Rekonstruktion und Modellierung medizinisch relevanter biologischer Interaktionsnetzwerke

Reconstruction and modelling of medically relevant biological interaction networks

Zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)
Genehmigte Dissertation von Isabella-Hilda Mendler aus Arad, Rumänien
Tag der Einreichung: 15. Juni 2023, Tag der Prüfung: 16. Oktober 2023

1. Gutachten: Prof. Dr. Barbara Drossel
2. Gutachten: Prof. Dr. Marc-Thorsten Hütt
Darmstadt, Technische Universität Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Physik
Institut für Physik Kondensierter Materie

Rekonstruktion und Modellierung medizinisch relevanter biologischer Interaktionsnetzwerke
Reconstruction and modelling of medically relevant biological interaction networks

Genehmigte Dissertation von Isabella-Hilda Mendler

1. Gutachten: Prof. Dr. Barbara Drossel
2. Gutachten: Prof. Dr. Marc-Thorsten Hütt

Tag der Einreichung: 15. Juni 2023
Tag der Prüfung: 16. Oktober 2023

Darmstadt, Technische Universität Darmstadt

Bitte zitieren Sie dieses Dokument als:
URN: urn:nbn:de:tuda-tuprints-247532
URL: http://tuprints.ulb.tu-darmstadt.de/24753

Dieses Dokument wird bereitgestellt von tuprints,
E-Publishing-Service der TU Darmstadt
http://tuprints.ulb.tu-darmstadt.de
tuprints@ulb.tu-darmstadt.de

La pensée n'est qu'un éclair au milieu d'une longue nuit.
Mais c'est cet éclair qui est tout.

(Henri Poincaré)

Erklärungen laut Promotionsordnung

§8 Abs. 1 lit. c PromO

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§9 Abs. 2 PromO

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 15. Juni 2023

Isabella-Hilda Mendler

Zusammenfassung

Diese Dissertation befasst sich mit der Analyse zweier biologischer Systeme, die eine wichtige Rolle im Zusammenhang mit nichtübertragbaren Krankheiten wie Diabetes oder Krebs spielen: dem menschlichen Mikrobiom, d.h. den Mikroorganismen, die unseren Körper besiedeln, und dem p53-Netzwerk, einem komplexen Protein- und Genregulationsnetzwerk rund um das als „Wächter des Genoms“ bekannte Tumorsupressorprotein p53. Trotz intensiver Forschungsbemühungen sind diese medizinisch höchst relevanten Systeme noch immer nicht vollständig verstanden. Sie sollen daher im Rahmen dieser Arbeit mithilfe von Methoden aus der statistischen Physik, genauer gesagt durch die Modellierung als dynamische Netzwerke, untersucht werden.

Eine große Herausforderung zum besseren Verständnis des menschlichen Mikrobioms besteht darin, ausgehend von Mikrobiomproben, wie z.B. Speichel- oder Stuhlproben, auf die zugrunde liegenden mikrobiellen Interaktionsnetzwerke zu schließen. Eine mögliche Vorgehensweise für diese sog. Netzwerkinferenz ist die von Claussen et al. entwickelte Entropy-Shifts-of-Abundance-vectors-under-Boolean-Operations (ESABO)-Methode [34], bei der die binarisierten Mikrobiomdaten als Attraktoren eines Booleschen Netzwerks interpretiert werden. Die ESABO-Methode weist jedoch einige signifikante Schwachpunkte auf: Zum einen werden mutualistische Interaktionen zwischen häufig vorkommenden Spezies aufgrund der Gestalt der verwendeten Entropiefunktion oft als antagonistische Interaktionen rekonstruiert, zum anderen ist die Methode insofern inkonsistent, als dass die mit der ESABO-Methode inferierten Netzwerke im Allgemeinen nicht die zur Inferenz genutzten mikrobiellen An- und Abwesenheitsmuster reproduzieren.
Der erste Schwachpunkt konnte in dieser Arbeit durch zwei simple Modifikation der ESABO-Methode, nämlich durch eine analytischen Berechnung der ESABO-Scores und dem Vertauschen von Nullen und Einsen in Abundanzvektoren mit einer hohen relativen Häufigkeit von Einsen, behoben werden.
Zur Behebung des zweiten Schwachpunkts wurde die ESABO-Methode um einen evolutionären Algorithmus erweitert bzw. in diesen eingebettet. Hierbei zeigte sich, dass die Kombination aus evolutionärem Algorithmus und ESABO-Methode, die im Folgenden als ESABO-gestützte Evolution bezeichnet wird, signifikant bessere Rekonstruktionsergebnisse liefert als die ESABO-Methode allein oder ein evolutionärer Algorithmus mit zufällig gewählten Mutationen, d.h. ohne Unterstützung der ESABO-Methode. Die Untersuchung der ESABO-gestützten Evolution bei unvollständiger Kenntnis der Attraktoren (d.h. einer begrenzten Menge an Mikrobiomproben, wie es auch in der Realität zu erwarten ist) ermöglichte es uns zudem, eine Beziehung zwischen dem Prozentsatz der bekannten Attraktoren und der durchschnittlichen Fitness, die durch die Inferenzmethode erreicht wurde, zu finden. Mithilfe dieser Beziehung konnte schließlich bei der Anwendung der ESABO-gestützten Evolution auf echte, empirische Speichelmikrobiomdaten die Vollständigkeit des betrachteten

Datensatzes abgeschätzt werden. Hierbei ergab sich, dass die verwendeten Daten weniger als 50% der Attraktoren des Systems abdeckten.

Im zweiten Teil der Arbeit wird mithilfe der dynamischen Netzwerkmodellierung eine wichtige Komponente des p53-Netzwerks, nämlich die Regulation des Proteins p21 durch das Protein p53 infolge von DNA-Doppelstrangbrüchen untersucht. Experimentelle Untersuchungen der Dynamik von p53 und p21 in einzelnen Brustepithelzellen zeigten, dass trotz einer relativ homogenen p53-Dynamik nach dem Auftreten von DNA-Doppelstrangbrüchen der Zeitpunkt und die Geschwindigkeit der p21-Induktion heterogen sind, wobei die Zellzyklusphase eine entscheidende Rolle zu spielen scheint. Zellen, die den Schaden in der S-Phase erfahren, zeigen eine verzögerte p21-Akkumulation, während Zellen, die in der G1-Phase bestrahlt werden und anschließend in die S-Phase übergehen, eine prompte pulsartige p21-Dynamik aufweisen. Das Ziel der Modellierung war es daher, den zellzyklusabhängigen molekularen Mechanismus zu identifizieren, der für diese Heterogenität verantwortlich ist. Durch den Vergleich von zwei minimalistischen Modellen mit den experimentellen p21-Daten und weiteren von Caibin Sheng durchgeführten Experimenten, konnte gezeigt werden, dass eine erhöhte p21-Abbaurate während der S-Phase in Kombination mit der Aktivierung von p21 durch p53 bereits ausreicht, um die auf Einzelzellebene beobachtete heterogene p21-Dynamik nach dem Auftreten von DNA-Doppelstrangbrüchen zu reproduzieren.

Abstract

This dissertation focuses on the analysis of two biological systems that play an important role in the context of noncommunicable diseases such as diabetes or cancer: the human microbiome, i.e. the microorganisms that colonise our body, and the p53 network, a complex protein and gene regulation network centered around the tumour suppressor protein p53, known as the „guardian of the genome“. Despite intensive research efforts, these medically highly relevant systems are still not fully understood. Therefore, in this thesis we will study them using methods from statistical physics, or to be more precise, by modelling them as dynamical networks.

A major challenge for a better understanding of the human microbiome is the deduction of the underlying microbial interaction networks from microbiome samples, such as saliva or stool samples. One possible approach for this so-called network inference is the Entropy Shifts of Abundance vectors under Boolean Operations (ESABO) method developed by Claussen et al. [34], where binarised microbiome patterns are interpreted as attractors of a Boolean network. However, the ESABO method has some significant weaknesses: First, mutualistic interactions between highly abundant species are often reconstructed as antagonistic interactions due to the shape of the used entropy function, and second, the method is inconsistent in the sense that networks inferred by the ESABO method do not generally reproduce the same microbial presence/absence patterns that were used for their inference.
The first weakness was addressed in this work by two simple modifications to the ESABO method, namely an analytical calculation of the ESABO scores and the swapping of zeros and ones in abundance vectors with a high relative frequency of ones.
To address the second weakness, the ESABO method was extended by or rather embedded in an evolutionary algorithm. This combination of an evolutionary algorithm and the ESABO method, which we will refer to as ESABO enhanced evolution, was found to provide significantly better reconstruction results than the ESABO method alone or an evolutionary algorithm with randomly chosen mutations, i.e. without the support of the ESABO method. Futhermore, the investigation of the ESABO enhanced evolution with incomplete knowledge of the attractors (i.e. a limited set of microbiome samples, as is to be expected in reality) allowed us to find a relationship between the percentage of known attractors and the average fitness obtained by the inference method. Finally, this relationship was used to estimate the completeness of a real empirical salivary microbiome dataset. It was found that the examined data covered less than 50% of the attractors of the system.

In the second part of this thesis, dynamic network modelling is used to investigate an important component of the p53 network, namely the regulation of the protein p21 by the protein p53 upon DNA double strand breaks (DSBs). Experimental studies of the dynamics of p53 and p21 in single breast epithelial cells showed that despite a relatively homogeneous p53

dynamics after the occurrence of DSBs, the timing and rate of p21 induction is heterogeneous, with the cell cycle phase playing a crucial role. Cells that experience the damage in S phase show a delayed p21 accumulation, whereas cells that are irradiated in G1 and subsequently progress to S phase show a prompt pulse-like p21 response. The aim of the modelling was therefore to identify the cell cycle-dependent molecular mechanism responsible for this heterogeneity. By comparing two minimalistic models with the experimental p21 data and through further experiments, performed by Caibin Sheng, it was shown that an increased p21 degradation rate during S phase in combination with the activation of p21 by p53 is sufficient to reproduce the heterogeneous p21 dynamics observed at the single cell level after the occurrence of DSBs.

Inhaltsverzeichnis

Abkürzungsverzeichnis

ATM Ataxia Telangiectasia Mutated (Proteinkinase)

AUC Fläche unter der ROC-Kurve (*Englisch*: area under the ROC curve)

BrdU 5-Bromo-2'-deoxyuridine

Cas9 CRISPR associated 9

Cdk Cyclin-abhängige Kinase (*Englisch*: cyclin-dependent kinase)

CFP cyan fluoreszierendes Protein (*Englisch*: cyan fluorescent protein)

Chk2 Checkpoint Kinase 2

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

DGL Differentialgleichung

DNA Desoxyribonukleinsäure (*Englisch*: deoxyribonucleic acid)

dNTP Desoxyribonucleosidtriphosphat

DSB DNA-Doppelstrangbruch

EdU 5-Ethynyl-2'-deoxyuridine

ESABO Entropy-Shifts-of-Abundance-vectors-under-Boolean-Operations

FPR Falsch-Positiv-Rate (*Englisch*: false positive rate)

HMP Human-Microbiome-Project

Mdm2 Mouse double minute 2 homolog

mRNA Boten-RNA (*Englisch*: messenger RNA)

NCD nichtübertragbare Krankheit (*Englisch*: Noncommunicable disease)

NGS Next-Generation-Sequencing

OTU operative taxonomische Einheit (*Englisch*: operational taxonomic unit)

PCNA Proliferating Cell Nuclear Antigen

PCR Polymerasekettenreaktion (*Englisch*: polymerase chain reaction)

QIIME Quantitative Insights Into Microbial Ecology

RFP rot fluoreszierendes Protein (*Englisch*: red fluorescent protein)

RNA Ribonukleinsäure (*Englisch*: ribonucleic acid)

ROC Receiver-Operating-Characteristics

rRNA ribosomale RNA

TPR Richtig-Positiv-Rate (*Englisch*: true positive rate)

Wip1 Wild-type p53-induced phosphatase 1

YFP gelb fluoreszierendes Protein (*Englisch*: yellow fluorescent protein)

1. Einleitung

1.1. Motivation

Eine der größten Bedrohungen für die menschliche Gesundheit stellen nichtübertragbare Krankheiten (*Englisch*: Noncommunicable diseases, NCDs) dar, die heutzutage die Haupttodesursache weltweit sind. Mit jährlich etwa 41 Millionen Todesopfern machen diese Krankheiten ungefähr 74% aller Todesfälle weltweit aus [160] und in Deutschland stellen NCDs sogar in schätzungsweise 91% aller Todesfälle die Ursache dar [38]. Die vier wichtigsten NCDs sind dabei Herz-Kreislauf-Erkrankungen (z.B. Herzinfarkt oder Schlaganfall), Krebs, chronische Atemwegserkrankungen (z.B. Asthma) und Diabetes [160].
Trotz der hohen medizinischen Relevanz dieser vier Erkrankungen gibt es in der Regel noch keine Möglichkeit zu ihrer Heilung, und die Mechanismen, die zu ihrer Entstehung führen, sind in viele Fällen noch unzureichend verstanden. Um die Entstehung dieser Krankheiten tiefgreifend zu verstehen und geeignete bzw. effiziente Behandlungsmethoden zu finden, ist ein Verständnis der für unsere Gesundheit essentiellen biologischen Systeme unerlässlich.

Zu diesen für unsere Gesundheit essentiellen biologischen Systemen, die im Kontext vieler nichtübertragbarer Krankheiten eine wichtige Rolle spielen [3, 138] und mit denen wir uns im Rahmen dieser Arbeit beschäftigen wollen, gehört u.a. das *menschliche Mikrobiom*, d.h. die Gesamtheit aller Mikroorganismen, die unseren Körper besiedeln.
Zahlreiche Studien der letzten 15 Jahre haben gezeigt, dass Veränderungen im Mikrobiom mit verschiedenen NCDs, wie z.B. Asthma [59, 148], Krebs [73, 137, 145] oder Herz-Kreislauf-Erkrankungen [156, 157] in Verbindung gebracht werden können. Besonders intensiv wurde dabei der Zusammenhang zwischen dem Darmmikrobiom und Diabetes-Typ-2 erforscht [92, 129, 80, 70, 162, 154]. Diese Untersuchungen haben gezeigt, dass Menschen mit Typ-2-Diabetes oft eine abweichende Zusammensetzung ihres Mikrobioms im Vergleich zu gesunden Personen aufweisen. Es scheint eine Reduktion bestimmter (Butyrat-produzierender) Bakterienarten zu geben, die für die Aufrechterhaltung einer gesunden Stoffwechselfunktion wichtig sind, während andere Bakterienarten (insbesondere opportunistische Krankheitserreger) übermäßig vorhanden sind [70, 80, 129].

Ein zweites für unsere Gesundheit essentielles biologisches System, welches in dieser Arbeit betrachtet werden soll, ist das sogenannte *p53-Netzwerk*. Hierbei handelt es sich um ein komplexes Protein- bzw. Genregulationsnetzwerk rund um den als „Wächter des Genoms“ [90] bekannten Tumorsupressor p53. p53 ist ein Protein, welches für die adäquate Reaktion unserer Zellen auf DNA-Schäden verantwortlich ist und so entscheidend dazu beiträgt, die Entstehung von Krebs zu verhindern. Dies wird insbesondere dadurch deutlich, dass das für p53 kodierende Gen (TP53) das am häufigsten mutierte bzw. inaktivierte Gen bei Krebserkrankungen ist [81, 68, 121, 159].

Trotz intensiver Forschungsbemühungen in den vergangenen Jahrzehnten [153, 128, 99, 91], sind diese beiden Systeme noch nicht vollständig verstanden [64, 16, 68] und aufgrund ihrer immensen medizinischen Bedeutung immer noch Gegenstand aktueller Forschung (siehe z.B. [154, 105] zur Mikrobiomforschung und [97, 98] zur p53-Forschung). Als Teil dieser Forschung verfolgt diese Arbeit das Ziel, einen Beitrag zum besseren Verständnis dieser zwei für unsere Gesundheit so wichtigen biologischen Systeme zu leisten.

Eine allgemeine Herausforderung bei der Untersuchung biologischer Systeme ist die Tatsache, dass es sich hierbei in der Regel um *komplexe Systeme* handelt [104, 117]. Dies sind im weitesten Sinne Systeme, die aus einer großen Zahl von relativ einfachen[1], miteinander interagierenden Bestandteilen aufgebaut sind, die keiner zentralen Kontrolle oder Steuerung unterliegen [116], und die ein *komplexes dynamisches Verhalten* zeigen. Für die Dynamik dieser Systeme ist dabei insbesondere der Begriff der *Nichtlinearität* wichtig [116]: Das Verhalten des Gesamtsystems lässt sich nicht durch die isolierte Analyse seiner elementaren Bestandteile vorhersagen [11] bzw. – um es mit den Worten von Aristoteles zu sagen – das Ganze ist mehr als die Summe seiner Teile. Aus diesem Grund ist ein intuitives Verständnis biologischer Systeme nahezu unmöglich und es müssen gezielte formale Methoden zu ihrer Analyse eingesetzt werden.

Eine vielversprechende Möglichkeit zur Analyse socher komplexen biologischen Systeme, die wir in dieser Arbeit nutzen wollen, besteht darin, sie als *Netzwerke* zu beschreiben [12, 13, 42, 94, 125]. Im mathematischen Kontext entspricht ein Netzwerk dabei einem *Graphen*, wobei jeder Graph aus einer Menge von *Knoten* und *Kanten* besteht. Jede Kante verbindet zwei Knoten und kann entweder gerichtet oder ungerichtet sein. Die genauen Eigenschaften der Knoten sowie die Bedeutung und Gestalt der Kanten bzw. Verbindungen hängen vom jeweiligen Kontext ab. [24]
Möchte man beispielsweise eine mikrobielle Gemeinschaft in Form eines Netzwerks beschreiben, so entsprechen die Knoten in der Regel den verschiedenen biologischen Spezies und die Kanten bilden die unterschiedlichen Interaktionen (wie z.B. Konkurrenz oder Mutualismus) zwischen diesen Spezies ab [31].
In zellulären Systemen wiederum können die Knoten unter anderem Gene, mRNA oder Proteine repräsentieren. Während sich gerichtete Kanten besonders zur Darstellung chemischer Umwandlungen oder regulatorischer Beziehungen eignen, werden ungerichtete Kanten meist zur Beschreibung gegenseitiger Wechselwirkungen, wie z.B. Protein-Protein-Bindungen, genutzt. Je nach vorliegendem Wissensstand werden die Kanten dabei durch simple Vorzeichen (z.B. positiv für Aktivierung und negativ für Inhibierung) oder Gewichte charakterisiert, die u.a. die Stärke der Interaktion, die Reaktionsgeschwindigkeit oder die Wahrscheinlichkeit für das Vorliegen dieser Interaktion angeben können. [5]

Der Teilbereich der Biologie, der sich mit der quantitativen Beschreibung biologischer Systeme auf Systemebene und somit mit biologischen Netzwerken beschäftigt, ist die sogenannte *Systembiologie* [85]. Das Ziel der Systembiologie ist es dabei das dynamische Verhalten biologischer, interagierender Systeme zu verstehen, vorherzusagen und, wenn möglich, zu steuern [5]. Hierzu kombiniert die Systembiologie vor Allem Prinzipien und Methoden aus der theoretischen Physik und Mathematik sowie dem Ingenieurwesen und der Bioinformatik [131],

[1] Einfach im Vergleich zum Gesamtsystem

um komplexe biologische Systeme auf verschiedenen Organisationsebenen – angefangen bei der molekularen und zellulären Ebene bis hin zum gesamten Organismus und sogar Ökosystemen – zu modellieren und zu analysieren.

Zwei Themengebiete bzw. Methoden, die für die Systembiologie dabei von entscheidender Bedeutung sind und die wir im Rahmen dieser Arbeit verwenden werden, sind die *Netzwerkinferenz* und die *dynamische Netzwerkmodellierung*.

Die *Netzwerkinferenz* befasst sich mit der Fragestellung, wie auf Grundlage von Informationen über die Identität und den Zustand der Elemente eines Systems auf Interaktionen oder funktionelle Beziehungen zwischen diesen Elementen geschlossen werden kann [5]. Das Ziel dabei ist es, den dem System zugrunde liegenden Interaktionsgraphen zu rekonstruieren. Dieser Interaktionsgraph kann dann beispielsweise im Zuge der sog. *Graphenanalyse*, bei der Methoden aus der Graphentheorie zur Analyse der Netzwerktopologie bzw. -struktur angewendet werden, weiter untersucht werden. Die Idee hierbei ist, dass man hofft, aus dem Wissen über die Struktur eines Netzwerks auch mehr Aufschluss über seine Funktion gewinnen zu können. So hat z.B. die Analyse von Genregulationsnetzwerken in Zellen zur Entdeckung sog. *Netzwerkmotive* geführt [8, 9]. Dies sind Muster von Verbindungen, d.h. Subgraphen, die in echten Genregulationsnetzwerken mit einer viel höheren Häufigkeit auftreten als es in zufälligen Netzwerken zu erwarten wäre [140] und denen bestimmte informationsverarbeitende Funktionen zugeordnet werden können [9].

Bei der *dynamischen Netzwerkmodellierung* geht es hingegen darum, nicht nur die Struktur des Netzwerks zu betrachten, sondern auch seine Dynamik. Hierzu wird zunächst jedem Knoten des Netzwerks eine Variable zugeordnet, die seinen Zustand, also z.B. die Populationsgröße oder Konzentration der entsprechenden Spezies, repräsentiert. Des Weiteren wird jedem Knoten eine sog. *Interaktionsfunktion* zugewiesen, die angibt, wie sich der Zustand dieses Knotens in Abhängigkeit des Verhaltens der anderen, mit ihm verbundenen Knoten im Laufe der Zeit verändert. Ausgehend von der Topologie des Netzwerks und einem bestimmten Anfangszustand seiner Knoten, lässt sich dann mit Hilfe der Interaktionsfunktionen die zeitliche Entwicklung des gesamten Systems bestimmen und unter verschiedenen Bedingungen analysieren. [2]

Im ersten Teil dieser Arbeit werden wir uns mit der Entwicklung einer neuen Methode zur Inferenz mikrobieller Interaktionsnetzwerke beschäftigen. Hierbei betrachten wir das Mikrobiom, im Gegensatz zu den meisten anderen Ansätzen, aus einer binären Perspektive, d.h. lediglich das Muster des Vorhanden- bzw. Nichtvorhandenseins der unterschiedlichen mikrobiellen Spezies in einer Mikrobiomprobe (z.B. einer Speichel- oder Stuhlprobe). Dies hat den Vorteil, dass wir Mikrobiomproben bzw. die gemessenen mikrobiellen An- und Abwesenheitsmuster als Attraktoren, oder genauer gesagt als Fixpunkte, eines Booleschen Netzwerks interpretieren können (siehe Abschnitt 2.2.2). Ausgehend von diesen Attraktoren soll dann mithilfe der hier entwickelten Methode, die auf der sog. ESABO-Methode [34] aufbaut und diese um einen evolutionären Algorithmus erweitert, ein Interaktionsnetzwerk zwischen den in den Mikrobiomproben vorhandenen Spezies inferiert werden. Die entscheidende Idee, die wir hierbei verfolgen, ist, dass das korrekt inferierte Netzwerk die ursprünglich beobachteten An- und Abwesenheitsmuster als seine Attraktoren reproduzieren können sollte. Der neu hinzugefügte evolutionäre Algorithmus dient daher dem Zweck, den Überlapp zwischen der Menge der Attraktoren des inferierten Netzwerks und der Menge der ursprünglichen (beobachteten) binären Häufigkeitsmustern zu maximieren.

Die Forschungsfragen, die in diesem ersten Teil der Arbeit, also im Kontext der Inferenz mikrobieller Interaktionsnetzwerke, beantwortet werden sollen, sind:

- Wie kann aus Mikrobiomdaten, die als Fixpunkte eines Booleschen Netzwerks interpretiert werden, die Netzwerktopologie rekonstruiert bzw. inferiert werden, sodass das resultierende Netzwerk wieder die zur Inferenz verwendeten Attraktoren aufweist?
- Wie gut ist die Qualität bzw. Güte der in dieser Dissertation neu eingeführten Inferenzmethode?
- Welche Auswirkungen hat es auf das Inferenzergebnis, wenn nicht alle Attraktoren des Netzwerks zu seiner Rekonstruktion zur Verfügung stehen?
- Lässt sich mithilfe der hier eingeführten Inferenzmethode die Vollständigkeit eines realen biologischen Datensatzes abschätzen?

Im zweiten Teil der Arbeit soll mithilfe der dynamischen Netzwerkmodellierung eine wichtige Komponente des p53-Netzwerks, nämlich die Regulation des Proteins p21 durch das Protein p53 betrachtet werden. p53 ist ein sog. Transkriptionsfaktor, d.h. ein Protein, welches die Expression anderer Gene reguliert, indem es an spezifische Stellen auf der DNA bindet und so die Rate der RNA-Synthese beeinflusst (siehe Abschnitt 2.1.5). Nach dem Auftreten von DNA-Schäden wird p53 durch verschiedene Proteine phosphoryliert und somit aktiviert. Nach seiner Aktivierung induziert p53 die Expression hunderter Zielgene [55]. Diese Zielgene werden dabei u.a. für die DNA-Reparatur, die Kontrolle des Zellzyklus und die Apoptose, d.h. den programmierten Zelltod, benötigt [68, 81]. Eines der wichtigsten und am längsten erforschten Zielgene von p53 ist CDKN1A, welches das Protein p21 kodiert [41]. Dieses Protein spielt in unseren Zellen eine wichtige Rolle bei der Regulation des Zellzyklus: p21 kann den Zellzyklus infolge von DNA-Schäden anhalten und somit somit je nach Schwere des Schadens die Teilung der Zelle dauerhaft verhindern (sog. Seneszenz) oder den Zellzyklus lediglich vorübergehend stoppen, um der Zelle die notwendige Zeit zur Reparatur zu ermöglichen. Dies ist wichtig, um die Entstehung von Krebs zu verhindern, sodass die Regulation von p21 durch p53 von besonderer Relevanz ist.
Den Ausgangspunkt für die in dieser Dissertation durchgeführte Modellierung zur Regulation von p21 durch p53 stellen experimentelle Untersuchungen der Dynamik dieser zwei Proteine in einzelnen lebenden Brustepithelzellen dar, die hauptsächlich von Caibin Sheng in der Arbeitsgruppe von Prof. Dr. Alexander Löwer an der TU Darmstadt durchgeführt wurden.
Diese Untersuchungen zeigten, dass obwohl verschiedene Zellen eine relativ homogene p53-Dynamik nach dem Auftreten von DNA-Doppelstrangbrüchen aufweisen, ihre p21-Dynamik deutlich heterogener ist. Sowohl beim Zeitpunkt als auch bei der Geschwindigkeit der p21-Induktion gibt es selbst zwischen isogenen Zellen deutliche Unterschiede, wobei die Zellzyklusphase, in welcher der Schaden auftritt, eine entscheidende Rolle zu spielen scheint. Zellen, die den Schaden in der S-Phase erfahren, zeigen eine verzögerte p21-Akkumulation, während Zellen, die in der G1-Phase bestrahlt werden und anschließend in die S-Phase übergehen, eine prompte pulsartige p21-Dynamik aufweisen. Das Ziel, welches durch die hier durchgeführte Modellierung verfolgt wird, ist daher, den Grund für die beobachtete Heterogenität näher zu untersuchen. Die zentrale Forschungsfrage hierbei ist:

- Welcher zellzyklusabhängige molekulare Mechanismus ist für die Heterogenität der p21-Antwort infolge von DNA-Doppelstrangbrüchen verantwortlich?

