKTEM, H. ; PICKL, S.W. ; RADDE, N. ; WEBER, G.-W. ; YILMAZ, F.B.: Inference of gene expression patterns by using a hybrid system formulation - an algorithmic approach to local state transition matrices. In: *Anticipat. Predict. Mod. Syst. Sci. I* (2004), S. 63–66

[Gebert u. a. 2003b] GEBERT, J. ; PICKL, S. ; SHOKINA, N. ; WEBER, G.-W. ; WÜNSCHIERS, R.: Mathematical modeling and discrete approximation in evaluation and forecasting of expression-data. In: *Electr. Notes in Discr. Math.* 13 (2003), S. 52–56

[Gebert u. a. 2002] GEBERT, J. ; PICKL, S. W. ; SHOKINA, N. ; WEBER, G.-W. ; WÜNSCHIERS, R.: Algorithmic analysis of gene expression data with polyhedral structures. In: KRÖPLIN, B. (Hrsg.) ; RUDOLPH, S. (Hrsg.) ; HÄCKER, J. (Hrsg.): *Similarity Methods - 5th International Workshop*. 2002, S. 79–87

[Gebert u. Radde 2006] GEBERT, J. ; RADDE, N.: A new approach for modeling procaryotic biochemical networks with differential equations. In: *AIP Conf. Proc.* 839 (2006), S. 526–533

[Gebert u. a. 2007a] GEBERT, J. ; RADDE, N. ; FAIGLE, U. ; STROESSER, J. ; BURKOVSKI, A.: *Modeling and simulation of nitrogen regulation in Corynebacterium glutamicum*. 2007. – wird erscheinen in Discr. Appl. Math., Special Issue on Networks in Computational Biology

[Gebert u. a. 2007b] GEBERT, J. ; RADDE, N. ; WEBER, G.-W.: *Modeling gene regulatory networks with piecewise linear differential equations*. 2007. – wird erscheinen in Special Issue of Challenges of Continuous Optimization in Theory and Applications in Europ. J. Operat. Res. 18(3)

[Gillespie 1977] GILLESPIE, D.T.: Exact stochastic simulation of coupled chemical reactions. In: *J. Phys. Chem.* 81(25) (1977), S. 2340–2361

[Glass u. Kauffman 1973] GLASS, L. ; KAUFFMAN, S.A.: The logical analysis of continuous, non-linear biochemical control networks. In: *J. Theor. Biol. 39* (1973), S. 103–129

[Hashimoto u. a. 2004] HASHIMOTO, R.F. ; KIM, S. ; SHMULEVICH, I. ; ZHANG, W. ; BITTNER, M.L. ; DOUGHERTY, E.R.: Growing genetic regulatory networks from seed genes. In: *Bioinformatics* 20 (2004), S. 1241–1247

[Hirsch u. Smale 1974] HIRSCH, M.W. ; SMALE, S.: *Differential equations, dynamical systems, and linear algebra*. Academic Press, 1974

[Hoon u. a. 2003] HOON, M. D. ; IMOTO, S. ; KOBAYASHI, K. ; OGASAWARA, N. ; MIYANO, S.: Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. In: *Proc. Pac. Symp. Biocomp.* (2003), S. 17–28

[Hoon u. a. 2002] HOON, M. D. ; IMOTO, S. ; MIYANO, S.: Inferring gene regulatory networks from time-ordered gene expression data using differential equations. In: *Lect. Notes in Comp. Sci.* 2534 (2002), S. 267–274

[Ihmels u. a. 2002] IHMELS, J. ; FRIEDLANDER, G. ; BERGMANN, S. ; SARIG, O. ; ZIV, Y. ; BARKAI, N.: Revealing modular organization in the yeast transcriptional network. In: *Nature Genet.* 31 (2002), S. 370–377

[Jacob u. Monod 1961] JACOB, F. ; MONOD, J.: Genetic regulatory mechanisms in the synthesis of proteins. In: *J. Mol. Biol.* 3 (1961), S. 318–356