1.2. Aufbau der Arbeit

Die Arbeit ist wie folgt gegliedert:

In Kapitel 2 werden zunächst die zum allgemeinen Verständnis dieser Arbeit notwendigen biologischen und mathematischen Grundlagen erläutert.

Kapitel 3 beschäftigt sich mit der Inferenz mikrobieller Interaktionsnetzwerke. Nach einer allgemeinen Einführung in das Themengebiet (Kapitel 3.1 und 3.2) sowie der Vorstellung der von Claussen et al. eingeführten ESABO-Methode zur Inferenz mikrobieller Interaktionsnetzwerke (Kapitel 3.3 und 3.4), liegt der Fokus dieses Kapitels auf der Entwicklung einer neuen, auf der ESABO-Methode aufbauenden Inferenzmethode (Kapitel 3.5 und 3.6). Wir untersuchen sowohl die Güte dieser Methode für unterschiedliche Szenarien (Kapitel 3.7.1 und 3.7.2) als auch eine beispielhafte Anwendung auf echte biologische Daten für das menschliche Speichelmikrobiom (Kapitel 3.7.3). Das Kapitel endet schließlich mit einem Fazit und der Diskussion der Ergebnisse (Abschnitt 3.8).

Kapitel 4 widmet sich der Modellierung eines bedeutenden Bestandteils des p53-Netzwerks, nämlich der Regulation des Proteins p21 durch das Tumorsuppressorprotein p53. Das Kapitel beginnt zunächst mit einer allgemeinen Einführung zum p53-Netzwerk und stellt die beiden relevanten Proteine p53 und p21 vor. Im Anschluss werden in Abschnitt 4.2 die neuen, von Caibin Sheng (Arbeitsgruppe von Alexander Löwer, TU Darmstadt) erzielten experimentellen Ergebnisse vorgestellt, die den Anlass für die in dieser Dissertation durchgeführte Modellierung darstellen. Nach einer kurzen Erläuterung der sich aus den experimentellen Daten ergebenden Fragestellung (Abschnitt 4.3), werden in Kapitel 4.4 zwei verschiedene Modelle zur Erklärung der Beobachtungen entwickelt und diskutiert. Kapitel 4.5 präsentiert die Ergebnisse der Modellierung, die wiederum weitere experimentelle Untersuchungen anregten. Diese zusätzlichen Experimente sind in Kapitel 4.6 beschrieben und werfen eine neue Frage bezüglich des Abbaus von p21 auf (Kapitel 4.7), die mithilfe unseres Modells in Kapitel 4.8 untersucht wird. Abschließend werden in Kapitel 4.9 alle Ergebnisse zur zellzyklusabhängigen Regulation von p21 durch p53 nochmal zusammengefasst und diskutiert.

Zum Abschluss dieser Dissertation wird in Kapitel 5 ein Fazit gezogen.

2. Grundlagen

2.1. Zellbiologischer Hintergrund

Im folgenden Abschnitt werden zunächst die wichtigsten zellbiologischen Grundlagen erläutert, da die Zelle das Umfeld zahlreicher biologischer Netzwerke, wie beispielsweise von Genregulationsnetzwerken, Signalübermittlungsnetzwerken oder metabolischen Netzwerken, ist. Hierbei werden wir zunächst auf den Unterschied zwischen Eukaryoten und Prokaryoten eingehen und zentrale Zellbestandteile, wie DNA oder Proteine, den Prozess der Genexpression und -regulation sowie den Zellzyklus kennenlernen.
Wenn nicht gesondert angegeben, sind die nachfolgenden biologischen Grundlagen den Quellen [7] und [6] entnommen.

2.1.1. Zellen: Eukaryoten und Prokaryoten

Alle Lebewesen sind aus *Zellen* aufgebaut. Diese sind kleine, von einer Membran umschlossene Einheiten, die mit einer konzentrierten wässrigen Lösung von Chemikalien, dem sog. *Cytoplasma*, gefüllt sind und über die besondere Fähigkeit verfügen, durch Wachstum und anschließende Teilung Kopien von sich selbst zu erzeugen. Allen Zellen ist zudem gemeinsam, dass sie *Desoxyribonukleinsäure (*Englisch*: deoxyribonucleic acid, DNA)* enthalten, welche zur Speicherung ihrer genetischen Information bzw. Erbinformation dient (siehe Kapitel 2.1.2). Des Weiteren enthält jede Zelle auch *Ribonukleinsäure (*Englisch*: ribonucleic acid, RNA)*, ein der DNA ähnliches Makromolekül, das vor allem zur Genexpression benötigt wird (siehe Kapitel 2.1.3), und eine Vielzahl von *Proteinen*, die aus einer linearen Kette von Aminosäuren bestehen. Proteine katalysieren als Enzyme fast alle Reaktionen in der Zelle und haben viele weitere Funktionen (siehe Kapitel 2.1.4).

Im Allgemeinen lassen sich alle Lebewesen anhand ihrer Zellstruktur in zwei Gruppen unterteilen, nämlich in *Eukaryoten*, zu denen u.a. mehrzellige Organismen wie z.B. Pflanzen und Tiere gehören, und in *Prokaryoten*, die in der Regel Einzeller sind und Bakterien sowie Archaeen[1] umfassen. Die Zellen der Eukaryoten zeichnen sich dadurch aus, dass sie einen sog. *Zellkern* oder *Nukleus* besitzen, bei dem es sich um ein membranumschlossenes intrazelluläres Kompartiment handelt, in dem sich die (chromosomale) DNA des jeweiligen Organismus befindet. Prokaryotische Zellen haben im Gegensatz dazu keinen Zellkern. Sie sind bezüglich ihrer inneren Struktur mit nur einem einzigen Kompartiment, das die gesamte DNA, RNA und alle Proteine des Organismus enthält, deutlich einfacher aufgebaut als eukaryotische Zellen, die mehrere sog. *Zellorganellen* besitzen. Zudem sind prokaryotische Zellen, die meist

[1] Einzeller, die Bakterien ähnlich sehen, aber auf molekularer Ebene häufig, insbesondere in Bezug auf ihre Maschinerie zur Verarbeitung genetischer Informationen (Transkription, Translation, Replikation), enger mit den Eukaryoten verwandt sind.

kugel-, stäbchen- oder spiralförmig sind, mit einer Größe von typischerweise $1 - 10\,\mu m$ [87] deutlich kleiner als eukaryotische Zellen, die typischerweise $10 - 100\,\mu m$ [87] groß sind. Nichtsdestotrotz sind Prokaryoten auf molekularer Ebene weitaus vielfältiger als Eukaryoten und können ökologische Nischen mit z.B. extremen Temperaturen, Salzkonzentrationen oder begrenztem Nährstoffangebot besetzen (siehe hierzu z.B. [114]). Es gibt sogar Arten, die ihre gesamte Energie und Nährstoffe aus anorganischen chemischen Quellen beziehen können [7, 83, 84].

Auch der menschliche Körper, welcher aus etwa 10^{13} eukaryotischen Zellen besteht [22], ist von einer Vielzahl von Prokaryoten (Archaeen und Bakterien) sowie Pilzen und Viren besiedelt. Die Gemeinschaft dieser Mikroorganismen (bzw. im engeren Sinn deren Genome) wird im Allgemeinen als *menschliches Mikrobiom* bezeichnet. Das menschliche Mikrobiom enthält mindestens genauso viele Zellen wie unser Körper [64, 139] und spielt für unsere Gesundheit eine entscheidende Rolle. So helfen die Bakterien in unserem Darm (das sog. *intestinale Mikrobiom*) beispielsweise bei der Verdauung unserer Nahrung und spielen eine wichtige Rolle für die Entwicklung und Aufrechterhaltung eines gesunden Immunsystems. Lediglich eine kleine Minderheit von Bakterien ist den Pathogenen, d.h. krankheitserregenden Lebewesen, zuzuordnen. [135]

2.1.2. DNA

Wie bereits erwähnt, speichert jede Zelle ihre genetische Information in Form von *Desoxyribonukleinsäure (*Englisch*: deoxyribonucleic acid, DNA)*. Jedes DNA-Molekül besteht aus zwei *DNA-Strängen*, bei denen es sich um lange, unverzweigte, gepaarte Polymerketten handelt, die immer aus denselben vier Arten von Monomeren, den sog. *Nukleotiden*, aufgebaut sind. Ein Nukleotid besteht dabei aus einem Zucker (Desoxyribose), an den eine Phosphatgruppe gebunden ist, und einer *Base*, bei der es sich entweder um Adenin (A), Guanin (G), Cytosin (C) oder Thymin (T) handeln kann (siehe Abb. 2.1 (a)). Über ihre Zucker und Phosphate sind die Nukleotide eines DNA-Strangs kovalent miteinander verbunden und bilden so ein stabiles Grundgerüst aus alternierenden Zucker- und Phosphat-Molekülen, das sog. *Zucker-Phosphat-Rückgrat*, an welchem die Basen hängen. Die charakteristische dreidimensionale Struktur der DNA, bei der es sich um eine Doppelhelix handelt (siehe Abb. 2.1 (b)), entsteht durch die Zusammenlagerung zweier solcher DNA-Stränge. Zusammengehalten werden die beiden Stränge durch Wasserstoffbrücken zwischen ihren Basen, wobei (nahezu) immer Adenin mit Thymin und Guanin mit Cytosin verbunden ist. Durch diese eindeutige Basenpaarung haben beide DNA-Stränge denselben Informationsgehalt, und die Erbinformation jeder Zelle, die durch die Reihenfolge der vier verschiedenen Basen in einem Strang gespeichert ist, liegt in doppelter Ausführung vor. Diese Tatsache wird u.a. zur Vervielfältigung bzw. *Replikation* der DNA ausgenutzt. Im Zuge dieser werden zunächst die beiden DNA-Stränge voneinander getrennt, was aufgrund ihrer vergleichsweise schwachen Bindung zueinander (Wasserstoffbrücken im Vergleich zu kovalenten Bindungen des Zucker-Phosphat-Rückgrates) ohne Beschädigungen möglich ist. Anschließend wird jeder Strang als Vorlage (engl.: template) zur Erzeugung eines neuen komplementären DNA-Stranges verwendet. Die Synthese des neuen DNA-Stranges erfolgt dabei durch Enzyme (s. Abschnitt 2.1.4), die man als *DNA-Polymerasen* bezeichnet.

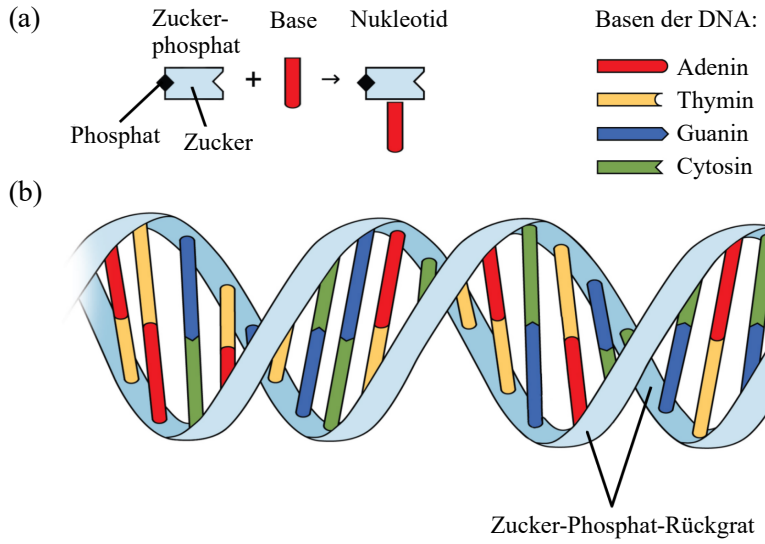


Abbildung 2.1.: Schematische Darstellung eines Nukleotids und der vier DNA-Basen (a) sowie der Doppelhelix-Struktur der DNA (b) nach [21].

2.1.3. Genexpression: Von der DNA zum Protein

Abschnitte der DNA, die für ein Protein oder eine Ribonukleinsäure (*Englisch*: ribonucleic acid, RNA) kodieren, bezeichnet man als *Gene*. Die sog. *Genexpression*, d.h. die Übersetzung der Nucleotidsequenz eines Gens in die Aminosäurensequenz eines Proteins ist ein zweistufiger Prozess, welcher aus der sog. *Transkription* und der sog. *Translation* (siehe Abbildung 2.2) besteht. Im Nachfolgenden werden beide Schritt kurz beschrieben, wobei sich die Ausführungen auf eukaryotische Zellen beziehen.
Bei der *Transkription* wird die genetische Information von der DNA auf die RNA übertragen. RNA ist ein der DNA strukturell ähnliches Makromolekül, bei dem der Zucker Desoxyribose durch Ribose sowie die Base Thymin (T) durch Uracil (U) ersetzt ist. Im Gegensatz zur DNA bildet die RNA allerdings keine Doppelhelix aus und kann sich daher zu vielfältigen Strukturen falten.
Der Transkriptionsprozess geschieht ähnlich wie die DNA-Replikation, indem zunächst die Stränge der DNA in einem kleinen Bereich aufgetrennt werden und ein Strang dann aufgrund der komplementären Basenpaarung (Adenin (A) paart mit Uracil (U) und Guanin (G) mit Cytosin (C)) als Schablone (Englisch: template) für die Synthese eines RNA-Moleküls dient. Die Synthese des RNA-Moleküls erfolgt dabei durch sog. *RNA-Polymerasen*, die zur Initiation der Transkription an speziellen Bindungsstellen der DNA, den sog. *Promotern*, binden.
Das Resultat der Transkription proteinkodierender Gene ist nach weiterer Prozessierung der RNA, im Rahmen derer beispielsweise nicht kodierende DNA-Bereiche (sog. Introns) entfernt werden, die sogenannte Boten-RNA (*Englisch*: messenger RNA, mRNA). Diese kann den Zellkern verlassen und wird anschließend im Zuge der *Translation* an den sog. *Ribosomen*[2] in ein Protein, bzw. dessen Aminosäurensequenz, übersetzt.

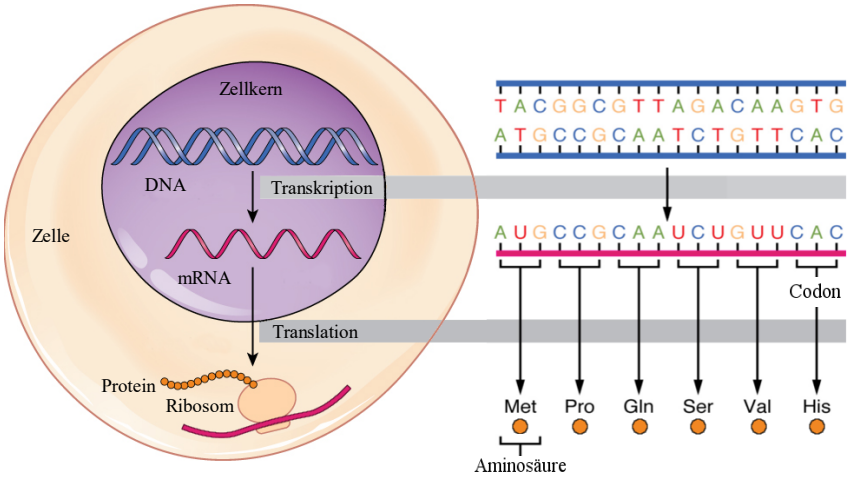


Abbildung 2.2.: Schematische Darstellung der Abläufe zur Synthese eines Proteins im Rahmen der Genexpression nach [21]. Während der Transkription wird die genetische Information zunächst von der DNA auf die mRNA übertragen. Bei der anschließenden Translation wird die Nucleotidsequenz der mRNA dann in die Aminosäurensequenz eines Proteins übersetzt.

[2] Multimolekulare Komplexe aus Proteinen und RNA, an denen im Cytoplasma die Proteinsynthese stattfindet [135].

Da es insgesamt 20 verschiedene Aminosäuren gibt, aus denen Proteine aufgebaut sein können, werden drei Basen der DNA bzw. mRNA benötigt, um eine Aminosäure zu kodieren. Eine solche Kombination von drei Basen wird als *Codon* bezeichnet. Neben den 61 Codons, die für eine Aminosäuren kodieren, gibt es auch drei sog. *Stopcodons*, welche das Ende der Translation markieren.

2.1.4. Proteine: Struktur und Funktion

Wie bereits erwähnt wurde, sind *Proteine* Polymere, die aus Aminosäuren zusammengesetzt sind. Die Aminosäuren sind dabei über kovalente Peptidbindungen miteinander verknüpft, weshalb Proteine auch als *Polypeptide* bezeichnet werden. Jedes Protein besitzt eine einzigartige Aminosäurensequenz, die sowohl seine endgültige dreidimensionale Struktur, die sog. *Konformation*, als auch seine biologische Aktivität maßgeblich bestimmt. Die Konformation eines Proteins wird dabei vor allem durch nicht kovalente Wechselwirkungen (Wasserstoffbrückenbindungen, Ionenbindungen, van-der-Waals-Kräfte oder hydrophobe Wechselwirkungen) zwischen verschiedenen Teilen der Polypeptidkette festgelegt.

Eine Gemeinsamkeit aller Proteine ist, dass sie an andere Moleküle, die man in diesem Kontext als sog. *Liganden* bezeichnet, binden. In der Regel ist diese Bindung sehr spezifisch bzw. selektiv, d.h. dass jedes Protein nur ganz bestimmte, wenige Molekülarten mit hoher Affinität bindet. Die Bindung eines Ligandenmoleküls, bei dem es sich beispielsweise um ein anderes Makromolekül oder ein Ion handeln kann, geschieht dabei über die gleichen nicht kovalenten Wechselwirkungen, die auch für die dreidimensionale Struktur des Proteins entscheidend sind. Da jede einzelne nicht kovalente Bindung sehr schwach ist, müssen viele solcher Bindungen ausgebildet werden, damit der Ligand fest an das Protein gebunden wird. Dies kann nur geschehen, wenn die Oberflächenkontur des Liganden perfekt zur Oberfläche des Proteins passt (sog. „Schlüssel-Schloss-Prinzip“). Den Bereich des Proteins, an welchen der Ligand gebunden wird, bezeichnet man auch als *Bindungsstelle* des Liganden. Solch eine Bindungsstelle – und damit auch die biologische Funktion eines Proteins – ist aufgrund des Schlüssel-Schloss-Prinzips sehr empfindlich gegenüber Änderungen der dreidimensionalen Proteinkonformation. Verursacht werden können solche Konformationsänderungen z.B. durch die (reversible) Bindung eines anderen, zweiten Liganden oder durch eine kovalente Modifikation des Proteins, wie beispielsweise der Addition einer geladenen Phosphatgruppe (sog. *Phosphorylierung*) [135]. Von Zellen wird dies häufig dazu genutzt, um die biologische Funktion eines Proteins genau zum benötigten Zeitpunkt zu aktivieren oder zu deaktivieren (siehe auch Abschnitt 2.1.5 und 2.1.6).

Proteine, die sich in vielzählige Klassen unterteilen lassen, erfüllen in der Zelle zahlreiche wichtige Funktionen. Einige dieser Funktionen sollen im Folgenden kurz vorgestellt werden, wobei kein Anspruch auf Vollständigkeit besteht.
Strukturproteine, wie Tubulin oder Aktin, dienen beispielsweise der mechanischen Stabilisierung der Zelle, indem sie u.a. das Cytoskelett bilden. Proteine, die man als *Motorproteine* bezeichnet, sind für die Erzeugung von Bewegungen in Zellen und Geweben verantwortlich. Das Motorprotein Kinesin interagiert z.B. mit Mikrotubuli (bestehend aus dem Strukturprotein Tubulin), um Organellen innerhalb der Zelle zu bewegen. *Transportproteine* wiederum, zu

Enzym	katalysierte Reaktion
Polymerasen	Katalysieren Polymerisationsreaktionen, wie z.B. die Synthese von DNA und RNA.
Kinasen	Katalysieren das Anfügen von Phosphatgruppen an Moleküle. Proteinkinasen fügen dementsprechend Phosphatgruppen an Proteine an.
Phosphatasen	Katalysieren die hydrolytische Entfernung einer Phosphatgruppe von einem Molekül.
Ligasen	Verbinden zwei Moleküle in einem Energie verbrauchenden Vorgang miteinander. Ubiquitin-Ligasen binden Ubiquitin[3] an ein Protein, wodurch dieses zum Abbau in einem Proteasom[4] markiert wird. Den Vorgang, der von einer Ubiquitin-Ligase katalysiert wird, nennt man Ubiquitinierung.

Tabelle 2.1.: Enzymarten und die zugehörigen katalysierten Reaktionen, die in dieser Arbeit relevant sind. Die Beschreibungen wurden größtenteils wörtlich aus [7] übernommen.

denen z.B. das im Blutkreislauf Sauerstoff transportierende Hämoglobin gehört, sind in der Lage, kleine Moleküle oder Ionen zu transportieren. *Signalproteine*, wie z.B. Hormone, übertragen extrazelluläre Signale von Zelle zu Zelle. So steuert das Hormon Insulin beispielsweise den Glucosespiegel im Blut. *Rezeptorproteine* erkennen wiederum solche Signale und leiten diese an den zelleigenen Reaktionsapparat weiter. Ein Beispiel hierfür ist der Insulinrezeptor, der es Zellen ermöglicht, auf Insulin zu reagieren, indem sie Glucose aufnehmen.
Zu den biologisch wichtigsten Proteinklassen, auf die wir hier näher eingehen wollen, gehören die sog. *Enzyme*. Enzyme sind Katalysatoren, die es der Zelle erlauben, kovalente Bindungen in kontrollierter Weise zu erzeugen oder zu spalten. Sie beschleunigen chemische Reaktionen oft um einen Faktor von 10^6 oder mehr und spielen daher für zahlreiche biologische Vorgänge, wie z.B. den Zellmetabolismus, eine entscheidende Rolle. Einige typische Enzymarten, die uns in dieser Arbeit in Kapitel 4 begegnen werden, und die von ihnen katalysierten Reaktionen sind in Tabelle 2.1 dargestellt.
Eine weitere wichtige Klasse von Proteinen stellen die sog. *Genregulatorproteine* dar, die in der Lage sind, an spezifische DNA-Sequenzen zu binden und dadurch die Expression eines Gens positiv oder negativ zu beeinflussen. Mit solchen Genregulatorproteinen und dem Prozess der Genregulation im Allgemeinen wollen wir uns im nächsten Abschnitt befassen.

[3] Kleines Protein, das kovalent an andere zelluläre Proteine gebunden wird [135].
[4] Großer Proteinkomplex im Cytoplasma, der an andere, zuvor mit Ubiquitin markierte Proteine bindet und diese spaltet [135].

2.1.5. Regulation der Genexpression und Genregulationsnetzwerke

Obwohl im Allgemeinen alle Zellen eines mehrzelligen Organismus dieselbe genetische Information in Form ihrer DNA enthalten, exprimiert jede Zelle nur einen individuellen Teil ihrer Gene. Das heißt, je nach Zelltyp und vorliegenden Umweltbedingungen, auf die Zellen dynamisch reagieren können, werden von den verschiedenen Zellen eines Organismus unterschiedliche RNA- und Proteinmoleküle synthetisiert. Die Kontrolle bzw. Regulation der Genexpression kann dabei auf verschiedenen Ebenen erfolgen, die in Abbildung 2.3 zusammengefasst sind.
Den wichtigsten Mechanismus im Rahmen der Genregulation stellt im Allgemeinen die Regulation der Transkription bzw. der Transkriptionsinitiation dar, da nur in diesem frühen Stadium der Genexpression vermieden werden kann, dass die Zelle unnötige Zwischenprodukte synthetisiert. Reguliert wird die Transkription eukaryotischer Gene dabei vor allem durch sog. spezifische *Transkriptionsfaktoren*, die zur Gruppe der Genregulatorproteine gehören. Diese sind in der Lage, die Expression bzw. Transkription eines Gens zu aktivieren (sog. *Aktivatoren*) oder zu inhibieren (sog. *Inhibitoren* oder *Repressoren*), indem sie an bestimmte DNA-Abschnitte (Promoter des Gens und sog. *cis-Elemente*), die wir in dieser Arbeit analog zur Bezeichnung in [7] als *Genkontrollregionen* bezeichnen wollen, binden.

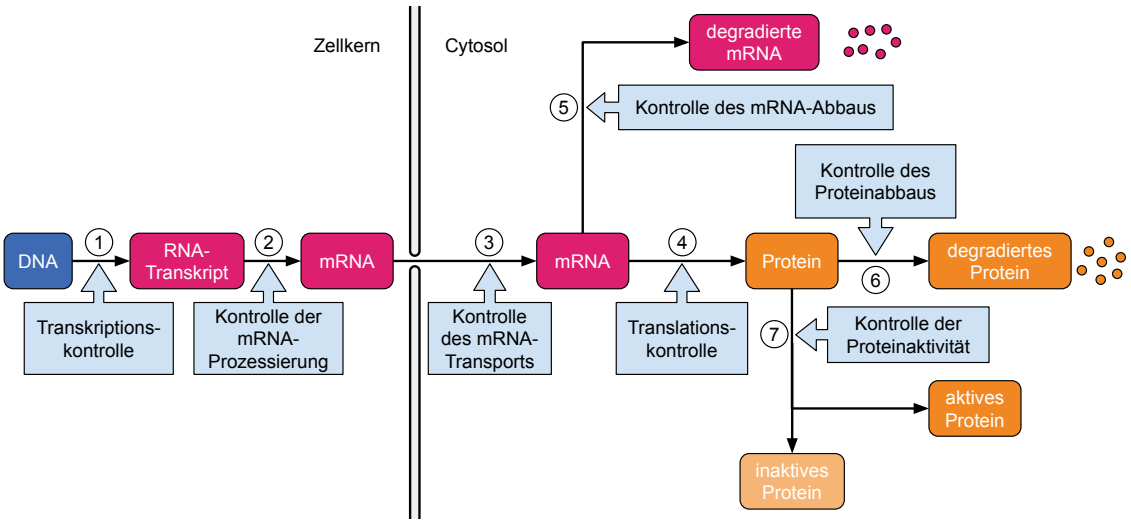


Abbildung 2.3.: Möglichkeiten einer (eukaryotischen) Zelle zur Kontrolle der Genexpression nach [6].
Eine Zelle kann zunächst im Rahmen der Transkriptionskontrolle (1) regulieren, wann und wie oft ein bestimmtes Gen transkribiert wird. Zudem kann sie kontrollieren, auf welche Art und Weise das primäre RNA-Transkript prozessiert wird (2) und auswählen welche fertiggestellten mRNAs vom Zellkern ins Cytoplasma exportiert und wohin genau sie transportiert werden (3). Im Rahmen der Translationskontrolle (4) kann die Zelle bestimmen, welche RNA-Moleküle im Cytoplasma durch Ribosomen in ein Protein übersetzt werden sollen. Anschließend kann sie sowohl den mRNA-Abbau (5) als auch den Proteinabbau (6) regulieren. Zudem kann eine Zelle spezifische Proteinmoleküle nach ihrer Synthese auf verschiedene Weisen (wie z.B. durch eine Proteinphosphorylierung) aktivieren oder deaktivieren (7).

Die Wirkung von Repressoren beruht dabei im Allgemeinen darauf, dass sie die Bindung der RNA-Polymerase an den Promoter blockieren, während Aktivatoren die Bindung der RNA-Polymerase an den Promoter erleichtern und somit die Transkriptionsrate erhöhen.

Wie wir gesehen haben, beeinflussen also Proteine, beispielsweise in Form von Transkriptionsfaktoren, die Expression von Genen. Da Proteine wiederum selbst von Genen kodiert werden, bedeutet das letztendlich, dass Gene ihre Expression gegenseitig regulieren. Man spricht in diesem Fall daher von einem *Genregulationsnetzwerk* [40]. In einem solchen Genregulationsnetzwerk repräsentieren die Knoten die beteiligten Gene, und eine gerichtete Verbindung von einem Gen i zu einem Gen j bedeutet, dass Gen i einen regulatorischen Einfluss auf die Expression von Gen j besitzt [24].

2.1.6. Der eukaryotische Zellzyklus

Damit eine Zelle sich vermehren kann, muss sie eine bestimmte Abfolge von Vorgängen durchlaufen, die als *Zellzyklus* bezeichnet wird. Der eukaryotische Zellzyklus besteht in der Regel aus vier verschiedenen Phasen (siehe Abbildung 2.4). Die zwei bedeutendsten Phasen sind dabei die *S-Phase* („S“ steht für Synthese), in der die nukleare DNA repliziert wird, und die *M-Phase*, in der sich zunächst der Zellkern teilt (sog. *Mitose*) und anschließend das Cytoplasma (sog. *Cytokinese*). Zwischen diesen Phasen gibt es zwei weitere sogenannte *Gap-Phasen* (engl. Gap = Lücke), die *G1-Phase* zwischen M- und S-Phase und die *G2-Phase* zwischen S- und M-Phase. Diese geben der Zelle zum einen mehr Zeit zum Wachsen, und zum anderen können während dieser Phasen verschiedene intra- und extrazelluläre Signale den Zellzyklusverlauf beeinflussen.

In eukaryotischen Zellen wird der Zellzyklus durch ein komplexes Proteinnetzwerk, dem sog. *Zellzyklus-Kontrollsystem*, überwacht und gesteuert. Dieses System stellt sicher, dass alle Vorgänge zum passenden Zeitpunkt, in der richtigen Reihenfolge und nur genau einmal je Zellzyklus ablaufen. Das Kontrollsystem kann hierzu den Zellzyklus an bestimmten Kontrollpunkten, den sog. *Checkpoints*, anhalten, falls z.B. vorhergehende Vorgänge nicht erfolgreich beendet wurden oder die Umweltbedingungen für den Eintritt in die S- oder M-Phase ungeeignet sind.

Eines der wichtigsten Bestandteile des Zellzyklus-Kontrollsystems sind die *Cyclin-abhängigen Kinasen (*Englisch*: cyclin-dependent kinases, Cdks)*. Diese werden im Laufe des Zellzyklus periodisch aktiviert und phosphorylieren dann andere intrazelluläre Proteine, die für das Anschalten oder die Steuerung wesentlicher Zellzyklusvorgänge verantwortlich sind. Die Aktivierung der Cdks erfolgt durch eine weitere für den Zellzyklusverlauf entscheidende Proteinklasse, den sog. *Cyclinen*. Cycline aktivieren Cdks indem sie mit diesen einen Komplex bilden. Das heißt, nur wenn Cdks an Cycline gebunden sind, sind sie als Proteinkinasen aktiv. Da die Konzentration der Cycline im Laufe des Zellzyklus zyklisch zu- und wieder abnimmt, während die Konzentration der Cdks relativ konstant bleibt, werden die Cyclin-Cdk-Komplexe dann vermehrt zusammengebaut, wenn eine hohe Konzentration des jeweiligen Cyclins vorliegt.

Neben der Regulierung durch die im Zellzyklusverlauf schwankenden Cyclinkonzentrationen, kann die Aktivität der Cyclin-Cdk-Komplexe aber auch auf andere Weisen beeinflusst werden. Hierzu gehört z.B. die Phosphorylierung bestimmter Stellen der Cdk-Untereinheit, welche die Cdk-Aktivität hemmen kann. Ein weiteres Beispiel sind sog. *Cdk-Inhibitor-Proteine*, die die Aktivität von Cyclin-Cdk-Komplexen inhibieren, indem sie an diese binden.

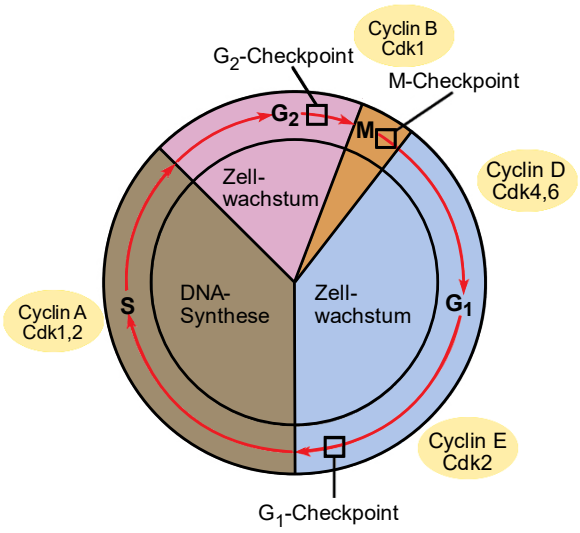


Abbildung 2.4.: Schematische Darstellung des Zellzyklus nach [21].
Während in der S-Phase die DNA repliziert wird, findet in der M-Phase die eigentliche Zellteilung statt. Zwischen diesen beiden Phasen gibt es zwei weitere sogenannte Gap-Phasen, G1 und G2, die der Zelle u.a. Zeit zum Wachsen geben. Überwacht wird das korrekte Durchlaufen des Zellzyklus durch das sog. Zellzyklus-Kontrollsystem, bei dem es sich um ein komplexes Netzwerk regulierender Proteine handelt, dessen zentrale Bestandteile Cyclin-abhängigen Kinasen (*Englisch*: cyclin-dependent kinases, Cdks) sind. Die Aktivität der Cdks hängt dabei von der Bindung regulierender Untereinheiten, der sogenannten Cycline, ab. Je nachdem, welche Cyclin-Cdk-Komplexe gerade aktiv sind, werden unterschiedliche Zellzyklusprozesse ausgelöst. Die für die jeweiligen Phasen im Zellzyklus entscheidenden Cyclin-Cdk-Komplexe sind in gelb neben dem Zellzyklus dargestellt. So werden beispielsweise aktive CyclinB-Cdk1-Komplexe zum Eintritt der Zelle in die M-Phase benötigt. Sind die Voraussetzungen für bestimmte Zellzyklusvorgänge (noch) nicht gegeben, kann das Zellzyklus-Kontrollsystem das Fortschreiten des Zellzyklus an bestimmten Kontrollpunkten, den sog. Checkpoints (an den Übergängen von der G1- zur S-Phase und von der G2- zur M-Phase sowie während der Mitose), stoppen.

2.2. Mathematische Modellierung biologischer Netzwerke

Biologische Netzwerke lassen sich auf vielfältige Art und Weise modellieren (siehe z.B. [40, 79] zur Modellierung von Genregulationsnetzwerken oder [51, 93] zur Modellierung mikrobieller Interaktionsnetzwerke), wobei wir uns in dieser Arbeit ausschließlich auf *deterministische Modelle* beschränken wollen. Diese sind immer dann eine gute Beschreibung der realen biologischen Prozesse, wenn die Anzahl oder die Konzentration der betrachteten Netzwerkspezies so groß ist, dass stochastische Effekte nur eine untergeordnete Rolle spielen. Deterministische Modelle lassen sich gemäß dem Detaillierungsgrad der Darstellung der Knotenzustände im Wesentlichen in zwei Gruppen unterteilen: in kontinuierliche und in diskrete (deterministische) Modelle.

Zu den beliebtesten und am häufigsten verwendeten Modellen im Bereich der Systembiologie gehören die *kontinuierlichen deterministischen Modelle*. Bei diesen Modellen repräsentieren die Knoten, deren Zustände kontinuierliche Werte annehmen können, in der Regel Konzentrationen, und ihre zeitliche Entwicklung wird mittels Differentialgleichungen modelliert. Der Vorteil von kontinuierlichen deterministischen Modellen, insbesondere in Form von gewöhnlichen Differentialgleichungen, ist, dass sie sich mathematisch sehr gut handhaben lassen. Sie erlauben theoretische Methoden wie die lineare Stabilitätsanalyse, mithilfe derer sich die Stabilität von Fixpunkten einfach bestimmen lässt, oder die Bifurkationsanalyse, im Rahmen derer betrachtet wird, wie sich die Dynamik des Systems in Abhängigkeit verschiedener Parameter (insbesondere qualitativ) verändert. Eine detaillierte Beschreibung dieser Methoden kann beispielsweise [149] entnommen werden.

Diskrete deterministische Modelle weisen im Vergleich zu kontinuierlichen (deterministischen) Modellen einen deutlich höheren Abstraktionsgrad auf, da die Netzwerkknoten in diesen Modellen nur wenige diskrete Zustände annehmen können.
Der Vorteil hiervon ist, dass zur Erstellung dieser Modelle verhältnismäßig wenige Informationen über die Modellparameter vorliegen müssen und sie in Fällen erstellt werden können, in denen kontinuierliche Modelle aufgrund der großen Anzahl von unbekannten Parametern ungeeignet sind. Der Nachteil ist, dass man mithilfe dieser Modelle im Wesentlichen nur qualitative Aussagen treffen kann [79], während kontinuierliche Modelle häufig auch quantitative Vorhersagen ermöglichen.
Zu den bekanntesten und am häufigsten verwendeten Modellen dieser Klasse gehören die *Booleschen Netzwerke*, bei denen jeder Knoten nur zwei Zustände annehmen kann. Boolesche Netzwerke sind insbesondere dann gut zur Modellierung eines biologischen Systems geeignet, wenn „An“- bzw. „Aus“-Zustände deutlich ausgeprägt sind. Dies ist beispielsweise im Kontext von Genregulationsnetzwerken der Fall, wo ein Gen in der Regel entweder exprimiert wird oder nicht. Zudem wird im Rahmen einer Booleschen Modellierung gewöhnlicherweise auch die Zeit diskret modelliert.

2.2.1. Kontinuierliche Modellierung

Im Rahmen kontinuierlicher deterministischer Modelle wird das zeitliche Verhalten eines biologischen Systems typischerweise mithilfe von gewöhnlichen Differentialgleichungen der Form

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i\left(x_1, \ldots, x_n\right), \quad i \in 1, \ldots, n, \tag{2.1}$$

bzw. in Vektorschreibweise

$$\frac{\mathrm{d}\vec{x}}{\mathrm{d}t} = \vec{f}(\vec{x}), \tag{2.2}$$

beschrieben. Die Variablen $x_1, \ldots, x_n$ stellen dabei in der Regel die Konzentrationen der modellierten Spezies (z.B. Proteine) dar und die Funktionen f_i geben deren zeitliche Entwicklung an.
Da die Funktionen f_i typischerweise nichtlinear in den Variablen x_i sind, ist es nur in seltenen Fällen möglich (und dann meist sehr schwierig) eine analytische Lösung für das Differentialgleichungssystem (2.1) zu finden [149]. Die Dynamik nichtlinearer Systeme lässt sich daher oft nur schwer vorhersagen und kann sehr komplex sein [46]. Zur Untersuchung solcher komplexen dynamischen Systeme verwendet man aus diesem Grund zum einen analytische Methoden, wie z.B. die lineare Stabilitätsanalyse von Fixpunkten oder die Bifurkationsanalyse, und zum anderen numerische Methoden bzw. Computersimulationen.

Massenwirkungsgesetz

Um für ein zelluläres biologisches System (wie z.B. ein Genregulationsnetzwerk) Differentialgleichungen der Form (2.1) aufzustellen, ist es häufig sinnvoll, zunächst die chemischen Reaktionen zu betrachten, die für das System relevant sind. Ausgehend von diesen chemischen Reaktionen bzw. den Reaktionsgleichungen lässt sich dann mithilfe des *Massenwirkungsgesetzes* ein Differentialgleichungssystem aufstellen. Das Massenwirkungsgesetz besagt dabei, dass die Geschwindigkeit einer Reaktion proportional zur Wahrscheinlichkeit eines Zusammenstoßes der Reaktionspartner ist [87]. Diese Wahrscheinlichkeit ist wiederum proportional zur Konzentration der Reaktanden potenziert mit ihrer Molekularität, d.h. mit der Anzahl, mit der sie in die jeweilige Reaktion eingehen [87].
Liegt also ein chemisches Reaktionssystem bestehend aus n verschiedenen Spezies x_i und N Reaktionen der Form

$$\sum_{i=1}^{n} \nu_{ij} x_i \xrightarrow{k_j} \sum_{i=1}^{n} \mu_{ij} x_i \quad \text{mit } j = 1, \ldots, N$$

vor, so gilt nach dem Massenwirkungsgesetz für die zeitliche Änderung der Konzentrationen der Spezies x_i [96, 45]:

$$\frac{\mathrm{d}\left[x_i\right]}{\mathrm{d}t} = \sum_{j=1}^{N} \left(\mu_{ij} - \nu_{ij}\right) k_j \prod_{l=1}^{n} \left[x_l\right]^{\nu_{lj}} . \tag{2.3}$$

Hierbei sind ν_{ij} und μ_{ij} die sog. stöchiometrischen Koeffizienten und k_j die Reaktionskonstanten. $[x_i]$ steht für die Konzentration von x_i.

Modellierung der transkriptionellen Genregulation: Hillfunktionen

In diesem Abschnitt soll nun die kontinuierliche Modellierung eines biologischen Prozesses mittels gewöhnlicher Differentialgleichungen am konkreten Beispiel der Genexpression erläutert werden. Die Ausführungen in diesem Abschnitt beruhen dabei weitestgehend auf den Quellen [2, 62].

Die Regulation während der Transkription eines Gens kann durch die nachfolgende Reaktion[5] beschrieben werden [2]

$$D + T \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} [DT] \xrightarrow{k_{tr}} D + T + mRNA.$$

Diese Reaktionsgleichung umfasst die Bindung eines Transkriptionsfaktors T an einen die Expression des Gens kontrollierenden DNA-Abschnitt D (z.B. Promoter oder Enhancer), wodurch ein Komplex $[DT]$ aus Genkontrollregion und Transkriptionsfaktor gebildet wird. Die Bildung dieses Komplexes führt dann zur Synthese eines mRNA-Moleküls ($mRNA$), wobei die Genkontrollregion und der Transkriptionsfaktor in der Zelle erhalten bleiben.

Die Dynamik der verschiedenen Spezies kann gemäß des Massenwirkungsgesetzes (vgl. Gl. (2.3)) durch die folgenden Gleichungen beschrieben werden:

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \frac{\mathrm{d}T}{\mathrm{d}t} = -k_1 \cdot D \cdot T + (k_{-1} + k_{tr})\,[DT] \tag{2.4}$$

$$\frac{\mathrm{d}\,[DT]}{\mathrm{d}t} = k_1 \cdot D \cdot T - (k_{-1} + k_{tr})\,[DT] \tag{2.5}$$

$$\frac{\mathrm{d}mRNA}{\mathrm{d}t} = k_{tr}\,[DT]\,, \tag{2.6}$$

wobei D, T, $[DT]$ und $mRNA$ jetzt (und in den nachfolgenden Gleichungen) für die Konzentrationen der jeweiligen Spezies stehen.

Da eine analytische Lösung dieses Gleichungssystems nicht möglich ist, müssen zusätzliche Annahmen getroffen werden, um das System auf eine geeignete Weise zu vereinfachen [87]. Eine sinnvolle Annahme in diesem Kontext ist, dass der Transkriptionsfaktor sehr schnell an die Genkontrollregion bindet und somit der Komplex $[DT]$ sehr viel schneller seinen Gleichgewichtszustand $\frac{\mathrm{d}[DT]}{\mathrm{d}t} = 0$ erreicht als die anderen Spezies.
Unter der Annahme, dass für den Komplex $[DT]$ also ein quasistationärer Zustand vorliegt, d.h. $\frac{\mathrm{d}[DT]}{\mathrm{d}t} = 0$ gilt, und unter Berücksichtigung der Tatsache, dass die Gesamtzahl D_{ges} der Genkontrollregionen konstant ist, d.h. $D + [DT] = D_{ges}$ gilt, ergibt sich aus Gleichung (2.5) für die Konzentration des $[DT]$-Komplexes:

$$[DT] = \frac{k_1 \cdot T \cdot D_{ges}}{k_{-1} + k_{tr} + k_1 T} = \frac{D_{ges} \cdot T}{\theta + T} \qquad \text{mit } \theta = \frac{k_{-1} + k_{tr}}{k_1}.$$

[5] Die hier dargestellte Gleichung entspricht im Wesentlichen der bekannten Reaktionsgleichung

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} [ES] \xrightarrow{k_{tr}} E + P$$

für die enzymatische Katalyse, bei der ein Enzym E ein Substrat S unter Bildung eines Enzym-Substrat-Komplexes $[ES]$ in ein Produkt P umwandelt. In unserem Fall wäre das Substrat S aber ebenfalls ein Produkt der Reaktion.

Setzt man nun dieses Ergebnis in Gleichung (2.6) für die Dynamik der mRNA-Konzentration ein, so ergibt sich:

$$\frac{\mathrm{d}mRNA}{\mathrm{d}t} = \frac{k_{tr} \cdot D_{ges} \cdot T}{\theta + T} = \frac{m \cdot T}{\theta + T} \tag{2.7}$$

mit $m = k_{tr} D_{ges}$. Diese Gleichung wird im Kontext der Enzymkinetik als *Michaelis-Menten-Gleichung* bezeichnet.

Sind mehrere Einheiten des Transkriptionsfaktors T erforderlich, um die Transkription zu aktivieren, spricht man in der Biologie von sog. *Kooperativität*. Nimmt man nun an, dass die Bildung der Komplexe aus n Transkriptionsfaktoren, die entweder vor oder nach der Bindung an die Genkontrollregion stattfinden kann, viel schneller abläuft als andere Reaktionen im Transkriptionsprozess, erhält man aus der Reaktionsgleichung

$$D + nT \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} [DT_n] \xrightarrow{k_{tr}} D + nT + mRNA$$

analog zur Herleitung der Michaelis-Menten-Gleichung den folgenden Ausdruck für die mRNA-Produktionsrate:

$$\frac{\mathrm{d}mRNA}{\mathrm{d}t} = \frac{m \cdot T^n}{\theta^n + T^n} \quad \text{mit } \theta = \left(\frac{k_{-1} + k_{tr}}{k_1}\right)^{\frac{1}{n}} \tag{2.8}$$

Hierbei handelt es sich um eine sog. *Hill-Funktion* mit Hill-Koeffizientem n, welcher den Grad der Kooperativität des Transkriptionsprozesses angibt. Der Parameter m stellt die maximale Transkriptionsrate dar und der Parameter θ, der auch als Aktivierungsschwelle bezeichnet wird, gibt die Konzentration des Transkriptionsfaktors an, die notwendig ist, um die Transkription des Gens signifikant zu aktivieren.

Abbildung 2.5 zeigt die durch Gleichung (2.8) beschriebene Hillfunktion für unterschiedliche Werte des Hill-Koeffizienten n. Wie man sieht, wird der Verlauf der Hill-Funktion mit zunehmendem Koeffizienten n immer steiler und nähert sich für $n \to \infty$ einer Stufenfunktion mit Sprung bei der Aktivierungsschwelle θ. Dieser Grenzfall stellt insbesondere im Kontext der Genregulation die Motivation für eine sog. *Boolesche Modellierung* dar, bei der die Variablen anstatt kontinuierlicher Werte nur zwei diskrete Werte, nämlich 0 („aus“) und 1 („an“), annehmen können. Mit diesen abstrakten Booleschen Modellen wollen wir uns im nächsten Abschnitt beschäftigen.

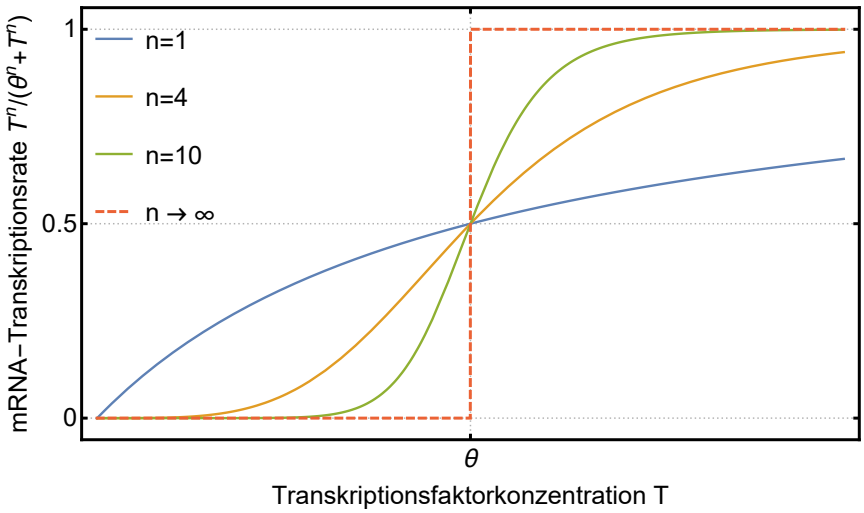


Abbildung 2.5.: Hill-Funktion (eines Aktivators) zur Beschreibung der $mRNA$-Transkriptionsrate in Abhängigkeit der Transkriptionsfaktorkonzentration T für verschiedene Werte des Hill-Koeffizienten n.

2.2.2. Boolesche Modellierung

Die hier dargestellten Grundlagen zur Booleschen Modellierung beruhen, wenn nicht gesondert angegeben, auf den Quellen [47, 152, 115] und wurden in großen Teilen nahezu wörtlich der Masterarbeit [24] der Autorin entnommen. Das gezeigte Beispielnetzwerk (Abb. 2.6) und sein Zustandsraum (Abb. 2.7) hingegen wurden speziell für diese Arbeit konzipiert.

Im Wesentlichen handelt es sich bei einem *Booleschen Netzwerk* um einen (meist) gerichteten Graphen mit N Knoten, von denen jeder einen Booleschen Wert s_i und eine Boolesche Aktualisierungsfunktion f_i zugewiesen bekommt.
Größere Bekanntheit erlangten Boolesche Netzwerke nachdem sie 1969 von Stuart A. Kauffman [82] zur Modellierung von Genregulationsnetzwerken verwendet wurden. In dem von Kauffman aufgestellten Modell repräsentieren die Knoten dementsprechend Gene, wobei jeder Knoten i zwei verschiedene mögliche Werte s_i annehmen kann: Entweder ein Gen wird exprimiert, dann ist $s_i = 1$, oder es wird nicht exprimiert, dann gilt $s_i = 0$. Eine gerichtete Verbindungen zwischen zwei Knoten stellt in Kauffmans Modell den regulatorischen Einfluss eines Gens auf die Expression bzw. Transkription des anderen dar. Zudem erhält jeder Knoten eine Boolesche Aktualisierungsfunktion f_i. Diese bestimmt den Zustand s_i von Knoten i im nächsten Zeitschritt in Abhängigkeit der Werte s_{i_1} bis $s_{i_{K_i}}$ seiner K_i Eingangsknoten, d.h.

$$s_i(t+1) = f_i\left[s_{i_1}(t), s_{i_2}(t), \ldots, s_{i_{K_i}}(t)\right]. \tag{2.9}$$

In Kauffmanns Modell wird die Unkenntnis über die genaue Art der Beeinflussung der Gene untereinander dadurch berücksichtigt, dass die Verknüpfungen und Aktualisierungsregeln der Knoten vollkommen zufällig gewählt werden, weshalb man auch von einem Zufallsmodell oder zufälligen Booleschen Netzwerken spricht.

Die Aktualisierung aller N Knoten erfolgt im klassischen Booleschen Zufallsmodell zudem synchron, sodass alle Knoten gleichzeitig gemäß ihrer Aktualisierungsfunktion f_i aktualisiert werden und die Dynamik somit deterministisch ist.
Im Folgenden beziehen sich alle Erläuterungen auf dieser Art der parallelen bzw. deterministischen Aktualisierung.

Aktualisierungsfunktionen

Wie bereits im vorherigen Abschnitt beschrieben, wird die Dynamik jedes einzelnen Knotens und somit die des gesamten Netzwerks durch die Booleschen Aktualisierungsfunktionen bestimmt.
Eine solche Boolesche Aktualisierungsfunktion, oder allgemeiner eine n-stellige Boolesche Funktion f, ist eine Funktion $f : \{0,1\}^n \to \{0,1\}$, die jeder möglichen Kombination von Booleschen Eingangswerten einen Booleschen Ausgangswert zuordnet [65]. Besitzt ein Knoten K Inputwerte, gibt es für ihn also genau 2^{2^K} verschiedene mögliche Boolesche Aktualisierungsfunktionen, die jeweils die Länge 2^K haben .
Typischerweise werden Boolesche Funktionen als eine Zeichenkette von Nullen und Einsen dargestellt. Man erhält diese Zeichenkette leicht mithilfe einer Wertetabelle, in der für unterschiedliche Kombinationen von Eingangswerten der Ausgangswert angegeben ist (vgl. Abbildung 2.6).

In dieser Arbeit wird in Kapitel 3.4 nur eine Unterklasse aller möglichen Booleschen Aktualisierungsfunktionen, jene der sogenannten *Schwellenwertfunktionen*, verwendet. Durch Schwellenwertfunktionen kann z.B. im Kontext der Genregulation die Tatsache berücksichtigt werden, dass die Transkription von Genen durch die Produkte anderer Gene aktiviert oder inhibiert bzw. positiv oder negativ reguliert werden kann, wobei nicht jede Änderung des Expressionslevels eines Gens einen unmittelbaren Effekt auf die Expression eines anderen Gens hat, sondern nur die zusammengenommene Aktivität aller Aktivatoren und Inhibitoren.

Die in dieser Arbeit verwendeten Schwellenwertfunktionen zur Aktualisierung der Knoten wurden bereits vielfach genutzt (beispielsweise zur erfolgreichen Modellierung der chronologischen Abläufe während des Zellzyklus der Bäckerhefe [100] oder der Spalthefe *Schizosaccharomyces pombe* [39]) und sind von der Form

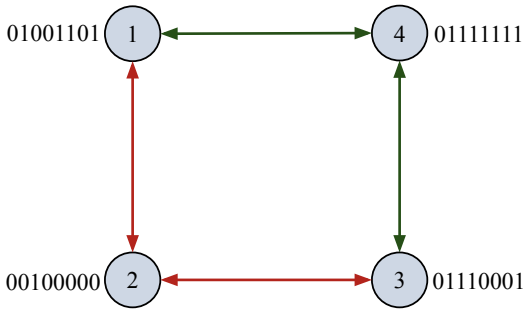
$$s_i(t+1) = \begin{cases} 1, & \sum_{j=1}^{N} G_{ij} s_j(t) > 0 \\ s_i(t), & \sum_{j=1}^{N} G_{ij} s_j(t) = 0 \\ 0, & \sum_{j=1}^{N} G_{ij} s_j(t) < 0 \end{cases} . \tag{2.10}$$

Hierbei ist G die Adjazenzmatrix des betrachteten Netzwerkgraphen und die Einträge G_{ij} dieser Matrix entsprechen Kantengewichten, wobei gilt

$$G_{ij} = \begin{cases} +1, & \text{für eine positive Interaktion zwischen } i \text{ und } j \\ -1, & \text{für eine negative Interaktion zwischen } i \text{ und } j \\ 0, & \text{wenn es keine Interaktion zwischen Knoten } i \text{ und } j \text{ gibt.} \end{cases}$$

Ein Knoten wird somit eingeschaltet ($s_i(t+1) = 1$), wenn die Summe seiner Eingänge über dem Schwellenwert liegt und ausgeschaltet ($s_i(t+1) = 0$), wenn die Summe seiner Eingänge unterhalb des Schwellenwertes liegt. Ergibt die Summe der Eingänge genau den Schwellenwert, so ändert der Knoten seinen Zustand im nächsten Zeitschritt nicht ($s_i(t+1) = s_i(t)$).