[Jakoby u. a. 2000] JAKOBY, M. ; NOLDEN, L. ; MEIER-WAGNER, J. ; KRÄMER, R. ; BURKOVSKI, A.: AmtR, a global repressor in the nitrogen regulation system of *Corynebacterium glutamicum*. In: *Mol. Microbiol.* 37(4) (2000), S. 964–977

[Janning u. Knust 2004] JANNING, W. ; KNUST, E.: *Genetik*. Thieme, 2004

[Jimenez u. Marzal 1999] JIMENEZ, V. M. ; MARZAL, A.: An algorithm for efficient computation of k shortest paths. In: *Proc. 3rd Intern. Workshop on Alg. Eng., Lecture Notes in Computer Science* (1999)

[de Jong 2002] JONG, H. de: Modeling and simulation of genetic regulatory systems: A literature review. In: *J. Comp. Biol. 9* (2002), S. 69–105

[de Jong u. a. 2004] JONG, H. de ; GOUZÉ, J.-L. ; HERNANDEZ, C. ; PAGE, M. ; SARI, T. ; GEISELMANN, J.: Qualitative simulation of genetic regulatory networks using piecewise-linear models. In: *Bull. Math. Biol.* (2004)

[Kaderali u. a. 2006] KADERALI, L. ; ZANDER, T. ; FAIGLE, U. ; WOLF, J. ; SCHULTZE, J.L. ; SCHRADER, R.: CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. In: *Bioinformatics* 22(12) (2006), S. 1495–1502

[Kaminski u. Friedman 2002] KAMINSKI, N. ; FRIEDMAN, N.: Practical approaches to analyzing results of microarray experiments. In: *Am. J. Respirat. Cell and Mol. Biol.* 27 (2002), S. 125–132

[Kell u. Knowles 2006] KELL, D.B. ; KNOWLES, J.D.: The role of modeling in systems biology. In: SZALLASI, Z. (Hrsg.) ; STELLING, J. (Hrsg.) ; PERIWAL, V. (Hrsg.): *System modeling in cellular biology, from concepts to nuts and bolts*. MIT Press, 2006, S. 1–18

[Kendall 1938] KENDALL, M.G.: A new measure of rank correlation. In: *Biometrika* 30 (1938), S. 81–93

[Kloster u. a. 2005] KLOSTER, M. ; TANG, C. ; WINGREEN, N.S.: Finding regulatory modules through large-scale gene-expression data analysis. In: *Bioinformatics* 21(7) (2005), S. 1172–1179

[Knippers 2001] KNIPPERS, R.: *Molekulare Genetik*. 8. Auflage. Georg Thieme Verlag, 2001

[Lawson u. Hanson 1974] LAWSON, C. ; HANSON, R.: *Solving least squares problems*. Prentice Hall, 1974

[Lerman 1980] LERMAN, P.M.: Fitting segmented regression models by grid search. In: *Applied Statistics* 29 (1980), S. 77–84

[Liang u. a. 1998] LIANG, S. ; FUHRMAN, S. ; SOMOGYI, R.: Reveal: a general reverse engineering algorithm for inference of genetic network architectures. In: *Pac. Symp. Biocomp.* (1998), S. 18–29

[Little u. Mount 1982] LITTLE, J.W. ; MOUNT, D.W.: The SOS regulatory system of *Escherichia coli*. In: *Cell* 29 (1982), S. 11–22

[Lodish u. a. 2001] LODISH, H. ; BERK, A. ; ZIPURSKY, S.L. ; MATSUDAIRA, P. ; BALTIMORE, D. ; DARNELL, J.E.: *Molekulare Zellbiologie*. 4. Auflage. Spektrum, Akad. Verl., 2001

[Markowetz 2006] MARKOWETZ, F.: *Probabilistische Modelle für RNA-Interferenz-Daten*, Freie Universität Berlin, Diss., 2006

[Mawuenyega u. a. 2005] MAWUENYEGA, K. G. ; FORST, C.V. ; DOBOS, K. M. ; BELISLE, J. T. ; CHEN, J. ; BRADBURY, E. M. ; BRADBURY, A. R. ; CHEN, X.: *Mycobacterium tuberculosis* functional network analysis by global subcellular protein profiling. In: *Mol. Biol. of the Cell* 16 (2005), S. 396–404