Abbildung 2.6 zeigt ein beispielhaftes Boolesches Schwellenwertnetzwerk der Netzwerkgröße $N = 4$. Rechts neben dem Netzwerk ist für Knoten 1 die zugehörige Aktualisierungsregel f_1, die sich gemäß Gleichung (2.10) ergibt, in Form einer Wertetabelle gezeigt.



σ_1	σ_2	σ_4	f_1
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	1
1	0	1	1
1	1	0	0
1	1	1	1

Abbildung 2.6.: Boolesches Netzwerk mit $N = 4$ Knoten, von denen jeder $K = 2$ Eingänge besitzt. Die Zahlen in den Knoten geben ihre Indizes an und die Zeichenketten neben den Knoten stellen ihre jeweilige Boolesche Funktion dar. Grüne Kanten repräsentieren eine positive Interaktion ($G_{ij} = 1$) und rote Kanten eine negative Interaktion ($G_{ij} = -1$) zwischen den jeweiligen Knoten. Rechts neben dem Netzwerk ist für Knoten 1 beispielhaft seine Aktualisierungsfunktion f_1 in Form einer Wertetabelle gezeigt. Links in der Tabelle stehen die verschiedenen möglichen Kombinationen der Eingangswerte des Knotens und rechts der zugehörige Ausgangswert. Da gemäß Gleichung (2.10) der Zustand jedes Knotens im nächsten Zeitschritt auch von seinem eigenen aktuellen Wert abhängt, hat jede Boolesche Aktualisierungsfunktion eine Länge von $2^3 = 8$.

Zustandsraum

Da in einem Booleschen Netzwerk mit N Knoten jeder Knoten i genau zwei verschiedene Werte s_i ($s_i = 1$ oder $s_i = 0$) annehmen kann, gibt es insgesamt 2^N unterschiedliche *Netzwerkzustände*, die in Form eines Vektors

$$\vec{s}(t) = \{s_1(t), s_2(t), \ldots, s_N(t)\}$$

dargestellt werden können. Alle Netzwerkzustände sind dabei durch ihre zeitliche Abfolge fest miteinander verbunden und bilden den *Zustandsraum*. Aufgrund der Endlichkeit dieses Zustandsraums und der deterministischen Aktualisierung der Zustände (vgl. Gl. 2.9) ist es unvermeidlich, dass sich ein Zustand nach spätestens 2^N Zeitschritten wiederholt und die Trajektorien im Phasenraum irgendwann periodisch werden. Eine solche Folge sich wiederholender Zustände, auf die ein oder mehrere andere Zustände führen, wird dabei als *Attraktor* bezeichnet. Die Zustandsabfolgen, die zu einem Attraktor hinführen, nennt man *Transienten*. Als Transientenlänge bezeichnet man die Anzahl der transienten Zustände in einer Trajektorie. Als Länge des Attraktors definiert man die Anzahl der Zustände, die Teil dieses Attraktors sind. Die Transienten bilden zusammen mit den Attaktorzuständen das sogenannte Einzugsgebiet bzw. *Bassin* des Attraktors. Die Größe des Bassins eines Attraktors ist die Anzahl der Zustände, die Teil dieses Bassins sind.

In Abbildung 2.7 ist der Zustandsraum des Netzwerkes aus Abbildung 2.6 gezeigt. Die Attraktorzustände sind in rot und die transienten Zustände in blau dargestellt. Wie man sieht, gibt es vier Attraktoren, die alle die Länge 1 besitzen, d.h. Fixpunkte sind. Während drei der Attraktoren nur ein kleines Bassin besitzen, welches lediglich aus dem jeweiligen Fixpunkt besteht (Bassingröße 1), hat der vierte Attraktor mit einer Bassingröße von 13 Zuständen ein sehr großes Einzugsgebiet.

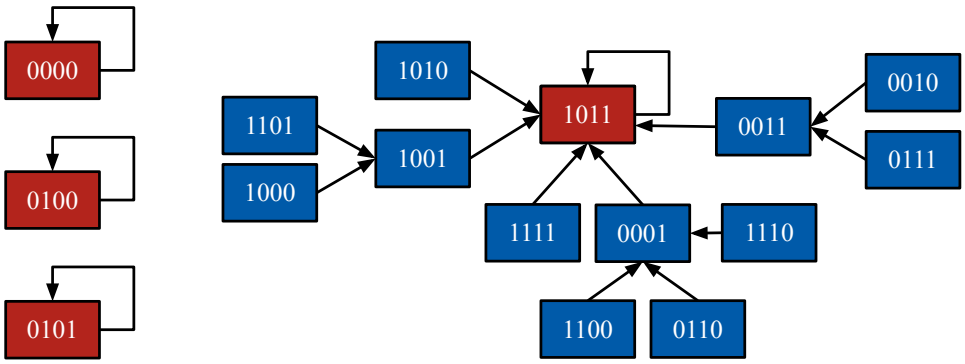


Abbildung 2.7.: Zustandraum des Netzwerks aus Abbildung 2.6. Die Attraktorzustände sind in rot und die transienten Zustände in blau dargestellt.

3. Inferenz mikrobieller Interaktionsnetzwerke mithilfe der ESABO-Methode

Dieses Kapitel der Dissertation basiert zu großen Teilen auf [113]. Neben der Autorin haben auch Barbara Drossel und Marc-Thorsten Hütt an diesem Manuskript mitgewirkt.

3.1. Einleitung: mikrobielle Interaktionsnetzwerke und ihre Inferenz

Mikroorganismen wie beispielsweise Bakterien leben nicht alleine, sondern bilden komplexe Gemeinschaften [51]. Spezies, die Teil einer solchen mikrobiellen Gemeinschaft sind, können auf verschiedene Weisen miteinander interagieren. Manche Bakterien weisen z.B. eine mutualistische Interaktion auf, indem sie beispielsweise gemeinsam einen Biofilm bilden [136], während andere eine antagonistische Beziehung zueinander haben, da sie z.B. um Nährstoffe konkurrieren [74, 63].

Da Mikrobiome und die Interaktionen zwischen ihren Mitgliedern eine entscheidende Rolle für die Gesundheit des jeweiligen Wirts spielen, hat die Analyse mikrobieller Auftretenshäufigkeiten (sog. *Abundanzen*) in den letzten zehn Jahren große Beachtung gefunden [153]. Für die unterschiedlichsten mikrobiellen Gemeinschaften, von Bodenproben bis hin zur menschlichen Haut, dem Mund oder dem Darm, ist die auf DNA-Sequenzierung beruhende Abundanzschätzung mikrobieller Taxa zu einem weit verbreiteten Instrument geworden, um Informationen über das zugrunde liegende Ökosystem zu sammeln. Die immense medizinische Bedeutung und das damit verbundene klinische Potenzial der Mikrobiomanalyse wurden dabei im Laufe der Zeit immer deutlicher [28, 35, 48, 105].

Trotz dieser großen Relevanz fehlt noch immer ein tiefes theoretisches Verständnis der beobachteten mikrobiellen Abundanzmuster. Ein anfänglicher – und zugleich stark kritisierter – Versuch war die Hypothese von drei charakteristischen Darmmikrobiomtypen, den sogenannten *Enterotypen*, die sich durch die Dominanz unterschiedlicher Bakteriengattungen auszeichnen [10]. Allerdings ist unter anderem bereits die Definition eines solchen Mikrobiomzustands ein schwieriges Problem [61].

Ein wichtiger Schritt zum theoretischen Verständnis von mikrobiellen Gemeinschaften im Allgemeinen und insbesondere des menschlichen Mikrobioms besteht darin, die zugrunde liegenden mikrobiellen Interaktionsnetzwerke ausgehend von Abundanzmustern zu schätzen. Hierzu existiert bereits eine Vielzahl von Inferenzmethoden, welche meist auf kontinuierlichen Abundanzdaten beruhen und beispielsweise in Faust und Raes [51] oder in Matchado et al. [109] zusammengefasst sind.

Trotz all dieser Methoden, stellt die Inferenz mikrobieller Interaktionsnetzwerke aus Abundanzdaten jedoch nach wie vor eine schwierige Aufgabe dar, deren Herausforderungen unter anderem von Röttjers und Faust in [133] oder aktueller von Faust in [50] diskutiert werden. Zu diesen Herausforderungen gehört beispielsweise die Tatsache, dass bei der Analyse kontinuierlicher Abundanzdaten in der Regel zunächst eine Normierung durchgeführt werden muss, da für jede Probe eine unterschiedliche Menge an die Abundanz bestimmendem Material, beispielsweise an extrahierter oder sequenzierter DNA, vorliegt [51]. Solch eine Normierung (auf die Gesamtzahl der Counts in der Probe) führt jedoch dazu, dass keine Absolutwerte, sondern relative Häufigkeiten betrachtet werden, die sich stets zu Eins aufsummieren müssen und somit nicht unabhängig sind. Für solche Daten, die auch als *Kompositionsdaten* (*compositional data*) [4, 53] bezeichnet werden, sind die Standardmethoden zur Berechnung von Korrelationen nicht mehr gültig und die Anwendung dieser kann zu falschen bzw. verzerrten Ergebnissen führen [57, 51]. So neigen die normierten Abundanzdaten beispielsweise dazu, unabhängig von der tatsächlichen Korrelation zwischen den zugrunde liegenden absoluten Häufigkeiten eine negative Korrelation aufzuweisen [57]. Ein weiteres Problem ist zudem der hohe Prozentsatz von Nullen in den Abundanzdaten (sog. *data sparsity*), da eine Null zwei unterschiedliche Bedeutungen haben kann [51]: (1) Die Spezies ist in der Probe tatsächlich nicht vorhanden oder (2) die Häufigkeit der Spezies liegt lediglich unterhalb der Nachweisgrenze. Dies ist insbesondere bei der Untersuchung selten vorkommender Mikroorganismen problematisch, da sie oft gänzlich aus der Analyse ausgeschlossen werden müssen, weil sie in zu vielen Proben fehlen [51]. Eine weitere Herausforderung besteht schließlich darin, zwischen direkten (es ist tatsächlich eine Netzwerkkante zwischen Spezies A und B vorhanden) und indirekten Verbindungen (Spezies A und B sind lediglich über eine dritte Spezies C miteinander verbunden, d.h. es gibt lediglich eine Kante zwischen A und C sowie B und C, aber nicht zwischen A und B) zu unterscheiden [109].
Methoden, die ausschließlich auf Korrelationen oder anderen einfachen statistischen Zusammenhängen basieren, können – wie bereits beschrieben– diesen Herausforderungen in der Regel nicht gerecht werden [23].

In dieser Arbeit wollen wir daher einen anderen Ansatz verfolgen und schlagen dazu zunächst vor, zwei Informationsebenen zu differenzieren, die in mikrobiellen Abundanzdaten enthalten sind: (1) die quantitativen Abundanzwerte und (2) das Muster des Vorhandenseins und der Abwesenheit mikrobieller Organismen, wobei wir uns im Folgenden mit Letzterem beschäftigen werden. Die Unterscheidung dieser Ebenen ist wichtig, denn es besteht ein erheblicher Unterschied in der Art des systemischen Einblicks, den kontinuierliche Abundanzmuster und binäre Muster von An- und Abwesenheiten bieten können. So ist beispielsweise in einem transkriptomischen Datensatz das absolute Expressionsniveau eines Gens oft ein Hinweis auf die Funktion des Genprodukts (typischerweise sind die Expressionsniveaus von Genen, die für Transkriptionsfaktoren kodieren, viel geringer als derer, die für Stoffwechselenzyme kodieren), während das „An“- und „Aus“-Muster der Genexpression hingegen häufig Rückschlüsse auf die zugrunde liegenden regulatorischen Netzwerke ermöglicht [101, 25]. Auch beim Mikrobiom ist es naheliegend, dass die exakte Abundanz einer Spezies für deren grundsätzliche Rolle im Netzwerk unwichtig ist und lediglich für verschiedene Rollen unterschiedliche Abundanzwerte typisch sind.

Die Betrachtung des Mikrobioms aus einer binären Perspektive hat den großen Vorteil, dass sie den Weg für neue mathematische Ansätze zur Untersuchung des Mikrobioms ebnet, indem wir Mikrobiomdaten, also das Vorhanden- bzw. Nichtvorhandensein der verschiedenen Spezies in einer Mikrobiomprobe, als Attraktoren, oder genauer gesagt als Fixpunkte, Boolescher Netzwerke interpretieren. Ausgehend von diesen Attraktoren ist es dann unser Ziel, ein Interaktionsnetzwerk zwischen den in den Mikrobiomproben vorhandenen Spezies zu inferieren. Um dies zu erreichen, führen wir eine neue Inferenzmethode ein, die auf der von Claussen et al. [34] eingeführten ESABO-Methode aufbaut und diese um einem evolutionären Algorithmus erweitert. Der Kerngedanke unseres Ansatzes ist dabei, dass das inferierte Netzwerk die ursprüngliche Menge der (beobachteten) binären Abundanzmustern als seine Attraktoren reproduzieren können muss.

Im Rahmen unserer Inferenzmethode, die wir als ESABO-gestützte Evolution bezeichnen wollen, verwenden wir daher das mit der ursprünglichen ESABO-Methode (bzw. einer verbesserten Version dieser) rekonstruierte Netzwerk lediglich als Ausgangspunkt für den Inferenzprozess und nutzen einen evolutionären Algorithmus, um den Überlapp zwischen den Attraktoren des inferierten Netzwerks und den ursprünglichen (beobachteten) binären Häufigkeitsmustern zu maximieren.

Im nachfolgen Kapitel 3.2 wird zunächst erläutert, wie mikrobielle Abundanzdaten, die den Ausgangspunkt für die Untersuchung von Mikrobiomen darstellen, experimentell mittels 16S-rRNA-Sequenzierung, eines der gängigsten Verfahren zur Ermittlung der taxonomischen Zusammensetzung von Mikrobiomen, gewonnen werden. Anschließend wird in Kapitel 3.3 die ursprüngliche ESABO-Methode beschrieben und in Kapitel 3.4 erklärt, wie diese anhand von simulierten Daten getestet werden kann. In Kapitel 3.5 werden dann zwei Verbesserungen vorgestellt, die ich an der ESABO-Methode vorgenommen habe, um ihre Geschwindigkeit sowie die Reproduzierbarkeit der Ergebnisse zu verbessern und die falsche Vorhersage von negativen Kanten im Falle positiver Interaktionen zu verhindern. Anschießend folgt in Kapitel 3.6 die Erläuterung der zentralen Erweiterung der ESABO-Methode um einen evolutionären Algorithmus zur ESABO-gestützten Evolution. Im ersten Teil der Ergebnisse (Kapitel 3.7.1) zeigen wir dann, dass die ESABO-gestützte Evolution zu rekonstruierten Netzwerken führt, die eine hohe (topologische) Ähnlichkeit mit den ursprünglichen Netzwerken aufweisen. Durch die nachfolgende Untersuchung der Methode in Situationen, in denen die Daten nicht alle Attraktoren umfassen, d.h. bei unvollständiger Kenntnis der Attraktoren (siehe Kapitel 3.7.2), finden wir eine Beziehung zwischen dem Prozentsatz der bekannten Attraktoren und der durchschnittlichen Genauigkeit, die durch unsere Inferenzmethode erreicht wird. Im letzten Ergebnisteil (Kapitel 3.7.3) wenden wir unsere Methode schließlich auf echte empirische Daten an. Abschließend folgt in Kapitel 3.8 ein Fazit sowie eine Diskussion der Ergebnisse.

3.2. Analyse des menschlichen Mikrobioms mittels 16S-rRNA-Gensequenzierung

In diesem Abschnitt wollen wir uns damit vertraut machen, wie die mikrobiellen Abundanzdaten, die den Ausgangspunkt für die Untersuchung eines Mikrobioms darstellen, prinzipiell gewonnen werden. Die Ausführungen in den Abschnitten 3.2.2 und 3.2.3 beruhen dabei, wenn nicht gesondert angegeben, auf Quelle [120].

3.2.1. Hochdurchsatzsequenzierung

Um das menschliche Mikrobiom, d.h. die Gesamtheit aller den Menschen bewohnenden Mikroorganismen (bzw. im engeren Sinn deren Genome), zu untersuchen, werden heutzutage in der Regel kultivierungsunabhängige Verfahren genutzt, die darauf beruhen, dass aus Proben extrahierte DNA analysiert bzw. sequenziert wird. Hierzu nutzt man (seit 2005) Methoden der Hochdurchsatzsequenzierung, welche unter dem Begriff „Next-Generation-Sequencing (NGS)“ zusammengefasst werden [135]. Da es viele verschiedene solcher Methoden gibt (siehe z.B. [141, 142]), die wir an dieser Stelle nicht alle ausführlich beschreiben können, werden wir uns auf die Gemeinsamkeiten der Methoden konzentrieren und lediglich zwei Methoden genauer betrachten: Zum einen die sog. *Pyrosequenzierung* [107], welche beispielsweise zur Sequenzierung der mikrobiellen Genome im Rahmen des Human-Microbiome-Projects (HMPs) verwendet wurde und zum anderen die sog. *Illumina-Sequenzierung* [20], die heutzutage dominierende Technik.

In der Regel wird bei jedem NGS-Verfahren zunächst ein (großes) DNA-Molekül in kurze Fragmente zerteilt. Dies geschieht z.B. mit Hilfe von Enzymen oder physikalisch durch das Erzeugen mechanischer Scherkräfte, die die DNA aufbrechen. Anschließend werden kurze synthetische Oligonucleotide, sog. *Adapter*, an die Enden dieser Fragmente gebunden. Mit Hilfe der Adapter können die DNA-Fragmente an einem Träger (z.B. an Mikroperlen (*Englisch:* beads) bei der Pyrosequenzierung oder an einer speziellen Platte bei der Illumina-Sequenzierung) befestigt werden. Im nächsten Schritt werden die DNA-Fragmente dann durch die Polymerasekettenreaktion (*Englisch*: polymerase chain reaction, PCR) amplifiziert bzw. vervielfältigt. Die Amplifizierung wird dabei so durchgeführt, dass die neu generierten DNA-Kopien nicht frei beweglich sind, sondern in der Nähe des ursprünglichen DNA-Fragments am Träger gebunden werden. Auf diese Weise entstehen Cluster von meist etwa 1000 identischen DNA-Fragmenten. Diese Cluster werden nach abgeschlossener DNA-Amplifizierung alle gleichzeitig sequenziert. [135, 7]

Die Sequenzierung, d.h. die Bestimmung der Nucleotidabfolge der DNA-Fragmente, beginnt in der Regel unabhängig vom gewählten Verfahren damit, dass man zu den DNA-Fragmenten die zur DNA-Synthese benötigten Primer[1], die DNA-Polymerase und Nucleotide bzw. Desoxyribonucleosidtriphosphate (dNTPs) hinzugibt [135].

[1] Kurze, einzelsträngige Abschnitte einer Nucleinsäure, gewöhnlich RNA, die das erforderliche Starterfragment bzw. Template für die DNA-Polymerase zur Synthese eines neuen DNA-Strangs bilden [135].

Bei der Pyrosequenzierung, welche – entsprechend ihres Namens – auf der Detektion von Pyrophosphaten beruht, werden die vier Nukleotid- bzw. dNTP-Arten (dATP, dCTP, dGTP und dTTP) nacheinander angeboten, d.h. es wird immer nur eine einzige Nukleotidart gleichzeitig zu den DNA-Fragmenten hinzugegeben. Das passende Nukleotid wird dann von einer DNA-Polymerase eingebaut, wobei Pyrophosphat freigesetzt wird. Das Pyrophosphat führt über eine mehrstufige Reaktion, an der ein Luciferase-Enzym beteiligt ist, zur Emission von Licht, welches von einem Detektor aufgefangen wird. Aus dem Zusammenhang von angebotenem Baustein und Lichtemission lässt sich so das eingebaute Nukleotid ermitteln und die gesamte Sequenz eines DNA-Fragments ergibt sich aus der Reihenfolge der Nukleotide. [58]

Bei der Illuminia-Sequenzierung ist jedes Nukleotid mit einem wieder entfernbaren fluoreszierenden Farbstoff (für jede der vier Basen eine andere Farbe) sowie einer ebenfalls wieder entfernbaren Blockierungsgruppe (*Englisch:* terminator), d.h. einer chemischen Gruppe, die die Elongation des DNA-Stranges durch die DNA-Polymerase blockiert und somit sicherstellt, dass nur eine einzige Base eingebaut werden kann, verbunden. Im Gegensatz zur Pyrosequenzierung können die vier (fluoreszenzmarkierten) Nukleotidarten daher alle gleichzeitig zu den DNA-Clustern hinzugefügt werden (und müssen nicht einzeln, nacheinander hinzugegeben werden). An jedem Cluster wird dann das passende Nukleotid (das komplementär zum nächsten Nukleotid im DNA-Fragment ist) kovalent eingebaut und die nicht eingebauten Nukleotide werden weggewaschen. Eine hochauflösende Digitalkamera nimmt dabei ein Bild auf, das registriert, welches der vier Nukleotide an jedem Cluster in die Kette eingebaut wurde. Die fluoreszierende Markierung und die Blockierungsgruppe werden anschließend (enzymatisch) entfernt und der Vorgang wird viele Male wiederholt. Auf diese Weise werden Milliarden von Sequenzierungsreaktionen gleichzeitig durchgeführt. Indem die Farbveränderungen an jedem Cluster verfolgt werden, kann die DNA-Sequenz ermittelt werden. Obwohl jede einzelne abgelesene Sequenz relativ kurz ist (wenige hundert Nukleotide), können die Milliarden von gleichzeitig durchgeführten Sequenzierungen in etwa einem Tag den Wert mehrerer menschlicher Genome ergeben. [7]

Am Ende jedes Sequenzierungsverfahrens steht schließlich die Datenanalyse, da die gewonnenen DNA-Sequenzen, die man auch als *Reads* bezeichnet, in der richtigen Reihenfolge zusammengesetzt werden müssen. Hierzu werden moderne bioinformatische Computeralgorithmen verwendet. [135]

3.2.2. 16S-rRNA-Gensequenzierung

Da es unpraktisch ist, das gesamte Genom aller Zellen in einer Mikrobiomprobe (z.B. Stuhl- oder Speichelprobe) zu sequenzieren, nutzt man zur taxonomischen Identifizierung der in der Probe vorhandenen Spezies häufig lediglich einen sog. *Marker*, d.h. eine vergleichsweise kurze DNA-Sequenz, die das sie enthaltende Genom möglichst eindeutig identifiziert, ohne dass man das gesamte Genom sequenzieren muss. Der bei Weitem am häufigsten genutzte solche Marker ist das Gen, welches die 16*S*-rRNA, den Hauptbestandteil der kleinen Untereinheit prokaryotischer Ribosomen, kodiert. Dieses Gen ist ca. 1500 Nukleotide lang und eignet sich besonders gut zur taxonomischen Identifizierung von Mikroorganismen, da es sowohl evolutionär hoch konservierte Sequenzbereiche enthält, als auch neun dazwischenliegende variable Regionen, die sich von Spezies zu Spezies unterscheiden.

Da eine solche 16S-rRNA-Gensequenzierung inzwischen sehr einfach und kostengünstig durchgeführt werden kann, gibt es inzwischen mehrere Datenbanken, wie beipielsweise GreenGenes [44], das Ribosomal-Database-Project [36] oder Silva [130], in denen die 16S-Sequenzen verschiedenster Spezies hinterlegt sind. Mit diesen lassen sich die aus den Proben gewonnen Gensequenzen abgleichen und so die in der Probe vorhandenen Taxa bestimmen.

3.2.3. Zusammenfassung von 16S-rRNA-Sequenzen zu OTUs

Eine bioinformatische Herausforderung, die sich bei der Analyse von 16S-rRNA-Genen ergibt, ist die genaue Definition einer „einzigartigen“ Sequenz. Obwohl ein Großteil des 16S-rRNA-Gens evolutionstechnisch hoch konserviert ist, sind mehrere der sequenzierten Regionen variabel oder hypervariabel, sodass sich eine geringe Anzahl von Basenpaaren in einem sehr kurzen (evolutionären) Zeitraum ändern kann. Des Weiteren erschweren verschiedene bei Mikroorganismen auftretende Phänomene, wie beispielsweise der horizontale Gentransfer[2], die präzise Definition einer „Art“ und limitieren die Möglichkeiten, diese technisch zu bestimmen. Da zudem 16S-rRNA-Genregionen in der Regel jeweils nur einmal sequenziert werden, besteht eine recht hohe Wahrscheinlichkeit, dass sie mindestens einen Sequenzierungsfehler enthalten. Dies bedeutet, dass die Forderung nach einer 100-prozentigen Übereinstimmung von betrachteten Sequenzen äußerst konservativ ist und dazu führen kann, dass klonale Genome als unterschiedliche Organismen behandelt werden. Um dies zu vermeiden, werden ähnliche Sequenzen (mit einer Übereinstimmung von typischerweise mindestens 95, 97 oder 99 Prozent) zu Gruppen zusammengefasst, die als *operative taxonomische Einheiten (*Englisch*: operational taxonomic units, OTUs)* oder *Phylotypen* bezeichnet werden. OTUs treten in vielen Mikrobiom-Diversitätsanalysen an die Stelle von Arten, da für bestimmte Markersequenzen oft keine benannten Artengenome verfügbar sind. Die Zuordnung von Sequenzen zu OTUs wird dabei als *Binning* bezeichnet und kann beispielsweise durch unüberwachtes Clustering ähnlicher Sequenzen oder durch phylogenetische Modelle, die Mutationsraten und evolutionäre Beziehungen berücksichtigen, erfolgen.
Das Binning ermöglicht somit die Analyse einer mikrobiellen Gemeinschaft in Form von diskreten Bins oder OTUs, wobei ein Pool von Mikrobiomsequenzen als Histogramm von Bin-Zahlen dargestellt wird. Alternativ dazu kann dieses Histogramm binarisiert werden, sodass nur die Information über die An- oder Abwesenheit einer Spezies (unabhängig von ihrer Häufigkeit) in einer Probe betrachtet wird. Solche binären An- und Abwesenheitsmuster stellen auch den Ausgangspunkt für die ESABO-Methode zur Inferenz mikrobieller Interaktionsnetzwerke dar, welche im folgenden Kapitel ausführlich beschrieben wird.

[2] Die Übertragung von Genen von einem Individuum auf ein anderes, gleichzeitig lebendes Indiviuum [135], also innerhalb derselben Generation.

3.3. Die ESABO-Methode

Bei der Entropy-Shifts-of-Abundance-vectors-under-Boolean-Operations (ESABO)-Methode handelt es sich um eine Methode zur Inferenz mikrobieller Interaktionsnetzwerke aus den Auftretenshäufigkeiten (Abundanzen) von verschiedenen Bakterien. Sie wurde ursprünglich 2017 von Claussen et al. [34] entwickelt.
Die Besonderheit dieser Methode besteht darin, dass nur binäre Abundanzvektoren (d.h. das Vorhandensein oder die Abwesenheit einer mikrobiellen Spezies in den vorhandenen Proben) betrachtet werden, was die Methode besonders für die Untersuchung des Niedrigabundanzsegments des Mikrobioms, d.h. selten vorkommende Bakterienstämme, interessant macht [34].

Um die ESABO-Methode zu verwenden, stellen wir die gemessenen Spezies-Abundanzen als Boolesche Matrix, $A \in \mathbb{B}^{N_A \times N}$ mit $\mathbb{B} = \{0; 1\}$, dar, wobei jede Spalte einer Spezies und jede Zeile einer Probe entspricht. Die Spaltenvektoren dieser Matrix werden auch als Abundanzvektoren der Spezies bezeichnet.
Der ESABO-Score zweier Spezies i und j, der im Rahmen der ESABO-Methode die Art ihrer Interaktion (positive gegenseitige Beeinflussung bzw. Mutualismus oder negative gegenseitige Beeinflussung bzw. Konkurrenz) angibt, lässt sich dann anhand ihrer Abundanzvektoren $\vec{b}_i$ und $\vec{b}_j$ wie folgt berechnen [34]:

(1) Die beiden Abundanzvektoren $\vec{b}_i$ und $\vec{b}_j$ werden komponentenweise per logischer AND-Operation verknüpft, d.h.

$$\left(\vec{x}_{ij}^{\text{AND}}\right)_k = \left(\vec{b}_i\right)_k \text{AND} \left(\vec{b}_j\right)_k .$$

(2) Die Entropie des resultierenden Vektors $\vec{x}_{ij}^{\text{AND}}$ wird berechnet über

$$H(\vec{x}_{ij}^{\text{AND}}) = -\sum_{l \in \{0,1\}} p_l\left(\vec{x}_{ij}^{\text{AND}}\right) \ln\left(p_l\left(\vec{x}_{ij}^{\text{AND}}\right)\right),$$

wobei $p_l\left(\vec{x}_{ij}^{\text{AND}}\right)$ die relative Häufigkeit des Eintrags $l \in \{0, 1\}$ im Vektor $\vec{x}_{ij}^{\text{AND}}$ darstellt.

(3) Zunächst werden die Einträge des Abundanzvektors $\vec{b}_j$ zufällig permutiert, wodurch sich ein neuer Vektor $\vec{b}_j^*$ ergibt. Anschließend wird wieder die AND-Operation zwischen $\vec{b}_i$ und $\vec{b}_j^*$ ausgeführt und die Entropie $H(\vec{x}_{ij*}^{\text{AND}})$ von $\left(\vec{x}_{ij*}^{\text{AND}}\right)_k = \left(\vec{b}_i\right)_k \text{AND} \left(\vec{b}_j^*\right)_k$ berechnet. Dieser Schritt wird $R = 1000$ Mal wiederholt, um eine Verteilung von Entropiewerten zu erhalten (welche auf verschiedenen Permutationen von $\vec{b}_j$ basiert), die als Nullmodell dient und deren Mittelwert μ sowie Standardabweichung σ berechnet werden.

(4) Der Entropiewert $H(\vec{x}_{ij}^{\text{AND}})$ des ursprünglichen Vektors sowie Mittelwert (μ) und Standardabweichung (σ) des Nullmodells werden genutzt, um einen z-Score zu berechnen:

$$Z_{ij} = \frac{H(\vec{x}_{ij}^{\text{AND}}) - \mu}{\sigma}. \tag{3.1}$$

Dieser z-Score entspricht dem sogenannten ESABO-Score für die zwei Spezies i und j.

Im Falle einer positiven Interaktion zwischen Spezies i und j erwarten wir einen positiven ESABO-Score, während wir im Falle einer negativen Interaktion zwischen i and j einen negativen ESABO-Score antizipieren [34].
Um ein endgültiges rekonstruiertes Netzwerk zu erhalten, ist es notwendig eine ESABO-Score-Schwelle Θ zu wählen, die angibt, ab welchem z-Score-Betrag eine Kante in das inferierte Netzwerk eingefügt werden soll. In [34] wurde vorgeschlagen, dass alle Kanten mit $|Z_{ij}| > 1$, d.h. Kanten, bei denen der Entropiewert um mehr als eine Standardabweichung über bzw. unter dem Entropiemittelwert des Nullmodells liegt, gesetzt werden sollten.

3.4. Erzeugung künstlicher binärer Abundanzmuster zum Test der ESABO-Methode

Ein Vorteil der ESABO-Methode ist, dass sie leicht getestet werden kann. Dazu erzeugen wir zunächst ein zufälliges ungerichtetes Boolesches Schwellennetzwerk mit N Knoten, welche N mikrobielle Spezies repräsentieren. Diese Knoten sind durch L_+ positive sowie L_- negative Kanten bzw. Interaktionen verbunden, welche den Einfluss eines Bakterienstamms auf die Präsenz des anderen beschreiben. Die Knoten in einem Booleschen Netzwerk können, wie bereits in Kapitel 2.2.2 beschrieben, lediglich zwei verschiedene Werte annehmen, entweder 1, d.h. die betrachtete Spezies ist in der Probe vorhanden, oder 0, was die Abwesenheit der Spezies bedeutet. Der Zustand oder die Dynamik eines jeden Knotens wird durch seine Boolesche Aktualisierungsfunktion bestimmt. In unserem Fall verwenden wir einfache Boolesche Schwellenwertfunktionen, bei denen der Wert s_i jedes Knotens bzw. jeder Spezies i von der Summe seiner Eingangssignale abhängt und in jedem Zeitschritt gemäß

$$s_i(t+1) = \begin{cases} 1, & \sum_{j=1}^{N} G_{ij} s_j(t) > 0 \\ s_i(t), & \sum_{j=1}^{N} G_{ij} s_j(t) = 0 \\ 0, & \sum_{j=1}^{N} G_{ij} s_j(t) < 0 \end{cases} \tag{3.2}$$

aktualisiert wird. G ist hierbei die Adjazenzmatrix des Interaktionsgraphen, wobei gilt

$$G_{ij} = \begin{cases} +1, & \text{für eine positive Interaktion (Mutualismus) zwischen } i \text{ und } j \\ -1, & \text{für eine negative Interaktion (Konkurrenz) zwischen } i \text{ und } j \\ 0, & \text{wenn es keine Interaktion zwischen } i \text{ und } j \text{ gibt.} \end{cases}$$

Da wir ungerichtete Netzwerke betrachten, ist G symmetrisch ($G_{ij} = G_{ji}$) und wir nehmen an, dass $G_{ii} = 0 \, \forall i$ ist, d.h. wir berücksichtigen keine Selbst-Inputs. Außerdem werden in dieser Arbeit stets alle Knoten des Netzwerks synchron aktualisiert.

Nach der Erzeugung eines Netzwerks bestimmen wir seine Attraktoren, die wir als stationäre Mikrobiomzusammensetzungen interpretieren, mithilfe des in [75] beschriebenen Algorithmus. Um sicherzustellen, dass alle Attraktoren des Netzwerks gefunden werden, aktualisieren wir das Netzwerk ausgehend von jedem seiner 2^N möglichen Zustände (im Gegensatz zur ursprünglichen Herangehensweise in [34], wo nur 1000 der $2^{15} = 32768$ Zustände zur Attraktorsuche verwendet wurden).

Für den seltenen Fall, in dem ein gefundener Attraktor kein Fixpunkt, sondern ein zyklischer Attraktor ist, wählen wir den ersten angetroffenen Attraktorzustand als die ermittelte stationäre Mikrobiomzusammensetzung.
Schließlich verwenden wir die Attraktoren des Netzwerks, um es mit der ESABO-Methode zu rekonstruieren und überprüfen die Qualität dieser Rekonstruktion (siehe Kapitel 3.7.1).

3.5. Verbesserung der ESABO-Methode

Bevor wir nun zur Rekonstruktion von Netzwerken mithilfe der ESABO-Methode übergehen, sollen in diesem Abschnitt zwei wichtige Änderungen vorgestellt werden, die ich an der ESABO-Methode vorgenommen habe. Während die erste dieser Änderungen, nämlich die analytische Berechnung der ESABO-Scores, hauptsächlich zur Verbesserung der Performanz der Methode dient und die exakte Reproduzierbarkeit der Ergebnisse sicherstellt, behebt die zweite Veränderung, das Vertauschen von Nullen und Einsen unter bestimmten Bedingungen, das Problem, dass in der bisherigen Version der ESABO-Methode für mutualistische Wechselwirkungen in seltenen Fällen große negative ESABO-Scores auftreten (siehe [34]).

3.5.1. Analytische Formel für die Berechnung des ESABO-Scores bzw. μ und σ

Um die Berechnung der ESABO-Scores Z zu beschleunigen und die Zufälligkeit der erhaltenen Ergebnisse zu vermeiden, führen wir eine analytische Formel für die Berechnung des Mittelwertes μ und der Standardabweichung σ der Entropieverteilung ein, die man erhält, wenn man die Entropie für alle $N_A!$ möglichen Permutationen $\pi(j)$ der Einträge von $\vec{b}_j$ berechnet.
Die Standardabweichung lässt sich wie folgt berechnen:

$$\sigma^2 = \langle H^2 \rangle - \mu^2 \tag{3.3}$$

mit dem Mittelwert

$$\begin{aligned} \mu &= \frac{1}{N_A!} \sum_{\pi \in P} H\left(\vec{x}^{\text{AND}}_{i\pi(j)}\right) \\ &= \frac{1}{N_A!} \sum_{z} H_z \, w(z) \end{aligned} \tag{3.4}$$

und

$$\langle H^2 \rangle = \frac{1}{N_A!} \sum_{z} H_z^2 \, w(z). \tag{3.5}$$

N_A ist dabei die Anzahl der Einträge des Abundanzvektors $\vec{b}_j$, $\pi(j)$ ist eine Permutation dieser Einträge und P ist die Menge aller möglichen Permutationen der Einträge von $\vec{b}_j$. $z(\pi(j))$ ist die Anzahl der Einsen in $\vec{x}^{\text{AND}}_{i\pi(j)}$ und $w(z_0)$ ist die Anzahl der Permutationen π, die zu $z(\pi) = z_0$ Einsen in $\vec{x}^{\text{AND}}_{i\pi(j)}$ führen.

Die Anzahl $w(z)$ der Permutationen, die zu z Einsen in $\vec{x}^{\text{AND}}_{i\pi(j)}$ führen, kann (für $n \geq m$) berechnet werden durch

$$w(z) = \binom{n}{z}\binom{N_A - n}{m - z} m!\,(N_A - m)!, \tag{3.6}$$

falls $z \in [\text{Max}(0, n + m - N_A), \text{Min}(n, m)]$.
Ansonsten ist $w(z) = 0$.

n ist hier die Anzahl der Einsen in $\vec{b}_i$ und m ist die Anzahl der Einsen in $\vec{b}_j$.
$\text{Max}(0, n+m-N_A)$ entspricht der minimal möglichen und $\text{Min}(n, m)$ der maximal möglichen Anzahl an Einsen in $\vec{x}^{\text{AND}}_{i\pi(j)}$, die sich ergibt, wenn bei der AND-Operation möglichst wenige bzw. möglichst viele Einsen der beiden Vektoren $\vec{b}_i$ und $\vec{b}_j$ aufeinander fallen.

Zur Herleitung von Formel (3.6) müssen wir uns mithilfe der Kombinatorik überlegen bzw. abzählen, wie viele der $N_A!$ möglichen Permutationen der Einträge des Vektors $\vec{b}_j$ zu einer bestimmten Anzahl z von Einsen im resultierenden Vektor $\vec{x}^{\text{AND}}_{i\pi(j)}$ führen.
Hierzu nehmen wir zunächst o.B.d.A. an, dass $n \geq m$ ist (ansonsten vertauschen wir die beiden Vektoren) und verteilen die Einträge bzw. die Einsen von $\vec{b}_j$ so auf die jeweiligen Vektorkomponenten von $\vec{b}_i$ (Plätze), dass das jeweilige z erreicht wird. Hierbei gehen wir wie folgt vor:

1. Als erstes verteilen wir z Einsen von $\vec{b}_j$ auf die n Einsen von $\vec{b}_i$, um nach der AND-Operation die gewünschten z Eins-Einträge zu erhalten. Hierfür gibt es insgesamt $\binom{n}{z}$ Möglichkeiten.

2. Als zweites verteilen wir die restlichen $(m - z)$ Einsen von $\vec{b}_j$ auf die $(N_A - n)$ Nullen von $\vec{b}_i$, damit bei der AND-Operation keine weiteren Einsen entstehen, als die z Einträge von oben. Hierfür gibt es $\binom{N_A - n}{m - z}$ Möglichkeiten.

3. Nun müssen wir noch berücksichtigen, dass die m Einsen von $\vec{b}_j$ permutiert werden können, da die Komponenten von $\vec{b}_j$ unterscheidbar sind. Dies führt auf den Term $m!$.

4. Zuletzt permutieren wir auch die $(N_A - m)$ Nullen von $\vec{b}_j$, wofür es $(N_A - m)!$ Möglichkeiten gibt.

Dieses Vorgehen wird in Abbildung 3.1 anhand eines einfachen Beispiels mit $N_A = 4$, $n = 3$ und $m = 2$ veranschaulicht.

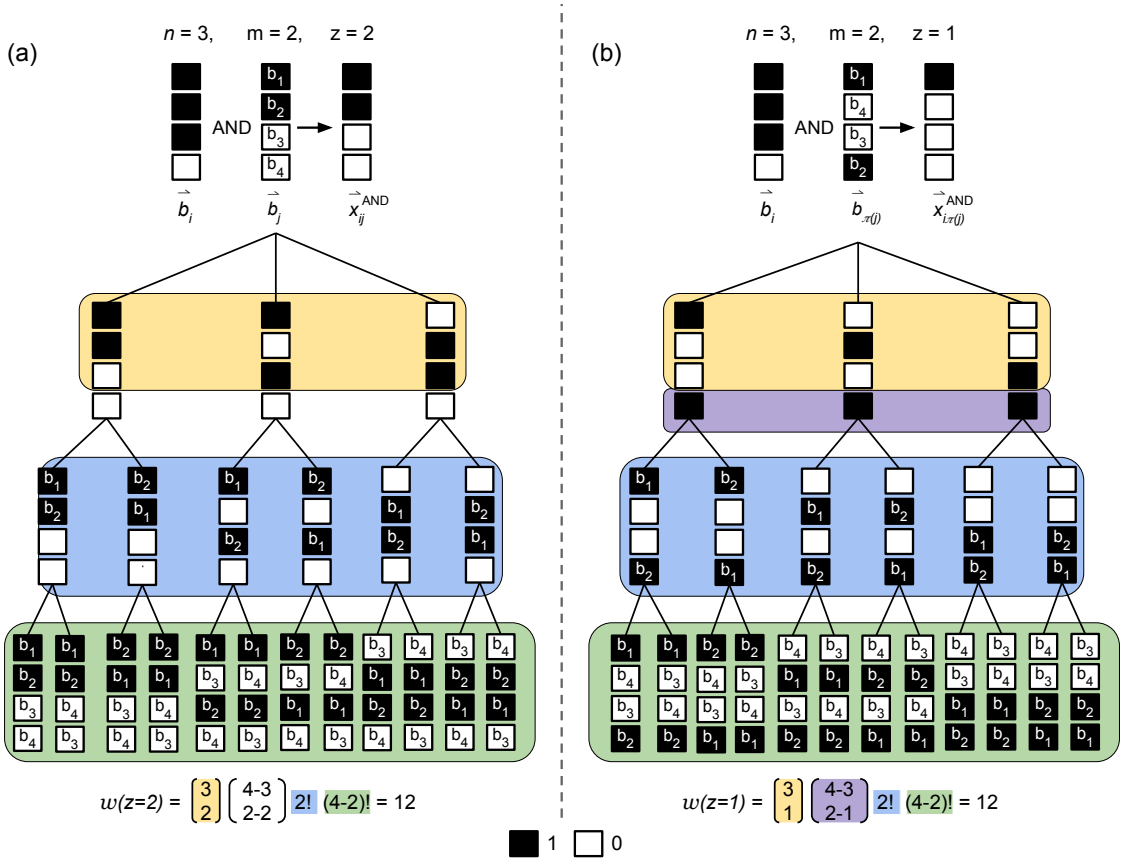


Abbildung 3.1.: Beispiel zur Berechnung der Anzahl $w(z)$ der Permutationen von $\vec{b}_j$, die gemäß Gleichung (3.6) zu z Einsen im resultierenden Vektor $\vec{x}^{\text{AND}}_{i\pi(j)}$ führen.
In diesem Beispiel hat jeder Abundanzvektor $N_A = 4$ Einträge, sodass es insgesamt $N_A! = 24$ mögliche Permutationen des Vektors $\vec{b}_j$ gibt. Da im Vektor $\vec{b}_i$ $n = 3$ Einsen (schwarz dargestellt) vorliegen und der zu permutierenden Vektor $\vec{b}_j$ $m = 2$ Einsen aufweist, können nach Durchführung der AND-Operation im resultierenden Vektor $\vec{x}^{\text{AND}}_{i\pi(j)}$ minimal $\text{Max}(0, n + m - N_A) = 1$ und maximal $\text{Min}(n, m) = 2$ Einsen auftreten, sodass gilt $z \in [1, 2]$. In Abbildung (a) sind die Permutationen gezeigt, die zu $z = 2$ Einsen in $\vec{x}^{\text{AND}}_{i\pi(j)}$ führen und in Abbildung (b) diejenigen, die zu $z = 1$ Einsen führen.
Im ersten Schritt, der gelb hinterlegt ist, sind alle $\binom{n}{z}$ Möglichkeiten gezeigt, auf die man z (hier noch als ununterscheidbar angenommene) Einsen von $\vec{b}_j$ auf die n Einsen von $\vec{b}_i$ verteilen kann. Sowohl in (a) als auch in (b) gibt es hierfür 3 Möglichkeiten. Während in (a) für $z = 2$ hiermit alle Einsen von $\vec{b}_j$ verteilt sind, muss in (b) noch eine Eins auf die Null von $\vec{b}_i$ fallen, damit wir nach der AND-Operation im resultierenden Vektor $z = 1$ erhalten. Hierfür gibt es genau eine Möglichkeit, die lila hinterlegt ist. Im nächsten Schritt, der blau markiert ist, berücksichtigen wir nun, dass die m Einsen als Vektorkomponenten von $\vec{b}_j$ unterscheidbar sind und auf $m! = 2$ Arten permutiert werden können. Im letzten Schritt permutieren wir nun auch die Nullen von $\vec{b}_j$, wofür es $(N_A - m)! = 2$ Möglichkeiten gibt. Die sich so insgesamt ergebenden Permutationen von $\vec{b}_j$, die zum gewünschten z ($z = 2$ (a) oder $z = 1$ (b)) führen, sind grün hinterlegt. Wie man sieht, gibt es sowohl um $z = 2$ als auch um $z = 1$ Einsen im Vektor $\vec{x}^{\text{AND}}_{i\pi(j)}$ zu erhalten, 12 verschiedene mögliche Permutationen des Vektors $\vec{b}_j$ (inklusive der identischen Permutation).

3.5.2. Vertauschen von Nullen und Einsen in Abundanzvektoren mit einer hohen relativen Häufigkeit von Einsen vor Durchführung der AND-Operation

In diesem Abschnitt wollen wir uns mit der Frage beschäftigen, wieso bei Verwendung der ESABO-Methode manchmal für mutualistische Wechselwirkungen anstatt der erwarteten positiven ESABO-Scores hohe negative ESABO-Score-Werte auftreten. Dieses Problem wurde bereits von den Autoren in [34] beschrieben und soll hier genauer beleuchtet werden.
Durch die analytische Berechnung von ESABO-Scores für verschiedene Speziespaare in unterschiedlichen Netzwerken ergab sich die Vermutung, dass negative z-Scores für mutualistische Interaktionen immer dann auftreten, wenn die zwei betrachteten Spezies besonders häufig in den Proben vorhanden sind.
Um diese Hypothese zu bestätigen, betrachten wir den Grenzfall zweier Spezies mit identischen Abundanzvektoren und berechnen den ESABO-Score ihrer Interaktion für unterschiedliche relative Häufigkeiten von Einsen in ihren Abundanzvektoren. Das Ergebnis dieser Untersuchung ist in Abbildung 3.2 dargestellt.

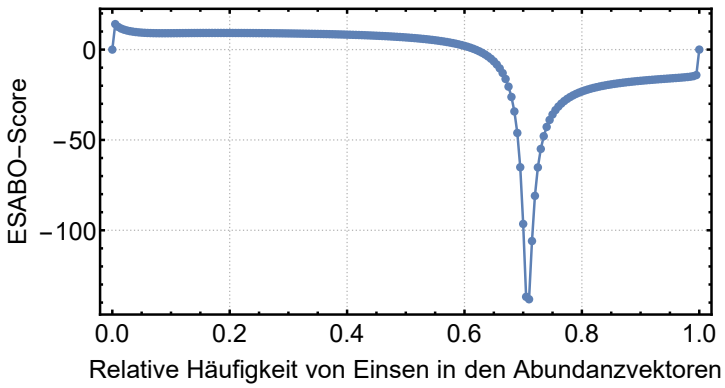


Abbildung 3.2.: ESABO-Score für die mutualistische Interaktion zwischen zwei Spezies mit identischen Abundanzvektoren in Abhängigkeit von der relativen Häufigkeit von Einsen in den Abundanzvektoren. In diesem Beispiel hat jeder Abundanzvektor 200 Einträge (bzw. besteht aus 200 Proben).

In der Tat ergeben sich negative ESABO-Scores, sobald eine hohe relative Häufigkeit von Einsen ($p_1 \gtrapprox 0.61$) in den Abundanzvektoren der Spezies vorliegt.

Wir können dieses Phänomen leicht verstehen, wenn wir zwei Spezies mit identischen Abundanzvektoren betrachten, die häufig in den Proben vorhanden sind, und dann die unterschiedlichen Entropiewerte berechnen, die durch eine Permutation der Einträge des zweiten Abundanzvektors erhalten werden können. Abbildung 3.3 zeigt beispielhaft die Entropiewerte, die man nach einer Umordnung eines Abundanzvektors mit 10 Einträgen (dies entspricht 10 Proben) und $p_1 = 0.8$ erhalten kann.
Wir stellen fest, dass die Entropie durch eine Permutation der Einträge des zweiten Abundanzvektors $\vec{b}_j$ nur zunehmen oder konstant bleiben kann, aber niemals abnimmt. Daher ist der Mittelwert μ immer größer als der Entropiewert $H(\vec{x}_{ij}^{\text{AND}})$, was zu einem negativen ESABO-Score führt.

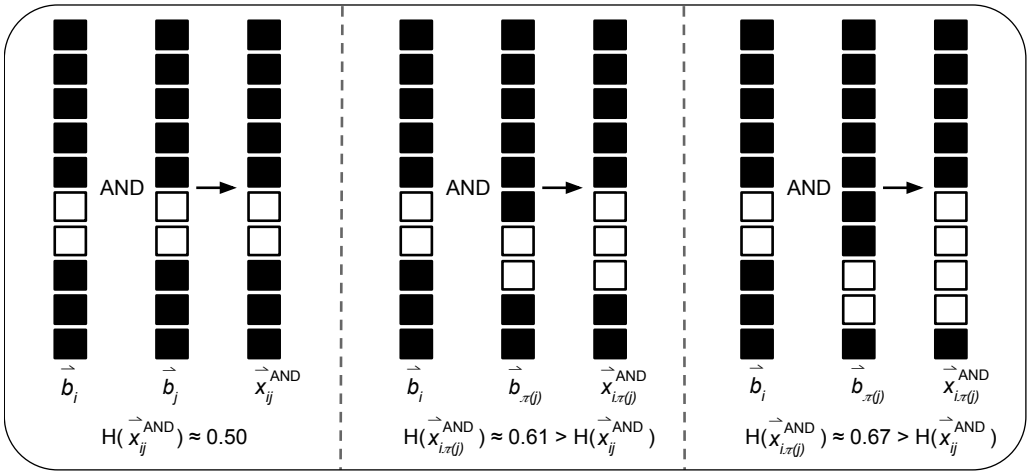


Abbildung 3.3.: Schematische Darstellung der verschiedenen Entropiewerte, die sich nach einer Permutation der Einträge des zweiten Abundanzvektors $\vec{b}_j$ ergeben können für ein Beispiel mit $N_A = 10$ Proben und $m = n = 8$ Einsen (schwarz dargestellt) in den beiden Abundanzvektoren $\vec{b}_i$ und $\vec{b}_j$.

Dieses Problem lässt sich vermeiden, indem wir vor Durchführung der logischen AND-Operation die Nullen und Einsen in einem Paar von Abundanzvektoren vertauschen, falls einer der Vektoren eine hohe relative Häufigkeit von Einsen aufweist. Genauer gesagt, vertauschen wir Nullen und Einsen dann, wenn die relative Häufigkeit von Einsen in einem der beiden betrachteten Abundanzvektoren höher als 50% ist, d.h. wenn $p_1(\vec{b}_i) > 0.5$ oder $p_1(\vec{b}_j) > 0.5$. Dies hat zur Folge, dass zwei Spezies, die identische Abundanzvektoren mit einem hohen Anteil an Einsen haben, keinen negativen ESABO-Score Z mehr erhalten, wie in Abbildung 3.4 gezeigt ist.

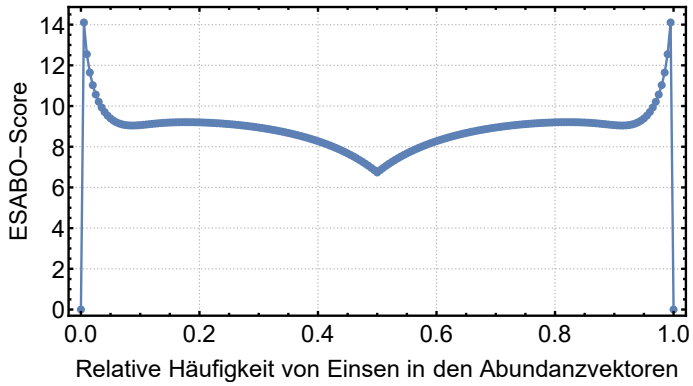


Abbildung 3.4.: ESABO-Score für die mutualistische Interaktion zwischen zwei Spezies mit identischen Abundanzvektoren in Abhängigkeit von der relativen Häufigkeit von Einsen (p_1) in den Abundanzvektoren. Für $p_1 > 0.5$ wurden Nullen und Einsen in den Abundanzvektoren vertauscht, bevor die logische AND-Operation durchgeführt wurde. Dies hat zur Folge, dass die ESABO-Scores immer positiv sind. In diesem Beispiel hat jeder Abundanzvektor 200 Einträge.