[Meier-Wagner u. a. 2001] MEIER-WAGNER, J. ; NOLDEN, L. ; JAKOBY, M. ; SIEWE, R. ; KRÄMER, R. ; BURKOVSKI, A.: Multiplicity of ammonium uptake systems in *Corynebacterium glutamicum*: Role of Amt and AmtB. In: *Microbiology* 147(1) (2001), S. 135–143

[Murphy u. Mian 1999] MURPHY, K. ; MIAN, S.: Modeling gene expression data using dynamic bayesian networks. 1999. – Technical report

[Natterer 1998] NATTERER, F.: *Vorlesungsskript Gewöhnliche Differentialgleichungen*. 1998

[Nolden u. a. 2001] NOLDEN, L. ; NGOUOTO-NKILI, C.-E. ; BENDT, A. K. ; KRÄMER, R. ; BURKOVSKI, A.: Sensing nitrogen limitation in *Corynebacterium glutamicum*: the role of *glnK* and *glnD*. In: *Mol. Microbiol.* 42(5) (2001), S. 1281–1295

[Öktem 2005] ÖKTEM, H.: A survey on piecewise-linear models of regulatory dynamical systems. In: *Nonlinear Analysis* 63 (2005), S. 336–349

[Pearl 1988] PEARL, J.: *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988

[Perrin u. a. 2003] PERRIN, B. ; RALAIVOLA, L. ; MAZURIE, A. ; BOTTANI, S. ; MALLET, J. ; BUC, F. d´Alché: Gene networks inference using dynamic bayesian networks. In: *Bioinformatics 19* (2003), S. 138–148

[Ptashne 1986] PTASHNE, M.: *A genetic switch. Gene control and phage* λ. Blackwell, 1986

[Pucar u. Sjöberg 1998] PUCAR, P. ; SJÖBERG, J.: On the hinge finding algorithm for hinging hyperplanes. In: *IEEE Trans. Inf. Theory* 44(3) (1998), S. 1310–8

[Radde u. a. 2006] RADDE, N. ; GEBERT, J. ; FORST, C. V.: Systematic component selection for gene-network refinement. In: *Bioinformatics* 22(21) (2006), S. 2674–2680

[Radde u. Kaderali 2007] RADDE, N. ; KADERALI, L.: Bayesian inference of gene regulatory networks using gene expression time series data. In: HOCHREITER, S. (Hrsg.) ; WAGNER, R. (Hrsg.): *BIRD 2007*, Springer-Verlag Berlin Heidelberg, 2007, S. 1–15

[Rand u. a. 2003] RAND, L. ; HINDS, J. ; SPRINGER, B. ; SANDERS, P. ; BUXTON, R.S. ; DAVIS, E.O.: The majority of inducible DNA repair genes in *Mycobacterium tuberculosis* are induced independently of RecA. In: *Mol. Microbiol.* 50(3) (2003), S. 1031–42

[Sakamoto u. Iba 2001] SAKAMOTO, E. ; IBA, H.: Inferring a system of differential equations for a gene regulatory network by using genetic programming. In: *Proc. Congr. on Evolut. Comp.* (2001), S. 720–726

[Sartor u. a. 2003] SARTOR, M. ; MEDVEDOVIC, M. ; ARONOW, B.: Microarray data normalization: the art and science of overcoming technical variance to maximize the detection of biologic variance. In: BLALOCK, Eric (Hrsg.): *A beginner's guide to microarrays*. Kluwer Academic Publishers, 2003, S. 151–178

[Schena 1999] SCHENA, M.: *DNA microarrays*. Oxford University, 1999

[Schena 2003] SCHENA, M.: *Microarray analysis*. John Wiley & Sons, 2003

[Schliep u. a. 2003] SCHLIEP, A. ; COSTA, I.G. ; STEINHOFF, C. ; SCHÖNHUTH, A.: Analyzing gene expression time-courses. In: *IEEE/ACM Trans. Comp. Biol. and Bioinf. (TCBB)* 2(3) (2003), S. 179–193