Für alle weiteren Untersuchungen verwenden wir von jetzt an immer diese verbesserte Version der ESABO-Methode (analytische Berechnung der z-Scores mit einem Vertauschen von Nullen und Einsen, wenn $p_1(\vec{b}_i) > 0.5$ oder $p_1(\vec{b}_j) > 0.5$), welche in Abbildung 3.5 schematisch dargestellt ist.

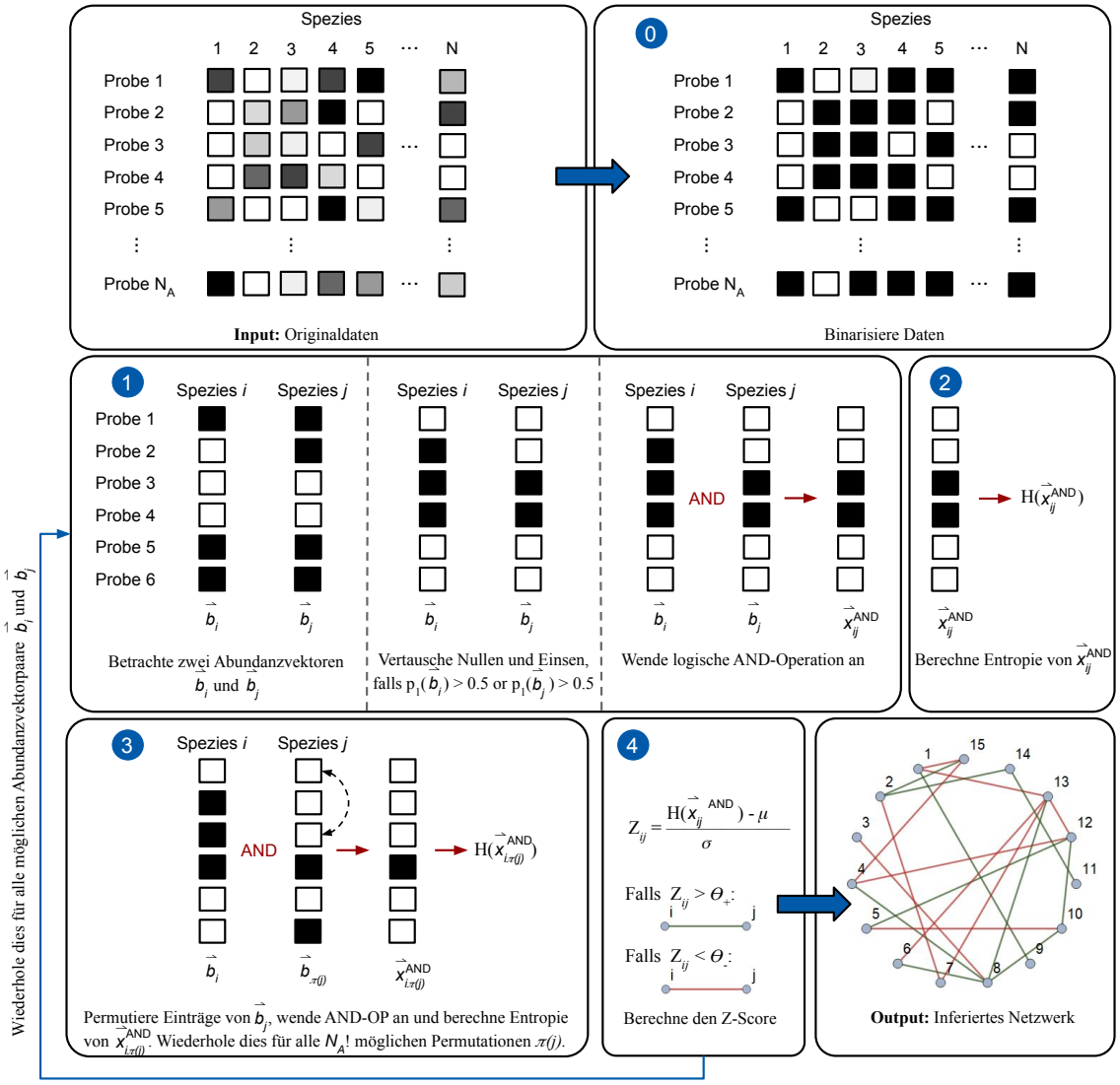


Abbildung 3.5.: Schematische Darstellung des verbesserten ESABO-Algorithmus. Im Gegensatz zur ursprünglichen ESABO-Methode werden die ESABO-Scores Z_{ij} analytisch berechnet sowie die Nullen und Einsen in einem Abundanzvektorpaar vor Durchführung der AND-Operation vertauscht, falls $p_1(\vec{b}_i) > 0.5$ oder $p_1(\vec{b}_j) > 0.5$.

3.6. Erweiterung der ESABO-Methode um einen evolutionären Algorithmus

Wie wir in Abschnitt 3.7.1 noch genauer sehen werden, weisen Netzwerke, die mit der ESABO-Methode rekonstruiert wurden, in der Regel nicht die gleichen Attraktoren auf, die zur Inferenz des Netzwerks verwendet wurden (vgl. Abb. 3.7). Aus diesem Grund unterwerfen wir die rekonstruierten Netzwerke einem einfachen evolutionären Algorithmus[3], der auf Mutation und Selektion basiert, um ihre Fähigkeit zu verbessern, die Attraktoren des ursprünglichen Netzwerks zu reproduzieren.
Der verwendete evolutionäre Algorithmus, den wir im Folgenden als ESABO-gestützte Evolution bezeichnen und der in C++ implementiert wurde, ist in Abbildung 3.6 schematisch dargestellt. Er besteht aus den folgenden vier zentralen Elementen: Der Erzeugung einer Population von M Netzwerken mit Hilfe der ESABO-Methode, der Bestimmung der Fitness aller Netzwerke in der Population, der fitnessproportionalen Selektion von M Netzwerken sowie der Mutation der selektierten Netzwerke, um die nächste Generation zu erzeugen. Diese Schritte werden dabei nacheinander so lange durchlaufen bis eine bestimmte Anzahl von Iterationen (bzw. Generationen) oder eine Fitness von $F = 1$ (entspricht einer vollständigen Rekonstruktion der ursprünglichen Attraktoren) erreicht wurde. Die konkrete Implementierung der vier wesentlichen Elemente des Algorithmus wird in den folgenden Abschnitten genauer erläutert.

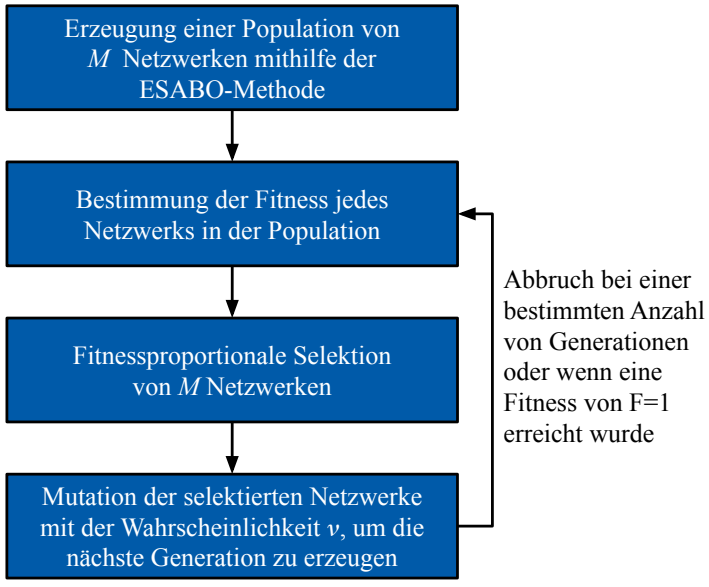


Abbildung 3.6.: Schematische Darstellung der um einen evolutionären Algorithmus erweiterten ESABO-Methode

[3] Ein dem Vorbild der natürlichen Evolution nachgebildetes stochastisches Optimierungsverfahren [27]. Für eine allgemeine Übersicht zu evolutionären Algorithmen sei beispielsweise auf [15] verwiesen.

3.6.1. Erzeugung einer Population von Netzwerken mithilfe der ESABO-Methode

Zu Beginn jeder Evolution wird eine Population von M Netzwerken erzeugt. Um diese initiale Population zu erzeugen, berechnen wir zunächst die ESABO-Scores für alle $\frac{N\cdot(N-1)}{2}$ möglichen Kanten im Netzwerk. Dann konstruieren wir M Netzwerke mit einer zunehmenden Anzahl von Kanten, indem wir mit dem Netzwerk beginnen, das die L_{min} Kanten enthält, die die höchsten z-Score-Beträge haben. Wir fügen sukzessive weitere Kanten entsprechend ihrem z-Score hinzu, bis wir Netzwerke mit $L_{min} + M - 1$ Verbindungen erhalten.

3.6.2. Fitnessbestimmung

Wenn alle Attraktoren des ursprünglichen Netzwerks für die Inferenz verwendet werden, entspricht die Fitness eines Netzes dem Jaccard-Index

$$F = \frac{|A_{\text{ori}} \cap A|}{|A_{\text{ori}} \cup A|}$$

zwischen seinen Attraktoren A und den Attraktoren A_{ori} des ursprünglichen Netzes, bzw. den ursprünglichen Proben. Der Vorteil dieses Gütemaßes ist, dass es auch für biologische Daten ausgewertet werden kann, bei denen das reale Interaktionsnetzwerk unbekannt ist.

Bei realen Mikrobiomen können wir jedoch nicht davon ausgehen, dass alle Attraktoren des zugrunde liegenden Interaktionsnetzwerks in den verfügbaren Proben vorkommen. Insbesondere Attraktoren mit kleinen Bassingrößen könnten in den Daten nicht vertreten sein. Daher untersuchen wir, was passiert, wenn einige Attraktoren, insbesondere solche mit kleinen Bassingrößen, während der Netzwerkinferenz nicht berücksichtigt werden. Wir erzeugen dabei die gewünschte Stichprobengröße N_A (die als fester Anteil der Gesamtzahl der Attraktoren zu verstehen ist), indem wir aus dem ursprünglichen Netzwerk die N_A Attraktoren mit den größten Bassingrößen auswählen.
Zur Bestimmung der Fitness eines solchen rekonstruierten Netzwerks, bei dessen Rekonstruktion - wie eben beschrieben - nicht alle Attraktoren genutzt wurden, gehen wir wie folgt vor: Zuerst aktualisieren wir das Netzwerk ausgehend von bis zu $N_{\text{ini}} = 1000$ zufälligen Anfangszuständen, um seine Attraktoren zu finden. Sobald wir so viele Attraktoren wie die Anzahl der Proben gefunden haben, brechen wir die Suche ab. Anschließend berechnen wir den Jaccard-Index zwischen den gefundenen Attraktoren und den ursprünglichen Proben. Da der Jaccard-Index einer solchen rekonstruierten Attraktormenge mit jeder Auswertung der Attraktoren variiert, mitteln wir ihn über $r = 100$ Durchläufe (sofern nicht anders angegeben). Dieser mittlere Jaccard-Index entspricht der Fitness des Netzwerks.

3.6.3. Fitnessproportionale Selektion

Nachdem die Fitness der einzelnen Individuen der Population bestimmt wurde, findet die Selektion statt. Hierbei werden M Individuen mit einer Wahrscheinlichkeit proportional zu ihrer relativen Fitness selektiert. Das heißt, dass die Wahrscheinlichkeit W_i, mit dem das Individuum i als (alleiniger) Elter eines bestimmten Individuums der nächsten Generation

ausgewählt wird, durch

$$W_i = \frac{F_i}{\sum_{j=1}^{N} F_j}$$

gegeben ist.

3.6.4. Mutation der selektierten Netzwerke

Jeder Nachkomme erhält dann mit der Wahrscheinlichkeit ν eine Mutation gegenüber der Netzwerkstruktur seines Elters. In dieser Arbeit werden dabei die folgenden drei Arten von Mutationen verwendet, die alle mit der gleichen Wahrscheinlichkeit stattfinden:

(1) Eine zufällig ausgewählte Kante wird gelöscht.

(2) Eine neue Kante wird in das Netzwerk eingefügt (gemäß ihres ESABO-Scores):
Zunächst losen wir mit gleicher Wahrscheinlichkeit aus, ob eine positive oder eine negative Kante gesetzt werden soll. Das Vorzeichen einer Kante wird dabei durch das Vorzeichen ihres zugehörigen ESABO-Scores Z festgelegt. Anschließend ziehen wir zufällig eine entsprechende Kante (linear) gewichtet nach ihrem ESABO-Score.

(3) Eine Kante des Netzwerks wird geändert:
Wir führen zunächst Mutation (1) und anschließend Mutation (2) aus.

3.6.5. Diskussion der Vor- und Nachteile der Erweiterung

Die Verwendung eines solchen evolutionären Algorithmus hat den großen Vorteil, dass es nicht notwendig ist, manuell einen bestimmten Schwellenwert zu wählen, der definiert, wie viele Kanten in das Netzwerk eingefügt werden sollen. Dies ist eine entscheidende Aufgabe in vielen anderen Netzwerkinferenzmethoden wie z.B. SparCC [57] oder SPIEC-EASI [88] und kann hier vermieden werden.
Ein Nachteil, der mit dem evolutionären Algorithmus einhergeht, ist, dass sich diese Methode aufgrund der langen Rechenzeiten bei der Attraktorsuche nur für die Analyse relativ kleiner Netzwerke eignet, also z.B. für Untersuchungen auf der Phylum- oder Klassenebene.

3.7. Ergebnisse

3.7.1. Analyse von simulierten Daten unter der Annahme, dass alle Attraktoren eines Netzwerks bekannt sind

In diesem Abschnitt testen wir die erweiterte ESABO-Methode mithilfe simulierter Daten, so wie es in Abschnitt 3.4 beschrieben wurde. Hierzu betrachten wir 40 zufällige Netzwerke mit N = 15 Knoten und $L_+ = L_- = 10$ positiven sowie negativen Kanten. Alle diese Netzwerke haben mehr als 200 verschiedene Attraktoren und sind zusammenhängend.
Wir rekonstruieren diese Netzwerke mit unserer ESABO-gestützten Evolution mit einer Populationsgröße von $M = 50$ und einer Mutationswahrscheinlichkeit von $\nu = 0.25$. In der Anfangspopulation, d.h. nach der Rekonstruktion mithilfe der einfachen ESABO-Methode, jedoch vor der Evolution, hat das Netzwerk mit den wenigsten Verknüpfungen $L_{\text{min}} = 10$ Kanten und das Netzwerk mit den meisten Verknüpfungen hat $L_{\text{min}} + M - 1 = 59$ Kanten.

Wir bewerten die Güte unserer Methode anhand von zwei verschiedenen Maßen.
Zum einen betrachten wir die Fitness F der evolvierten Netzwerke. Die Fitness eines Netzes gibt an, inwieweit seine Attraktoren mit denen des ursprünglichen Netzes übereinstimmen.
Zum anderen überprüfen wir, wie gut die Netzwerktopologien, d.h. die Kanten des inferierten Netzwerks und die des ursprünglichen Netzwerks, übereinstimmen. Dazu werten wir den Jaccard-Index

$$J = \frac{|L_{\text{ori}} \cap L|}{|L_{\text{ori}} \cup L|}$$

zwischen den Kanten L des fittesten evolvierten Netzwerks und den Kanten L_{ori} des ursprünglichen Netzwerks aus.

Darüber hinaus vergleichen wir die ESABO-gestützte Evolution mit zwei verschiedenen Arten einer zufälligen Evolution, bei denen jeweils während Mutation (2) eine völlig zufällige Kante gesetzt wird, unabhängig von ihrem ESABO-Score oder ihrem erwarteten Vorzeichen. Während die erste Art der zufälligen Evolution von Netzwerken ausgeht, die mit der (einfachen) ESABO-Methode rekonstruiert wurden (genau wie bei der ESABO-gestützten Evolution), geht die zweite Art der zufälligen Evolution von einer Population von Zufallsnetzwerken aus.

Im Allgemeinen erlaubt uns unser ESABO-gestützter Evolutionsalgorithmus, Netze zu finden, die die gleichen Attraktoren aufweisen wie die ursprünglichen Netze, d.h. die eine Fitness von $F = 1$ haben. Dies ist in Abbildung 3.7 zu sehen, die den maximalen Fitnesswert zeigt, der für jedes der 40 untersuchten Netzwerke zu irgendeinem Zeitpunkt während einer Evolution von 10000 Generationen erreicht wurde. Während Netzwerke, die ausschließlich mit der ESABO-Methode rekonstruiert wurden, im Median nur eine Fitness von $F = 0.3$ aufweisen, haben Netzwerke, die einer ESABO-gestützten Evolution unterzogen wurden, im Median eine Fitness von $F = 1$. Die beiden anderen evolutionären Algorithmen erzielen bessere Ergebnisse als die ESABO-Methode ohne Evolution, zeigen aber eine breitere Verteilung und einen deutlich niedrigeren Medianwert der Fitness als die ESABO-gestützte Evolution.

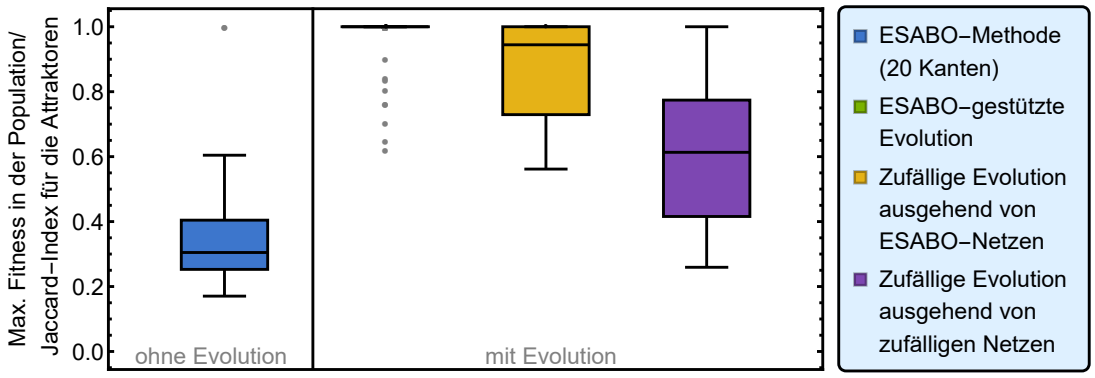


Abbildung 3.7.: Boxplots, die den Jaccard-Index zwischen den Attraktoren des ursprünglichen Netzwerks und den Attraktoren des inferierten oder fittesten evolvierten Netzwerks zeigen, das zu irgendeinem Zeitpunkt während der Evolution gefunden wurde. Netzwerke, die einer ESABO-gestützten Evolution unterzogen wurden, weisen im Median die höchste Attraktorübereinstimmung auf.
Es wurden 40 Netzwerke mit $N = 15$ Knoten und $L_+ = L_- = 10$ positiven und negativen Kanten untersucht. Die Evolution wurde für 10000 Generationen mit $M = 50$, $\nu = 0.25$ und $L_{\text{min}} = 10$ durchgeführt. Für die Netzwerkinferenz mit der ESABO-Methode wurden die 20 Kanten mit den höchsten ESABO-Score-Beträgen gesetzt.

Der Hauptgrund hierfür ist, dass die ESABO-gestützte Evolution viel schneller ist als eine zufällige Evolution. Dies wird in Abb. 3.8 deutlich, wo die Entwicklung der maximalen Fitness in der Population für die drei verschiedenen Versionen des evolutionären Algorithmus dargestellt ist. Während Abb. 3.8 (a) die Evolution der maximalen Fitness für zwei beispielhafte Netzwerke zeigt, zeigt Abb. 3.8 (b) den Median der maximalen Fitness für die 40 untersuchten Netzwerke und Abb. 3.8 (c) zeigt die Fitnessverteilungen in Form von Boxplots. Wie man sieht, weisen Netzwerke, die mit der ESABO-gestützten Evolution evolviert wurden, in der Regel einen steilen Fitnessanstieg auf (Abb. 3.8 (a)) und erreichen eine Median-Fitness von $F = 1$ in weniger als 500 Generationen (Abb. 3.8 (b), (c)), während der Fitnessanstieg bei beiden Arten der zufälligen Evolution deutlich langsamer verläuft. Zufällig evolvierte Netzwerke, bei denen die Evolution von einer Population zufälliger Netzwerke ausgeht, haben selbst nach 10000 Generationen eine deutlich geringere Median-Fitness von $F \approx 0.5$.

Außerdem haben Netzwerke, die mit der ESABO-gestützten Evolution evolviert wurden, nicht nur eine ähnliche Dynamik wie die Originalnetzwerke, sondern sie sind ihnen auch topologisch sehr ähnlich. Dies ist in Abbildung 3.9 zu sehen, in der der Jaccard-Index zwischen den Kanten des ursprünglichen Netzwerks und den Kanten des inferierten bzw. fittesten evolvierten Netzwerks, das zu irgendeinem Zeitpunkt während der Evolution gefunden wurde, für die 40 untersuchten Netzwerke dargestellt ist. Vergleicht man die Rekonstruktionsqualität, so stellt man fest, dass Netzwerke, die mit der ESABO-Methode inferiert wurden, im Median einen Jaccard-Index von $J = 0.74$ aufweisen, während Netzwerke, die der ESABO-gestützten Evolution unterzogen wurden, einen deutlich höheren Median-Jaccard-Index von $J = 1$ aufweisen.

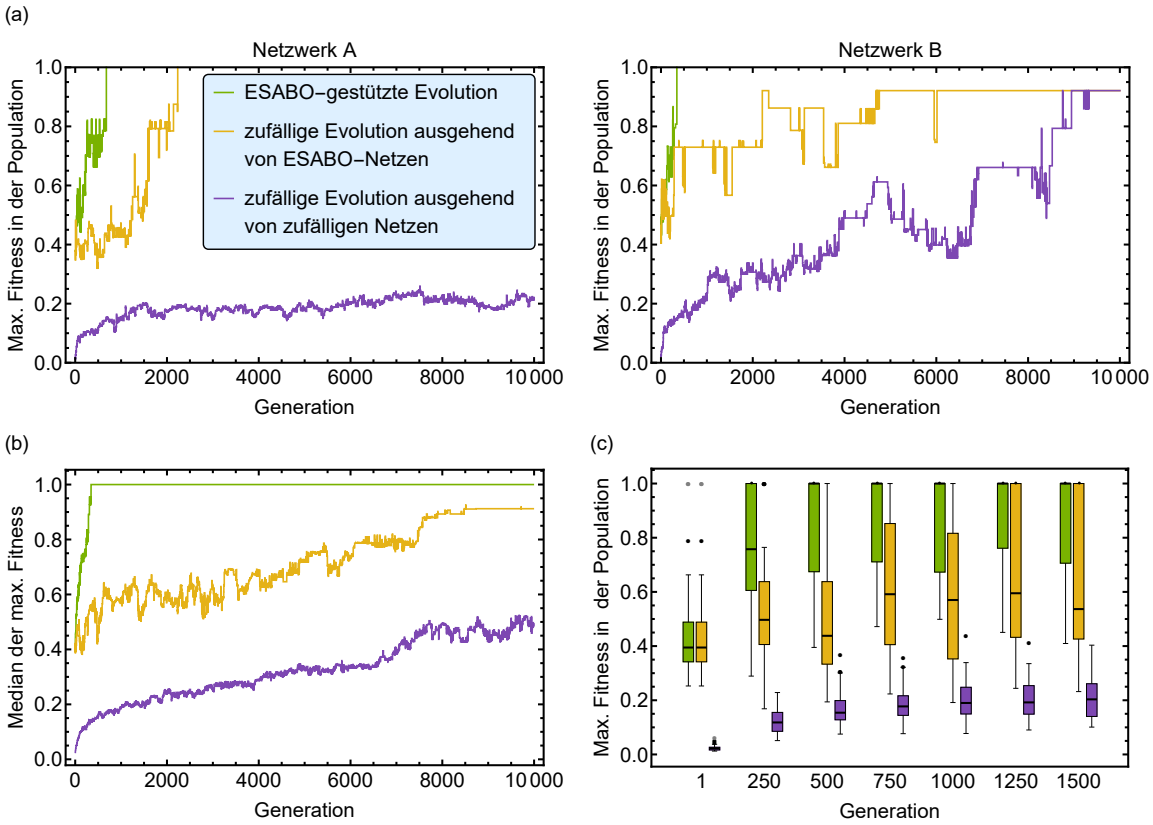


Abbildung 3.8.: Evolution der maximalen Fitness in der Population für die drei verschiedenen Versionen des evolutionären Algorithmus. Es zeigt sich, dass die Fitness im Falle der ESABO-gestützten Evolution viel schneller zunimmt als für die beiden anderen Versionen des evolutionären Algorithmus, bei denen die Mutationen zufällig erfolgen.
(a) Fitnesszunahme im Verlauf der Evolution für zwei beispielhafte Netzwerke.
(b) Median der maximalen Fitness für die 40 untersuchten Netzwerke im Verlauf der Evolution.
(c) Fitnessverteilung für die 40 untersuchten Netzwerke in Form von Boxplots.

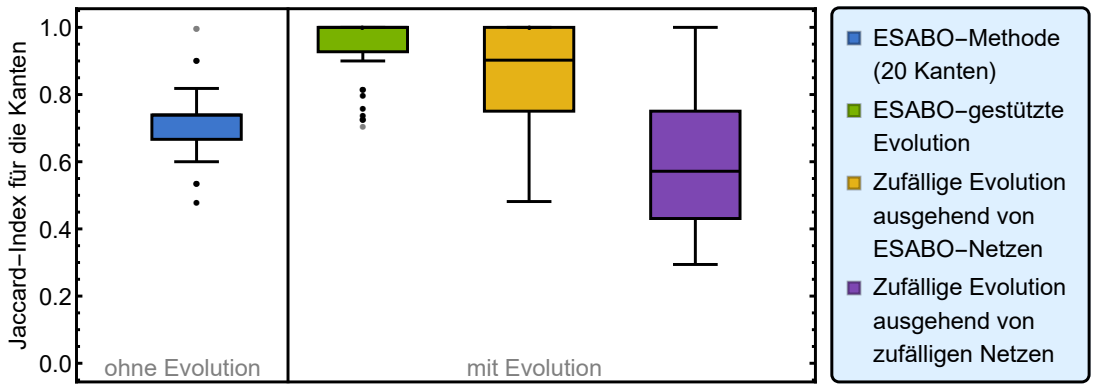


Abbildung 3.9.: Boxplots, die den Jaccard-Index zwischen den Kanten des ursprünglichen Netzwerks und den Kanten des inferierten oder fittesten evolvierten Netzwerks zeigen, das zu irgendeinem Zeitpunkt während der Evolution gefunden wurde. Man sieht, dass Netzwerke, die mit der ESABO-gestützten Evolution evolviert wurden, die größte topologische Ähnlichkeit zu den Originalnetzwerken haben.
Es wurden dieselben 40 Netzwerke mit $N = 15$ Knoten und $L_+ = L_- = 10$ positiven und negativen Kanten wie in Abb. 3.7 untersucht. Die Evolution wurde für 10000 Generationen mit $M = 50$, $\nu = 0.25$ und $L_{\min} = 10$ durchgeführt. Bei der Netzwerkinferenz mit der ESABO-Methode wurden die 20 Kanten mit den höchsten ESABO-Score-Beträgen gesetzt.

3.7.2. Analyse von simulierten Daten unter der Annahme, dass nur ein Teil aller Attraktoren bekannt ist

Wir betrachten erneut die 40 zufälligen Netzwerke aus dem vorherigen Kapitel und verwenden nun lediglich einen bestimmten prozentualen Anteil ihrer Attraktoren (immer diejenigen mit den größten Bassingrößen), um die Netzwerke mit dem ESABO-gestützten Evolutionsalgorithmus zu rekonstruieren.
Abbildung 3.10 zeigt die Qualität dieser Rekonstruktion, ausgedrückt als Jaccard-Index zwischen den Kanten des ursprünglichen Netzwerks und denen des rekonstruierten Netzwerks, wenn entweder alle Attraktoren, 75%, 50% oder 25% von ihnen für die Netzwerkinferenz verwendet wurden. Wie wir sehen können, ist die ESABO-gestützte Evolution der einfachen ESABO-Methode sowie den anderen Evolutionsarten stets überlegen. Sie funktioniert auch dann noch sehr gut, wenn nur 50% der Attraktoren als Grundlage für die Netzwerkinferenz verwendet werden. In diesem Fall erreicht die ESABO-gestützte Evolution im Median eine Rekonstruktionsqualität von $J = 0.87$. Werden hingegen nur 25% der Attraktoren für die Rekonstruktion eines Netzwerks berücksichtigt, zeigt die ESABO-gestützte Evolution einen starken Abfall der Inferenzleistung und erreicht nur eine Median-Inferenzqualität von $J = 0.52$.

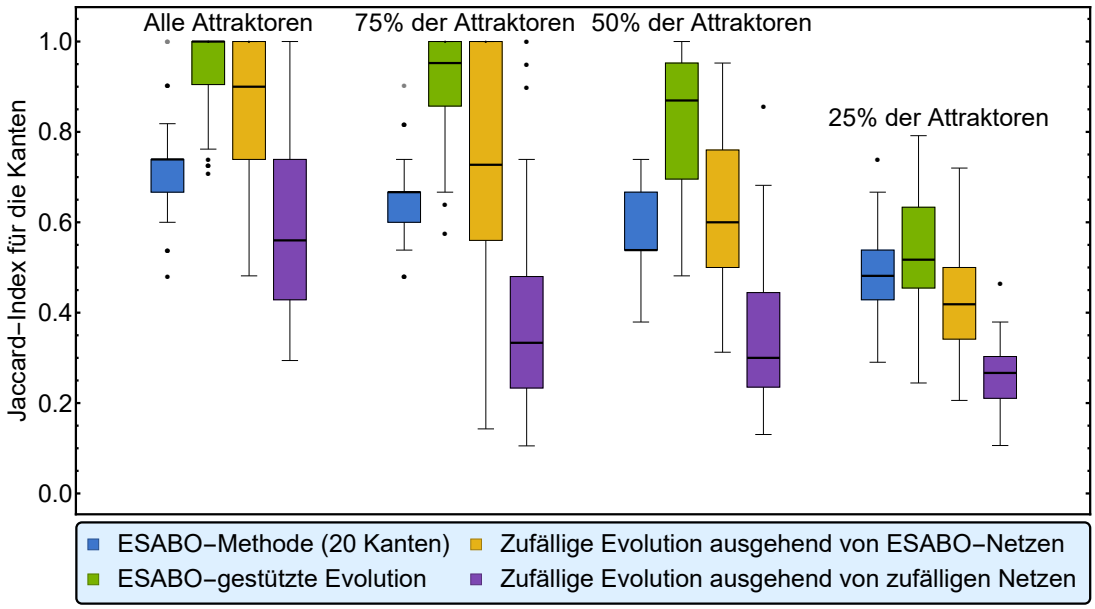


Abbildung 3.10.: Boxplots, die den Jaccard-Index zwischen den Kanten des ursprünglichen Netzwerks und den Kanten des inferierten oder fittesten evolvierten Netzwerks zeigen, das zu irgendeinem Zeitpunkt während der Evolution gefunden wurde. Netzwerke, die der ESABO-gestützten Evolution unterzogen wurden, haben stets die größte topologische Ähnlichkeit zu den Originalnetzwerken.
Es wurden dieselben 40 Netzwerke mit $N = 15$ Knoten und $L_+ = L_- = 10$ positiven und negativen Kanten wie in Abb. 3.7, unter Verwendung der dort beschriebenen Parameter, untersucht.

Bevor wir unsere Inferenzmethode auf echte biologische Abundanzdaten anwenden, weiten wir unsere Untersuchung auf größere zufällige Netzwerke mit $N = 22$ Knoten aus, um unseren Algorithmus für die gleiche Anzahl von Knoten bzw. biologische Klassen zu testen, wie sie später in den untersuchten Abundanzdaten vorhanden sind (siehe Abschnitt 3.7.3). Da diese Netzwerke einen viel größeren Zustandsraum haben ($2^N = 2^{22} = 4194304$) als Netzwerke mit 15 Knoten ($2^{15} = 32768$), erhöhen wir die maximale Anzahl der Anfangszustände, die verwendet werden, um die Attraktoren eines Netzwerks während des evolutionären Prozesses zu finden, auf $N_{\text{ini}} = 100000$. Um die Rechenzeit trotz dieser Änderung in einem vertretbaren Rahmen zu halten, reduzieren wir die Wiederholungen der Jaccard-Index-Berechnung auf $r = 10$, und modifizieren den evolutionären Algorithmus so, dass wir das fitteste Netzwerk immer beibehalten, d.h. ohne Mutation in die nächste Generation übernehmen. Die fitnessproportionale Selektion mit einer (möglichen) anschließenden Mutation erfolgt somit nur bei den anderen $M - 1$ Netzwerken der Population.
Wie in Abb. 3.11 zu sehen ist, führt die ESABO-gestützte Evolution auch bei diesen größeren Netzwerken zu besseren Ergebnissen als die einfache ESABO-Methode. Selbst nach einer kurzen Evolutionszeit von nur 2000 Generationen erzielt sie im Median höhere Jaccard-Index-Werte, d.h. Netzwerke, die den ursprünglichen Netzwerken topologisch ähnlicher sind als die Netzwerke, die mit der einfachen ESABO-Methode inferiert wurden. Eine längere Evolutionszeit würde mit hoher Wahrscheinlichkeit die Ergebnisse noch weiter verbessern, ist aber mit langen Rechenzeiten verbunden, insbesondere für den Fall, in dem alle Attraktoren zur Netzwerkrekonstruktion berücksichtigt werden.

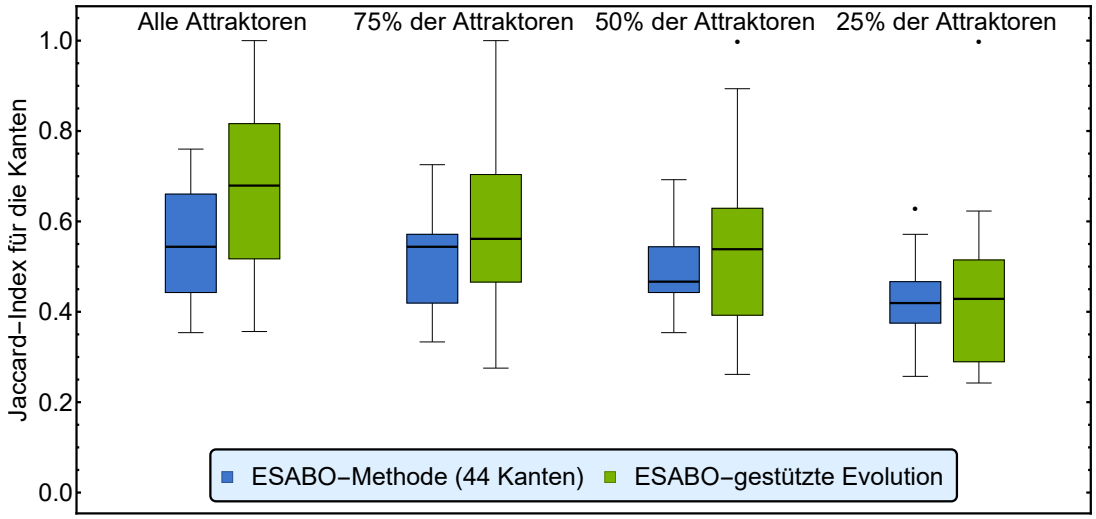


Abbildung 3.11.: Boxplots, die den Jaccard-Index zwischen den Kanten des ursprünglichen Netzwerks und den Kanten des inferierten oder fittesten evolvierten Netzwerks zeigen. Es wurden 25 Netzwerke mit $N = 22$ Knoten und $L_+ = L_- = 22$ positiven und negativen Kanten untersucht. Die Evolution wurde für 2000 Generationen mit $M = 50$, $\nu = 0.25$ und $L_{\text{min}} = 10$ durchgeführt. Bei der Netzwerkinferenz mit der ESABO-Methode wurden die 44 Kanten mit den höchsten ESABO-Score-Beträgen gesetzt. Auch bei diesen größeren Netzwerken führt die ESABO-gestützte Evolution zu besseren Ergebnissen als die einfache ESABO-Methode.

Da die Überlegenheit der ESABO-gestützten Evolution in diesem Fall allerdings nicht mehr ganz so stark ausgeprägt ist wie für die kleineren Netzwerke, wollen wir noch ein zweites Maß für die Qualität unserer Netzwerkrekonstruktion heranziehen, nämlich sog. *Receiver-Operating-Characteristics (ROC)-Diagramme* oder *ROC-Kurven*.

Einschub: ROC-Diagramme und -Kurven

ROC-Diagramme sind ursprünglich aus der Signaldetektionstheorie bekannt und werden u.a. in der Medizin zur Evaluierung diagnostischer Tests [150, 151, 163] oder im Bereich des maschinellen Lernens [126, 52] eingesetzt, um die Leistung von Klassifikatoren bzw. Klassifikationsverfahren zu visualisieren, zu bewerten und zu vergleichen.
Ein ROC-Diagramm ist im Allgemeinen ein zweidimensionales Diagramm, bei welchem die sog. *Richtig-Positiv-Rate (*Englisch*: true positive rate, TPR)*, die je nach Kontext auch als Sensitivität, Recall oder Trefferquote bezeichnet wird, auf der Y-Achse und die *Falsch-Positiv-Rate (*Englisch*: false positive rate, FPR)* auf der X-Achse aufgetragen wird [52]. Die Begriffe Richtig-Positiv-Rate und Falsch-Positiv-Rate gehen dabei auf ein binäres Klassifikationsproblem zurück, bei welchem ein untersuchtes Objekt lediglich einer sog. positiven Klasse oder einer negativen Klasse angehören kann. Folglich versteht man unter der Richtig-Positiv-Rate die Anzahl der richtig klassifizierten positiven Objekte geteilt durch die Gesamtzahl aller positiven Objekte und unter der Falsch-Positiv-Rate die Anzahl der negativen Objekte, die fälschlicherweise als positiv klassifiziert wurden, geteilt durch die Gesamtzahl aller negativen Objekte (siehe Abbildung 3.12).

		Tatsächliche Klasse: positiv	Tatsächliche Klasse: negativ
Vorhergesagte Klasse	positiv	richtig positiv (TP)	falsch positiv (FP)
	negativ	falsch negativ (FN)	richtig negativ (TN)

$TPR = TP / (TP+FN)$ $FPR = FP / (FP+TN)$

Abbildung 3.12.: Konfusionsmatrix für einen binären Klassifikator und Definition der Richtig-Positiv-Rate (*Englisch*: true positive rate, TPR) sowie der Falsch-Positiv-Rate (*Englisch*: false positive rate, FPR). Im Kontext eines medizinischen Test würde die vorhergesagte Klasse dem Testergebnis entsprechen (positiv oder negativ) und die tatsächliche Klasse würde angeben, ob der Patient tatsächlich krank ist (positive Klasse) oder nicht (negative Klasse).

Da die Vorhersage nahezu aller Klassifikatoren auf der Wahl eines bestimmten Schwellenwertes beruht, ergeben sich für unterschiedliche Wahlen des Schwellenwertes auch unterschiedliche Richtig-Positiv- und Falsch-Positiv-Raten. Trägt man diese nach dem Schwellenwert sortiert in ein ROC-Diagramm ein, so ergibt sich die sog. *ROC-Kurve* (siehe Abb. 3.14). Diese verbindet im ROC-Diagramm die Punkte (0,0), bei welchem der Klassifikator alle Objekte der negativen Klasse zuordnet, und (1,1), wo alle Objekte der positiven Klasse zugeordnet werden, und verläuft gewöhnlicherweise oberhalb der Winkelhalbierenden y = x, welche dem zufälligen Erraten einer Klasse entspricht. Die ROC-Kurve für einen perfekten Klassifikator verläuft vom Punkt (0,0) senkrecht hoch zu (0,1) und dann vom Punkt (0,1) waagerecht bis zum Punkt (1,1) [163].
Zur Behandlung von Klassifizierungsproblemen mit mehr als zwei Klassen besteht eine gängige Methode darin, für jede Klasse eine eigene ROC-Kurve zu erstellen. Die betrachtete Klasse wird hierbei als die positive Klasse aufgefasst und alle anderen Klassen werden zur negativen Klasse zusammengefasst. [52]
Um verschiedene Klassifikatoren anhand ihrer ROC-Kurven miteinander zu vergleichen, berechnet man meist die Fläche unter der ROC-Kurve (*Englisch*: area under the ROC curve, AUC) [26]. Diese hat immer einen Wert zwischen 0 und 1, wobei ein AUC-Wert von 0.5 der zufälligen Klassifikation entspricht und somit von jedem real genutzten Klassifikationsalgorithmus übertroffen werden sollte.
Weitere Details zur Interpretation von ROC-Kurven und effiziente Algorithmen, um solche Kurven zu erzeugen, können beispielsweise dem Übersichtsartikel von Tom Fawcett [52] entnommen werden, auf welchem die hier dargestellten Informationen beruhen.

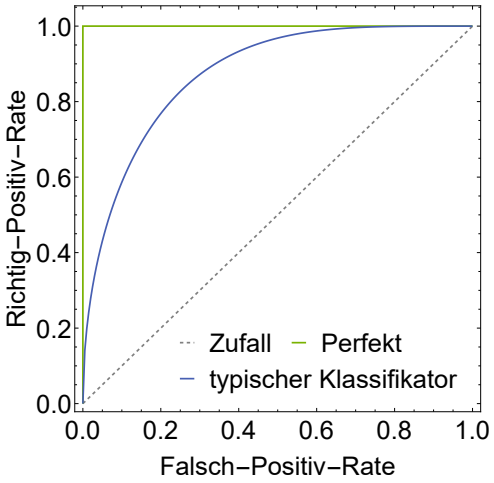


Abbildung 3.13.: Drei hypothetische ROC-Kurven nach [163]. Die grüne Kurve repräsentiert einen perfekten Klassifikator mit einem AUC-Wert von 1 und die grau gestrichelte Linie entspricht der zufälligen Klassifikation mit einem AUC-Wert von 0.5. Eine typische ROC-Kurve mit einem AUC-Wert von 0.87 ist in blau dargestellt. Wenn sich der Klassifikationsalgorithmus verbessert, bewegt sich die ROC-Kurve in Richtung der grünen Kurve und die Fläche unter der ROC-Kurve nähert sich dem Wert 1.

Um ROC-Diagramme zur Beurteilung der Erkennung positiver oder negativer Kanten sowie von Kanten im Allgemeinen (unabhängig von ihrem Vorzeichen) zu erstellen, berechnen wir zunächst für jedes (mithilfe aller Attraktoren) inferierte oder evolvierte Netzwerk aus Abbildung 3.11 die Richtig-Positiv-Rate (*Englisch*: true positive rate, TPR) und die Falsch-Positiv-Rate (*Englisch*: false positive rate, FPR) auf die nachfolgende Weise:

- Für die Erkennung von positiven Kanten gilt:

$$TPR_+ = \frac{n_{++}}{L_+}, \quad FPR_+ = \frac{n_{+-} + n_{+0}}{\frac{N \cdot (N-1)}{2} - L_+} \tag{3.7}$$

- Für die Erkennung negativer Kanten gilt:

$$TPR_- = \frac{n_{--}}{L_-}, \quad FPR_- = \frac{n_{-+} + n_{+0}}{\frac{N \cdot (N-1)}{2} - L_-} \tag{3.8}$$

- Für die Erkennung von Kanten im Allgemeinen gilt:

$$\begin{aligned} TPR_{\text{Kante}} &= \frac{n_{++} + n_{--} + n_{+-} + n_{-+}}{L_+ + L_-}, \\ FPR_{\text{Kante}} &= \frac{n_{+0} + n_{-0}}{\frac{N \cdot (N-1)}{2} - (L_+ + L_-)} \end{aligned} \tag{3.9}$$

L_+ ist hierbei die Anzahl der positiven Kanten und L_- die Anzahl der negativen Kanten im ursprünglichen Netzwerk. N gibt die Anzahl der Knoten an. In der Abkürzung n_{xy} steht y für die tatsächliche Art der Kante (positiv (+), negativ (-) oder keine (0)) und x für die vorhergesagte Beziehung. Somit ist n_{++} (bzw. n_{--}) die Anzahl der positiven (bzw. negativen) Kanten, die richtigerweise als positiv (bzw. negativ) klassifiziert wurden. n_{+-} (bzw. n_{-+}) ist die Anzahl der negativen (bzw. positiven) Kanten, die fälschlicherweise als positiv (bzw. negativ) eingestuft wurden, und n_{+0} (bzw. n_{-0}) ist die Anzahl der positiven (bzw. negativen) Kanten, die im evolvierten oder inferierten Netzwerk vorhanden waren, obwohl es im ursprünglichen Netzwerk keine entsprechende Kante gab.

Anschließend erstellen wir für die (einfache) ESABO-Methode ROC-Kurven mithilfe des von Tom Fawcett in [52] beschriebenen Algorithmus. Hierzu fassen wir die Vorhersagen für die Kanten aller untersuchten Netzwerke zu einem großen Datensatz zusammen und ordnen sie nach ihrem ESABO-Score. Zur Beurteilung der Erkennung positiver Kanten sortieren wir die Kanten dabei nach absteigendem ESABO-Score Z und setzen nur positive Kanten. Zur Erstellung der ROC-Kurve für die Erkennung negativer Kanten sortieren wir alle Kanten nach aufsteigendem ESABO-Score und setzen nur negative Kanten. Für die Erkennung von Kanten im Allgemeinen (unabhängig von ihrem richtigen Vorzeichen) werden die Kantenvorhersagen nach absteigendem ESABO-Score-Betrag $|Z|$ sortiert.

Die ROC-Kurven, die sich für die einfache ESABO-Methode ergeben, und die Untersuchung der Richtig-Positiv-Raten sowie der Falsch-Positiv-Raten der evolvierten oder inferierten Netzwerke sind in Abbildung 3.14 gezeigt und bestätigen die Überlegenheit der ESABO-gestützten Evolution gegenüber der einfachen ESABO-Methode.
Wie man sieht, weist bereits die einfache ESABO-Methode mit AUC-Werten von 0.95 bzw. 0.93 eine gute Inferenzqualität für die Erkennung von positiven bzw. negativen Kanten auf. Bei der Unterscheidung, ob eine Kante überhaupt vorhanden ist oder nicht, unabhängig von ihrem Vorzeichen ist sie jedoch weniger erfolgreich (AUC-Wert von 0.89 für die Erkennung von Kanten im Allgemeinen). Dies spiegelt sich auch in der Tatsache wider, dass Netzwerke, die mit der einfachen ESABO-Methode inferiert wurden, indem alle Kanten mit einem ESABO-Score $|Z| > 1$ gesetzt wurden, wie in [34] vorgeschlagen, in der Regel eine TPR nahe 1, aber eine relativ hohe FPR (> 0.5 für die Erkennung von Kanten) haben. Wenn wir nur die 44 Verbindungen mit den höchsten ESABO-Score-Beträgen setzen, sinkt die FPR beträchtlich, aber auch die TPR sinkt, und - was entscheidend ist - wir haben unser Vorwissen über die Anzahl der im ursprünglichen Netzwerk vorhandenen Kanten verwendet. Die ESABO-gestützte Evolution führt im Allgemeinen zu höheren Richtig-Positiv-Raten bei vergleichbaren Falsch-Positiv-Raten wie die einfache ESABO-Methode, bei der die 44 Kanten mit den höchsten ESABO-Score-Beträgen gesetzt wurden. Zudem wird dieses Ergebnis ohne Vorwissen über die Anzahl der im ursprünglichen Netzwerk vorhandenen Kanten erreicht.

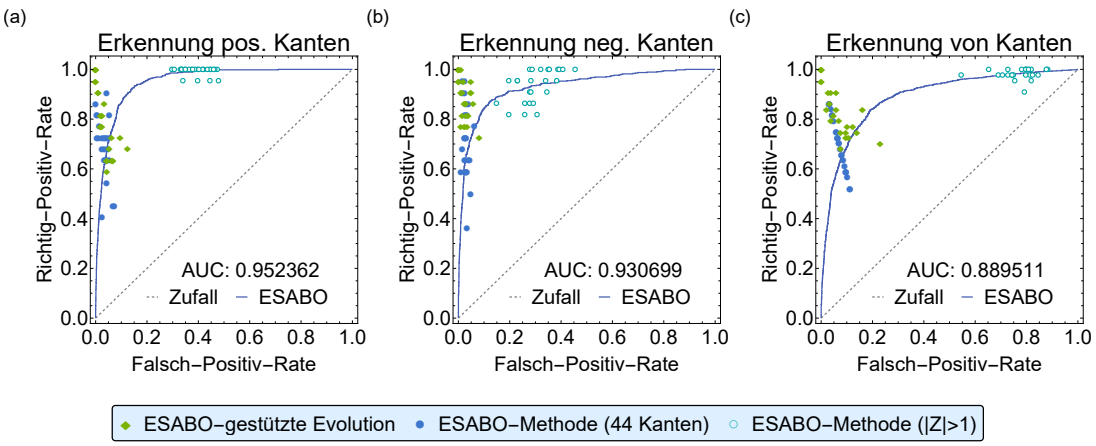


Abbildung 3.14.: Vergleich der Inferenzqualität zwischen der ESABO-gestützten Evolution und der ursprünglichen ESABO-Methode anhand von Receiver-Operating-Characteristics (ROC)-Diagrammen bzw. -Kurven. Um die ROC-Kurven für die ESABO-Methode zu erstellen, haben wir die Kanten-Vorhersagen (unter Berücksichtigung aller Attraktoren) für alle 25 untersuchten Netzwerke aus Abb. 3.11 zu einem großen Datensatz zusammengefasst und sie nach ihrem ESABO-Score geordnet. Außerdem haben wir die Richtig-Positiv-Rate und die Falsch-Positiv-Rate für Netzwerke ausgewertet, die entweder mit der ESABO-gestützten Evolution oder mit der einfachen ESABO-Methode inferiert wurden. Für Netzwerke, die mit der einfachen ESABO-Methode inferiert wurden, haben wir entweder die 44 Kanten mit den höchsten ESABO-Score-Beträgen oder alle Kanten mit einem ESABO-Score $|Z| > 1$ gesetzt. Der AUC-Wert gibt die Fläche unter der ROC-Kurve an.

Abbildung 3.15 zeigt die Fitness der evolvierten Netzwerke im Vergleich zur Fitness zufällig zusammengewürfelter Attraktoren, die nicht zu einem tatsächlichen Booleschen Netzwerk gehören. Obwohl die maximale Fitness, die während der Evolution erreicht wird, erwartungsgemäß mit abnehmendem Prozentsatz der für die Rekonstruktion verwendeten Attraktoren abnimmt, sind die erreichten Fitnesswerte ($F_{\text{Median}} \approx 0.33$) selbst in dem Fall, in dem nur 25% der ursprünglichen Attraktoren als Input für die Inferenzmethode verwendet wurden, mindestens eine Größenordnung größer als die der zufällig erzeugten Attraktoren ($F_{\text{Median}} \approx 0.02$). Zum einen bedeutet dies, dass die ESABO-gestützte Evolution auch im Falle sehr unvollständiger Datensätze die darin enthaltenen Informationen über das zugrunde liegende Interaktionsnetzwerk detektieren kann. Zum anderen können wir den hier gefundenen Zusammenhang zwischen dem Prozentsatz der zur Netzwerkrekonstruktion verwendeten Attraktoren und den erreichten Fitnesswerten nutzen, um grob abzuschätzen wie vollständig der in Kapitel 3.7.3 untersuchte biologische Datensatz für das menschliche Speichelmikrobiom ist.

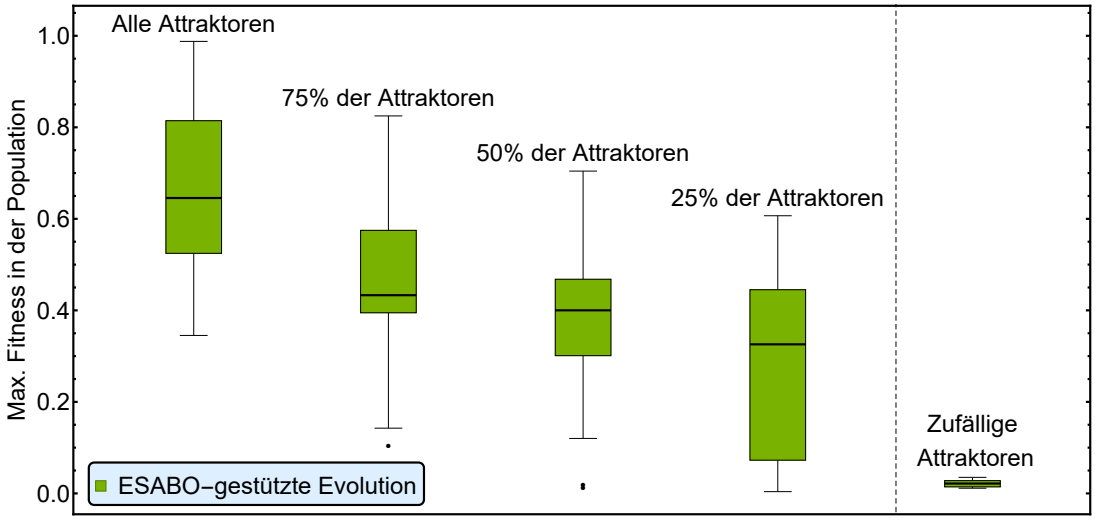


Abbildung 3.15.: Boxplots für die Fitness der evolvierten Netzwerke aus Abb. 3.11 verglichen mit der Fitness von 25 Netzwerken, die jeweils aus 138 verschiedenen zufällig zusammengewürfelten Attraktoren inferiert wurden. Um einen zufälligen Attraktor zu erzeugen, wurde für jeden seiner 22 Einträge zufällig der Wert 1 oder 0 mit gleicher Wahrscheinlichkeit ausgewürfelt.
Die maximale Fitness, die nach der Evolution erreicht wird, nimmt erwartungsgemäß mit abnehmendem Prozentsatz der für die Rekonstruktion verwendeten Attraktoren ab. Sie ist jedoch immer deutlich größer als die Fitness von Netzwerken, die ausgehend von zufällig zusammengewürfelten Attraktoren inferiert wurden.