[Seber u. Wild 1989] SEBER, G.A.F. ; WILD, C.J.: *Nonlinear regression*. John Wiley & Sons, 1989

[Shamir u. Sharan 2002] SHAMIR, R. ; SHARAN, R.: Algorithmic approaches to clustering gene expression data. In: *Current topics in computational molecular biology*. MIT Press, 2002, S. 269–300

[Shmulevich u. a. 2002a] SHMULEVICH, I. ; DOUGHERTY, E.R. ; KIM, S. ; ZHANG, W.: Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks. In: *Bioinformatics 18* (2002), S. 261–274

[Shmulevich u. a. 2002b] SHMULEVICH, I. ; DOUGHERTY, E.R. ; ZHANG, W.: From Boolean to probabilistic Boolean networks as models of genetic regulatory networks. In: *Proc. of the IEEE 90(11)* (2002), S. 1778–1792

[Siewe u. a. 1996] SIEWE, R.M. ; WEIL, B. ; BURKOVSKI, A. ; EIKMANNS, B.J. ; EIKMANNS, M. ; KRÄMER, R.: Functional and genetic characterization of the (methyl)ammonium uptake carrier of *Corynebacterium glutamicum*. In: *J. Biol. Chem.* 271(10) (1996), S. 5398–5403

[Slonim 2002] SLONIM, D.K.: From patterns to pathways: gene expression data analysis comes of age. In: *Nature Genet.* (2002), S. 32

[Smolen u. a. 2000] SMOLEN, P. ; BAXTER, D.A. ; BYRNE, J. H.: Modeling transcriptional control in gene networks - methods, recent results, and future directions. In: *Bull. Math. Biol. 62* (2000), S. 247–292

[Somogyi u. Sniegoski 1996] SOMOGYI, R. ; SNIEGOSKI, C.A.: Modeling the complexity of genetic networks: Understanding multigenic and pleiotropic regulation. In: *Complexity* 1(6) (1996), S. 45–63

[Stroesser u. a. 2004] STROESSER, J. ; LÜDKE, A. ; SCHAFFER, S. ; KRÄMER, R. ; BURKOVSKI, A.: Regulation of GlnK activity: modification, membrane sequestration and proteolysis as regulatory principles in the network of nitrogen control in *Corynebacterium glutamicum*. In: *Mol. Microbiol.* 54(1) (2004), S. 132–147

[Szallasi u. a. 2006] SZALLASI, Z. ; STELLING, J. ; PERIWAL, V.: *System modeling in cellular biology, from concepts to nuts and bolts*. MIT Press, 2006

[Thierolf 2003] THIEROLF, F.: *Mathematische Modelle und Methoden zur Genexpressionsanalyse in der Bioinformatik*. 2003. – Diplomarbeit, Technische Universität Darmstadt

[Thomas 1998] THOMAS, R.: Laws for the dynamic of regulatory networks. In: *Int. J. Dev. Biol. 42* (1998), S. 479–485

[Thomas u. Kaufman 2001] THOMAS, R. ; KAUFMAN, M.: Multistationarity, the basis of cell differentiation and memory. 1. Structural conditions of multistationarity and other nontrivial behavior. In: *Chaos* 11(1) (2001), S. 170–179

[Thulasiraman u. Swamy 1992] THULASIRAMAN, K. ; SWAMY, M.N.S.: *Graphs: theory and algorithms*. 1992

[Walker 1996] WALKER, G.C.: The SOS response of *Escherichia coli*. In: NEIDHARDT, F.C. (Hrsg.): *Escherichia coli and Salmonella*. ASM Press, 1996, S. 1400–1416

[Walter 1993] WALTER, W.: *Gewöhnliche Differentialgleichungen*. 5. Auflage. Springer-Verlag, 1993

[Watson u. Crick 1953] WATSON, J.D. ; CRICK, F.H.C.: Molecular structure of nucleic acids. A structure for desoxyribose nucleic acid. In: *Nature* 171 (1953), S. 737–738

[Wiebringhaus u. a. 2003] WIEBRINGHAUS, T. ; FAIGLE, U. ; SCHOMBURG, D. ; GEBERT, J. ; IGEL, C. ; WEBER, G.-W.: Protein fold class prediction using neural networks reconsidered. In: *Curr. in Comp. Mol. Biol., Sev. Ann. Int. Conf. on Res. in Comp. Mol. Biol. (RECOMB 2003)* (2003), S. 225–226

[Wiebringhaus u. a. 2004] WIEBRINGHAUS, T. ; IGEL, C. ; GEBERT, J.: Protein fold class prediction using neural networks with tailored early-stopping. In: *Int. Joint Conf. on Neur. Netw. (IJCNN 2004)* (2004), S. 1693–1697

[Wille 1982] WILLE, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: RIVAL, I. (Hrsg.): *Ordered sets*. Reidel, Dordrecht-Boston, 1982, S. 445–470

[Wünschiers 2004] WÜNSCHIERS, R.: Die Mikroarray-Maschine der febit AG. In: *CLB Chemie in Labor und Biotechnik* 55 (2004), S. 8–13

[Wünschiers u. a. 2001] WÜNSCHIERS, R. ; ZINN, T. ; BORZNER, S.: Herstellung und Verwendung von DNA-Mikroarrays. In: *CLB Chemie in Labor und Biotechnik* 52 (2001), S. 260–266

[Wolff 2000] WOLFF, K.E.: Towards a conceptual system theory. In: *Proc. World Multiconf. on Syst., Cybern. and Inf.* 2 (2000), S. 124–132

[Wolff 2001] WOLFF, K.E.: Temporal concept analysis. In: *Int. Workshop on Conc. Lattices-Based Th., Meth. and Tools for Knowl. Disc. in Datab.* (2001), S. 91–107

[Wolff 2002] WOLFF, K.E.: Transitions in conceptual time systems. In: *Int. J. Comp. Anticip. Syst.* 11 (2002), S. 398–412

[World Health Organization] WORLD HEALTH ORGANIZATION, 2004: *Anti-tuberculosis drug resistance in the world, Third Global Report*

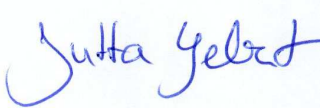
[Yagil 1975] YAGIL, G.: Quantitative aspects of protein induction. In: *Curr. Top. in Cell. Reg.* 9 (1975), S. 183–236

[Yagil u. Yagil 1971] YAGIL, G. ; YAGIL, E.: On the relation between effector concentration and the rate of induced enzyme synthesis. In: *Biophys. J. 11* (1971), S. 11–27

Stichwortverzeichnis

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat, dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. U. Faigle betreut worden.



Teilpublikationen:

2007 J. Gebert, N. Radde, U. Faigle, J. Strösser, A. Burkovski, Modelling and simulation of nitrogen regulation in *Corynebacterium glutamicum*, wird erscheinen in Discrete Applied Mathematics, Special Issue on Networks in Computational Biology.

2007 J. Gebert, S. Motameny, U. Faigle, C. V. Forst, R. Schrader, Identifying genes of gene regulatory networks using formal concept analysis, in Arbeit.

2007 N. Radde, J. Gebert, U. Faigle, R. Schrader, K. Schnetz, Modeling feedback loops in the regulation of the *bgl* operon in *Escherichia coli*, in Arbeit.

2007 J. Gebert, N. Radde, G.-W. Weber, Modeling gene regulatory networks with piecewiese linear differential equations, wird erscheinen in Special Issue of Challenges of Continuous Optimization in Theory and Applications in the European Journal of Operational Research 18(3).

2006 J. Gebert, N. Radde, A new approach for modeling procaryotic biochemical networks with differential equations, AIP Conference Proceedings 839, S. 526-533.

2006 N. Radde, J. Gebert, C. V. Forst, Systematic component selection for gene-network refinement, Bioinformatics 22(21), S. 2674-2680.

2006 J. Gebert, M. Lätsch, S.W. Pickl, G.-W. Weber, R. Wünschiers, An algorithm to analyze stability of gene-expression patterns, Discrete Appl. Math. 154(7), in: special issue Discrete Mathematics and Data Mining II, Anthony, M., Boros, E., Hammer, P.L.,and Kogan, A. (guest eds.), S. 1140-1156.

2005 M.U. Akhmet, J. Gebert, H. Öktem, S.W. Pickl, G.-W. Weber, An improved algorithm

for analytical modeling and anticipation of gene expression patterns, Journal of Computational Technologies 10(4), S. 3-20.
2004 J. Gebert, M. Lätsch, E.M.P. Quek, G.-W. Weber, Analyzing and optimizing genetic network structure via path-finding, Journal of Computational Technologies 9(3), S. 3-12.
2004 J. Gebert, H. Öktem, S.W. Pickl, N. Radde, G.-W. Weber, F.B. Yilmaz, Inference of gene expression patterns by using a hybrid system formulation - an algorithmic approach to local state transition matrices, Anticipative and Predictive Models in Systems Science I, IIAS, S. 63-66.
2004 T. Wiebringhaus, C. Igel, J. Gebert, Protein fold class prediction using neural networks with tailored early-stopping. International Joint Conference on Neural Networks (IJCNN 2004), S. 1693-1697, IEEE Press.
2003 T. Wiebringhaus, U. Faigle, D. Schomburg, J. Gebert, C. Igel, G.-W. Weber, Protein fold class prediction using neural networks reconsidered. In Currents in Computational Molecular Biology, The Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 2003), S. 225-226.
2003 J. Gebert, M. Lätsch, N. Radde, S.W. Pickl, G.-W. Weber, R. Wünschiers, Genetic networks and anticipation of gene-expression patterns, AIP Conference Proceedings 718, S. 474-485.
2003 J. Gebert, S.W. Pickl, N. Shokina, G.-W. Weber, R. Wünschiers, Mathematical modeling and discrete approximation in evaluation and forecasting of expression-data, Electronic Notes in Discrete Mathematics 13, S. 52-56.
2002 J. Gebert, S.W. Pickl, N. Shokina, G.-W. Weber, R. Wünschiers, Algorithmic analysis of gene expression data with polyhedral structures, Similarity Methods (5th International Workshop), Eds.: B. Kröplin, S. Rudolph, J. Häcker, S. 79-87.

Lebenslauf

Persönliche Daten

Name	Jutta Christa Gebert
Adresse	Cranachstrasse 1, 60596 Frankfurt am Main
geboren am	06.12.1976 in Nürnberg
Staatsangehörigkeit	deutsch
Familienstand	verheiratet mit Volker Kudszus
Eltern	Prof. Dr. Alfred Gebert und Christa Gebert, geb. Detje
Geschwister	Dr. Claus Gebert

Schulbildung

1983-1987	Grundschule Laiz in Sigmaringen-Laiz
1987-1996	Annette-von-Droste-Hülshoff-Gymnasium in Münster

Studium und Praktika

1996-2002	Studium der Mathematik mit Nebenfach Angewandte Kulturwissenschaften an der Westfälischen Wilhelms-Universität Münster, Abschluss: Diplom
11.10.1999-14.07.2000	Studentische Hilfskraft im Institut für medizinische Informatik und Biomathematik in Münster
14.02.-07.04.2000	Praktikum bei der Westdeutschen Landesbank in der Abteilung Mathematische Beratung/ Operations Research
seit 01.03.2002	wissenschaftliche Mitarbeiterin am Zentrum für Angewandte Informatik (ZAIK), Universität zu Köln

Auslandsaufenthalte

09.07.-25.07.2004	Sommerschule EURO Summer Institute in der Middle East Technical University Ankara, Türkei
03.09.-30.10.2005	Forschungsaufenthalt in den Los Alamos National Laboratories, New Mexico, USA

Köln, im Mai 2007