3.7.3. Anwendung auf echte biologische Daten: Analyse des menschlichen Speichelmikrobioms

Im letzten Teil dieses Kapitels wenden wir die ESABO-Methode nun auf echte biologische Abundanzdaten an, die im Rahmen des Human-Microbiome-Projects (HMPs) mittels 16S-rRNA-Gensequenzierung gewonnen wurden. Diese Daten wurden von Mitgliedern des Human-Microbiome-Project-Cosortiums mit dem Softwarepaket QIIME (Quantitative Insights Into Microbial Ecology) [30] verarbeitet, und die daraus resultierenden operativen taxonomischen Einheiten (*Englisch:* operational taxonomic units, OTUs) bzw. Abundanzdaten wurden unter `https://www.hmpdacc.org/hmp/HMQCP/` zur Verfügung gestellt. Ausführliche Informationen zur Gewinnung dieser Daten können der dazugehörigen Publikation [153] entnommen werden.
Im Folgenden betrachten wir lediglich Speichelproben, da man davon ausgehen kann, dass das Speichelmikrobiom eines erwachsenen Menschen im Laufe der Zeit relativ stabil ist [19, 95] und daher als ein Attraktorzustand bzw. Fixpunkt interpretiert werden kann. Zudem werten wir hier nur die Daten aus den variablen Regionen 3-5 (V35) des 16S-rRNA-Gens aus (da für diesen Bereich des Gens der größte Datensatz vorliegt) und analysieren die Abundanzdaten auf Klassenebene, d.h. OTUs, die zur selben biologischen Klasse[4] zugeordnet werden können, werden zu einer Gruppe zusammengefasst[5]. Als Binarisierungsschwellenwert wird der Wert 1 verwendet, d.h. eine Spezies bzw. Klasse, die in einer betrachteten Probe enthalten ist, erhält unabhängig von ihrer gemessenen Häufigkeit den Wert 1 und eine Klasse, die nicht enthalten ist, den Wert 0. Klassen, die in jeder der Proben vorkommen, werden nicht berücksichtigt, da unsere Methode eine gewisse Variation in der Anwesenheit einer Spezies erfordert, um eine Wechselwirkung für diese vorhersagen zu können.
Die Ergebnisse dieser Untersuchung sind in Abbildung 3.16 dargestellt.

Wie wir sehen können, führt die ESABO-gestützte Evolution zu einer relativ großen Zunahme der Fitness gegenüber der einfachen ESABO-Methode (Abb. 3.16 (a)). Während das fitteste Netzwerk, das mit der einfachen ESABO-Methode inferiert wurde (vgl. Generation 0 der Evolution), nur eine Fitness von $F \approx 0.01$ hat, hat das evolvierte Netzwerk eine Fitness von $F \approx 0.27$ (nach einer Evolutionsdauer von 3000 Generationen). Darüber hinaus ist die Fitness des evolvierten Netzwerks, das aus realen biologischen Daten (138 verschiedene binäre Proben) inferiert wurde, deutlich höher als die Fitness von Netzwerken, die aus zufällig zusammengewürfelten Attraktoren, die nicht zu einem tatsächlichen Booleschen Netzwerk gehören, inferiert wurden. Diese Netzwerke erreichen nach einer Evolution von 3000 Generationen lediglich eine Fitness von $F_{\text{Median}} \approx 0.02$. Das bedeutet, dass unsere Inferenzmethode erkennt, dass die biologischen Abundanzmuster nicht zufällig sind, sondern tatsächlich einem zugrunde liegenden Interaktionsnetzwerk zugeordnet werden können.

[4] Organismen werden in der Biologie anhand gemeinsamer Merkmale in Taxa genannte Kategorien eingeteilt. Die Taxa sind dabei hierarchisch gegliedert. Von der umfassendsten zur spezifischsten Kategorie lautet die Reihenfolge [58]:
Domäne (Bakterien, Archaeen oder Eukaryoten) -
Reich (existiert nur im Falle von Eukaryoten; Bakterien und Archaeen werden direkt in sog. Phyla eingeteilt) -
Phylum bzw. Stamm - **Klasse** - Ordnung - Familie - Gattung - Art

[5] Hierzu wurde das R-Package phyloseq [112] verwendet.

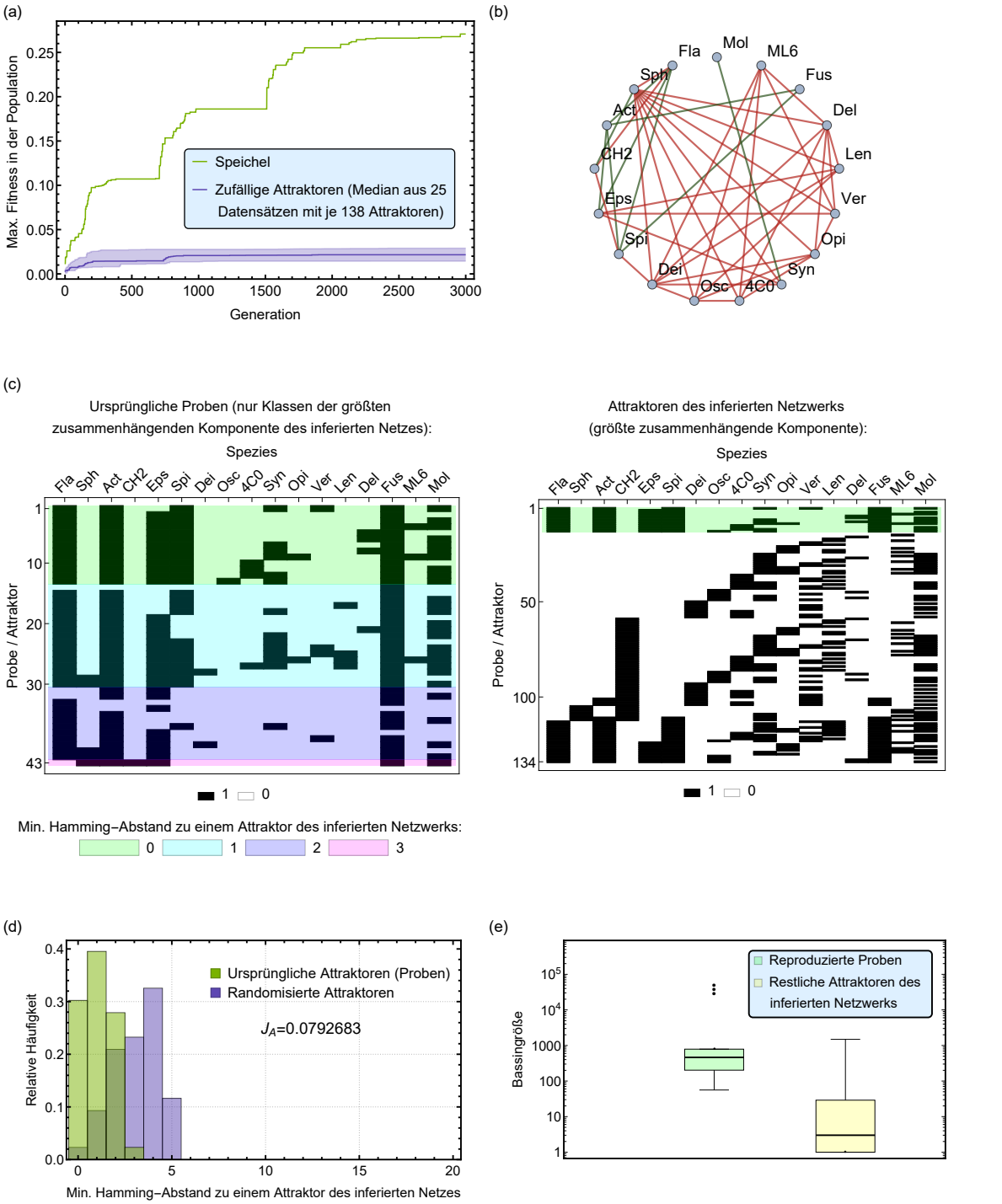

(a)
Max. Fitness in der Population
0.25
0.20
0.15
0.10
0.05
0.00
0
500
1000
1500
2000
2500
3000
Generation
Speichel
Zufällige Attraktoren (Median aus 25 Datensätzen mit je 138 Attraktoren)
(b)
Fla
Mol
ML6
Sph
Fus
Act
Del
CH2
Len
Eps
Ver
Spi
Opi
Dei
Syn
Osc
4C0
(c)
Ursprüngliche Proben (nur Klassen der größten zusammenhängenden Komponente des inferierten Netzes):
Spezies
Fla Sph Act CH2 Eps Spi Dei Osc 4C0 Syn Opi Ver Len Del Fus ML6 Mol
Probe / Attraktor
1
10
20
30
43
1 0
Attraktoren des inferierten Netzwerks (größte zusammenhängende Komponente):
Spezies
Fla Sph Act CH2 Eps Spi Dei Osc 4C0 Syn Opi Ver Len Del Fus ML6 Mol
Probe / Attraktor
1
50
100
134
1 0
Min. Hamming−Abstand zu einem Attraktor des inferierten Netzwerks:
0 1 2 3
(d)
Relative Häufigkeit
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
Min. Hamming−Abstand zu einem Attraktor des inferierten Netzes
Ursprüngliche Attraktoren (Proben)
Randomisierte Attraktoren
J_A=0.0792683
(e)
Bassingröße
10^5
10^4
1000
100
10
1
Reproduzierte Proben
Restliche Attraktoren des inferierten Netzwerks

Abbildung 3.16.: Analyse des menschlichen Speichelmikrobioms
(a) Entwicklung der maximalen Fitness für das rekonstruierte biologische Netzwerk im Vergleich zur Entwicklung der Median-Fitness von 25 Netzwerken, die aus 138 zufälligen Attraktoren inferiert wurden, welche nicht zu einem tatsächlichen Booleschen Netzwerk gehören (vgl. Abb. 3.15). Der schattierte Bereich zeigt den Bereich zwischen dem 10%- und dem 90%-Quantil an. Man erkennt, dass die Fitness bei der Inferenz des echten biologischen Netzwerks im Laufe der ESABO-gestützten Evolution deutlich ansteigt, während die Fitness von Netzwerken, die aus zufällig zusammengewürfelten Attraktoren rekonstruiert wurden, kaum zunimmt.
(b) Größte zusammenhängende Komponente des rekonstruierten Speichelnetzwerks (fittestes Netzwerk, das nach einer Evolution von 3000 Generationen gefunden wurde). Die anderen 5 Knoten, die nicht Teil dieser Komponente sind, sind überhaupt nicht verbunden, d.h. sie haben keine Kanten zu anderen Knoten. Abkürzungen: Fla: *Flavobacteria*, Sph: *Sphingobacteria*, Act: *Actinobacteria*, CH2: *CH21*, Eps: *Epsilonproteobacteria*, Spi: *Spirochaetes*, Dei: *Deinococci*, Osc: *Oscillatoriophycideae*, 4C0: *4C0d-2*, Syn: *Synergistia*, Opi: *Opitutae*, Ver: *Verrucomicrobiae*, Len: *Lentisphaerae*, Del: *Deltaproteobacteria*, Fus: *Fusobacteria*, ML6: *ML615J-28*, Mol: *Mollicutes*.
(c) Vergleich der ursprünglichen, binarisierten Abundanzdaten bzw. Proben (links) mit den Attraktoren des rekonstruierten Netzwerks (rechts). In beiden Fällen wurden nur die Klassen berücksichtigt, die in der größten zusammenhängenden Komponente des rekonstruierten Netzwerks vorhanden sind. Für jede Probe ist die minimale Hamming-Distanz zu einem Attraktor des rekonstruierten Netzwerks durch die entsprechende Farbe angegeben. 13 (grün markiert) der 43 Proben werden als Attraktoren des rekonstruierten Netzwerks reproduziert.
(d) Histogramm, das die minimalen Hamming-Distanzen der ursprünglichen Attraktoren (Proben) oder der randomisierten Versionen dieser Attraktoren zu den Attraktoren des rekonstruierten Netzwerks zeigt. Die Randomisierung eines Attraktors wurde durch eine Permutation seiner Einträge durchgeführt. J_A gibt den Jaccard-Index zwischen den Attraktoren des rekonstruierten Netzwerks und den ursprünglichen Proben aus (c) an. Man sieht, dass die echten biologischen Proben (grün) in der Regel einen deutlich kleineren Hamming-Abstand ($h \leq 3$) zu den Attraktoren des rekonstruierten Netzwerks aufweisen als die randomisierten Proben (lila).
(e) Verteilung der Bassingrößen der Attraktoren des inferierten Netzwerks. In grün sind die Bassingrößen der 13 Proben gezeigt, die als Attraktoren des inferierten Netzwerks reproduziert werden. In gelb sind die restlichen Attraktoren des inferierten Netzwerks dargestellt, d.h. diejenigen Attraktoren, die nicht exakt einer binarisierten biologischen Probe entsprechen. Man erkennt, dass die 13 Attraktoren, die einer Mikrobiomprobe entsprechen, im Median eine deutlich höhere Bassingröße aufweisen als die anderen 121 Attraktoren des Netzwerks. (Logarithmische Skala beachten!)

Die größte zusammenhängende Komponente dieses geschätzten Interaktionsnetzwerks, das nach einer Evolution von 3000 Generationen gefunden wurde, ist in Abb. 3.16 (b) dargestellt. Wir betrachten im Folgenden nur diese Komponente und nicht das gesamte inferierte Netzwerk, da die restlichen 5 Netzwerkknoten, die nicht Teil der größten zusammenhängenden Komponente sind, alle gänzlich unverbundenen sind (siehe Anhang A.2).
Wie wir sehen können, besitzt das inferierte Netzwerk bzw. seine größte zusammenhängende Komponente wesentlich mehr negative Kanten (34 negative Kanten) als positive Kanten (9 positive Kanten). Obwohl es interessant erscheint, dieses Netzwerk und insbesondere das Verhältnis von negativen zu positiven Kanten mit anderen Untersuchungen für das Speichelmikrobiom, bei denen ebenfalls ein Interaktionsnetzwerk inferiert wurde, zu vergleichen, ist ein solcher Vergleich nicht besonders zielführend. Dies liegt vor allem daran, dass sich unsere Analyse auf die seltenen Bakterienklassen konzentriert (wir haben die Klassen *Bacilli*, *Bacteroidia*, *Betaproteobacteria*, *Clostridia* und *Gammaproteobacteria* nicht berücksichtigt, da sie in jeder der Proben vorkamen), während viele andere Studien (wie z.B. [51]) hauptsächlich Interaktionen zwischen besonders häufig vorkommenden Phyla oder Klassen vorhersagen (siehe auch die Diskussion in [34]).

Stattdessen betrachten wir die dynamischen Eigenschaften bzw. die Attraktoren des resultierenden Netzwerks genauer. Wenn wir die Attraktoren des ursprünglichen Netzwerks, d.h. die binarisierten Proben, mit den Attraktoren des rekonstruierten Netzwerks vergleichen, stellen wir fest, dass sie einander sehr ähnlich sind. So werden 13 der 43 ursprünglichen Attraktoren exakt rekonstruiert und auch die übrigen 30 Proben haben einen sehr geringen Hamming-Abstand[6] von $h \leq 3$ zu den Attraktoren des rekonstruierten Netzwerks (siehe Abb. 3.16 (c)-(d)). Die Betrachtung der Verteilung dieser Hamming-Abstände in Abb. 3.16 (d) bestätigt zudem, dass sie für die echten biologischen Proben in der Regel deutlich kleiner sind (Maximum der Verteilung bei $h = 1$) als die Hamming-Abstände, die man bei vergleichbaren Attraktoren dieser Länge (hier durch eine Randomisierung der Proben erzeugt) erwarten würde (Maximum der Verteilung bei $h = 4$).
Ein Vergleich von Abbildung 3.16 (a) mit Abbildung 3.15 legt des Weiteren nahe, dass die derzeit verfügbaren Daten weniger als 50 Prozent der Attraktoren des Systems abdecken. Dies wird insbesondere durch die Tatsache bekräftigt, dass das rekonstruierte Netzwerk mit insgesamt 134 Attraktoren etwa 3-mal so viele Attraktoren besitzt wie der gemessene biologische Datensatz (43 unterschiedliche binarisierte Proben) und unsere Inferenzmethode somit viele zusätzliche, bisher experimentell noch nicht beobachtete Attraktoren voraussagt. Der geringe Anteil an gemessenen Attraktoren im biologischen Datensatz könnte dabei dadurch erklärt werden, dass vermutlich in der Regel nur Attraktoren mit einer großen Bassingröße experimentell beobachtet werden konnten, während Attraktoren mit einer geringen Bassingröße weitestgehend unbeobachtet blieben. Um diese Vermutung zu überprüfen, betrachten wir daher in Abbildung 3.16 (e) die Bassingrößen der Attraktoren des rekonstruierten Netzwerks.

[6] Der Hamming-Abstand [69] ist definiert als die Anzahl der Stellen, an denen sich zwei gleich lange (binäre) Vektoren unterscheiden. Die Vektoren $\vec{a_1} = \{0, 0, 1, 1\}$ und $\vec{a_2} = \{1, 1, 1, 1\}$ haben beispielsweise einen Hamming-Abstand von $h = 2$.

In der Tat zeigt sich, dass diejenigen Attraktoren des Netzwerks, die einer binarisierten Mikrobiomprobe entsprechen, eine signifikant höhere Bassingröße aufweisen als andere Attraktoren des Netzwerks. Während die 13 Attraktoren, die jeweils exakt einer gemessenen Probe entsprechen, nämlich im Median eine hohe Bassingröße von 463 Zuständen aufweisen, haben alle anderen Attraktoren im Median lediglich eine Bassingröße von 3. Zudem gehören auch die drei Attraktoren mit den am Abstand größten Bassins (Bassingröße von 49969, 39402 und 30285 Zuständen) zur Gruppe der reproduzierten biologischen Proben. Insgesamt umfassen die Bassins dieser 13 reproduzierten Proben 123264 Zustände und entsprechen somit etwa 94% aller $2^{17} = 131072$ Zustände im Zustandsraum des inferierten Netzwerks. Dies bedeutet, dass diese 13 Attraktoren, die im Rahmen der ESABO-gestützten Evolution reproduziert werden konnten, die Dynamik des rekonstruierten Netzwerkes entscheidend bestimmen. Insgesamt weist unsere Untersuchung zu den Bassingrößen der Netzwerkattraktoren auf eine hohe Plausibilität der hier durchgeführten Abschätzung zur Vollständigkeit des biologischen Datensatzes hin.

3.8. Fazit und Diskussion

In diesem Kapitel haben wir aufbauend auf der von Claussen et al. [34] eingeführten ESABO-Methode eine neue evolutionsbasierte Methode zur Inferenz mikrobieller Interaktionsnetzwerke, die ESABO-gestützte Evolution, entwickelt. Diese Methode verfolgt den simplen Ansatz, dass ein inferiertes Netzwerk die ursprünglichen, zur Inferenz genutzten binären Mikrobiomdaten – welche von uns als Fixpunkte der Dynamik interpretiert werden – als seine Attraktoren reproduzieren sollte. Um dies zu erreichen, wird im Rahmen der ESABO-gestützten Evolution das inferierte Netzwerk (mittels eines evolutionären Algorithmus) so optimiert, dass der Überlapp seiner Attraktoren mit den zur Inferenz verwendeten binarisierten Mikrobiomdaten maximiert wird. Auf diese Weise erreichen wir zwei Ziele, die für eine formale Beschreibung des Mikrobioms von grundlegender Bedeutung sind:

(1) Wir inferieren Netzwerke, die bereits aufgrund ihrer Konstruktion in der Lage sind, die experimentellen Daten auf einer binären Ebene zu reproduzieren.

(2) Wir ermitteln, wie diese evolutionäre Inferenzmethode durch unvollständige Informationen bezüglich der möglichen stabilen Mikrobiomzustände, d.h. wenn nur ein bestimmter Prozentsatz aller Netzwerkattraktoren experimentell beobachtet wurde, beeinflusst wird.

Die ursprüngliche ESABO-Methode, die durch die in Kapitel 3.5 beschriebenen Modifikationen (analytische Berechnung der z-Scores und Vertauschen von Nullen und Einsen in Abundanzvektorpaaren mit einem hohen Anteil an Einsen) deutlich verbessert wurde, fließt auf zwei Arten in den hier implementierten evolutionären Algorithmus ein: Das ESABO-Netzwerk dient zum einen als Ausgangspunkt für die Evolution und zum anderen dienen die ESABO-Scores zur Festlegung des Vorzeichens der Netzwerkkanten (positiv oder negativ) und erlauben eine Priorisierung der zu setzenden Kanten während der Evolution. Hierdurch wird die simulierte Evolution deutlich beschleunigt.

Ein Nachteil, der sich durch dieses Vorgehen ergibt, ist, dass durch eine falsche Vorhersage des Vorzeichens einer Kante bzw. des Interaktionstyps im Rahmen der ESABO-Methode

auch mittels der nachfolgenden Evolution das ursprüngliche Netzwerk nicht mehr perfekt rekonstruiert werden kann. Da die falsche Zuordnung des Vorzeichens einer Kante aber vergleichsweise selten passiert, überwiegt unseres Erachtens der Vorteil einer schnelleren Fitnesszunahme während der Evolution gegenüber dem Nachteil, dass eine Fitness von $F = 1$ in seltenen Fällen eventuell niemals erreicht werden kann. Eine mögliche Lösung für dieses Problem könnte in Zukunft darin bestehen, das für eine Kante (durch die ESABO-Methode) vorhergesagte Vorzeichen mit einer gewissen Wahrscheinlichkeit, z.B. proportional zu $1/(2 + |Z_{ij}|)$, zu vertauschen.

Der Fokus auf die Informationen, die in den An- und Abwesenheitsmustern der mikrobiellen Spezies enthalten sind, ist nicht als Alternative, sondern vielmehr als Ergänzung zu den auf kontinuierlichen Abundanzen basierenden Inferenzmethoden gedacht: Wie in der Einleitung dargelegt, weisen diese beiden Informationsebenen in einem Mikrobiom-Datensatz ganz unterschiedliche systemische Eigenschaften auf. Basierend auf den in unserer Untersuchung durchgeführten numerischen Experimenten und der Diskussion der ESABO-Methode in [34] glauben wir, dass die Boolesche Perspektive seltene Mikroorganismen und ihren Beitrag zur mikrobiellen Gemeinschaft sowie die *intrinsischen* Interaktionen zwischen Mikroorganismen hervorhebt, während die Abundanzperspektive einen stärkeren Schwerpunkt auf die dominanten Mikroorganismen legt (mit *Firmicutes* und *Bacteroidetes* als prominente Beispiele; siehe [108]) und ein zuverlässigerer Indikator für *externe* Stimuli ist, die große Teile der Gemeinschaft beeinflussen.

Die binäre Betrachtung der Mikrobiomzusammensetzung (d.h. die Betrachtung der An- und Abwesenheitsmuster von mikrobiellen Spezies) hat insbesondere den Vorteil, dass sie es uns – unter der Annahme, dass diese binären Zustände die Attraktoren bzw. Fixpunkte eines Booleschen Netzwerks darstellen – ermöglicht, das mikrobielle Interaktionsnetzwerk mit Mikrobiomzuständen (Attraktoren) auf eine im Wesentlichen parameterfreie Weise in Beziehung zu setzen.

Zudem trägt unser Ansatz im Gegensatz zu vielen bisherigen Inferenzmethoden, die lediglich strukturelle Netzwerkeigenschaften berücksichtigen (siehe z.B. [158] oder [123]), der Tatsache Rechnung, dass es sich beim Mikrobiom um ein dynamisches Netzwerk handelt und dass die gemessenen Daten stationäre Zustände eines solchen Netzwerks darstellen sollten. Diese eher neue Sicht auf Mikrobiomdaten wird beispielsweise auch in [161] beschrieben.

Obwohl wir unsere Ergebnisse unter Verwendung ganz bestimmter Boolescher Aktualisierungsfunktionen (siehe Gleichung 3.2) erhalten haben, können prinzipiell auch andere Boolesche Schwellenwertfunktionen, die explizit den aktuellen Knotenwert ($s_i(t)$) enthalten, wie z.B.

$$s_i(t+1) = \begin{cases} 1, & \sum_{j=1}^{N} G_{ij} s_j(t) > 0 \\ s_i(t), & \sum_{j=1}^{N} G_{ij} s_j(t) \leq 0 \end{cases} \tag{3.10}$$

gewählt werden und führen zu ähnlichen Ergebnissen (siehe Appendix A.3).

Die meisten realen Datensätze sind unvollständig, d.h. sie enthalten nicht alle möglichen Attraktoren des zugrunde liegenden Interaktionsnetzwerks. Der Prozentsatz der verfügbaren Attraktoren ist dabei im Allgemeinen vollkommen unbekannt. Im Rahmen der

ESABO-gestützten Evolution finden wir eine Beziehung zwischen dem Prozentsatz der bekannten Attraktoren und der durchschnittlichen Fitness, die am Ende der Evolution erreicht wird (siehe Abbildung 3.15 und Appendix A.1). Dies deutet auf die Möglichkeit hin, die Vollständigkeit eines Satzes von mikrobiellen Abundanzmustern zu schätzen.
Sobald mehr und mehr mikrobielle Abundanzdaten zur Verfügung stehen, wird die Einschränkung der Inferenzqualität durch unvollständige Datensätze weniger schwerwiegend sein, aber selbst jetzt kann unsere Methode, eine grobe Abschätzung des Vollständigkeitsgrades der verwendeten Daten liefern.

Bei der überwiegenden Mehrheit der gegenwärtig verfügbaren Datensätze handelt es sich zudem lediglich um Abundanz-„Schnappschüsse“ bzw. -„Momentaufnahmen“. Mit der Verfügbarkeit von mehr und mehr zeitabhängigen Daten bzw. von Zeitreihen solcher Abundanzmuster können ausgefeiltere Methoden der Netzwerkinferenz entwickelt werden (z.B. dynamische oder zeitvariable Netzwerkmodellierung; siehe [60]).

Unsere Ergebnisse legen nahe, die Veränderung von Attraktoren infolge kleiner Variationen des zugrunde liegenden Interaktionsnetzwerks genauer zu untersuchen, da ein tieferes Verständnis dieser Beziehung zur Entwicklung besserer Inferenzalgorithmen beitragen kann.

Wie schon in der Einleitung zur Dissertation dargelegt wurde, stellt die Inferenz mikrobieller Interaktionsnetzwerke im Allgemeinen nur den Ausgangspunkt für weitere Untersuchungen der Netzwerktopologie und -dynamik dar. In diesem Zusammenhang könnte es beispielsweise interessant sein, die Robustheit der rekonstruierten Netzwerke gegenüber externen Störungen der Mikrobiomzusammensetzung zu untersuchen.

4. Modellierung der zellzyklusabhängigen Regulation von p21 durch p53 nach DNA-Doppelstrangbrüchen

4.1. Einleitung: p53 und p21

Die Zellen unseres Körpers befinden sich in einer komplexen Umgebung und müssen daher ständig auf wechselnde Umweltbedingungen und verschiedene Formen von Stress reagieren. Eine der wichtigsten Ursachen für zellulären Stress ist dabei ionisierende Strahlung, die häufig DNA-Doppelstrangbrüche (DSBs) verursacht und somit die erfolgreiche Teilung oder sogar das Überleben der bestrahlten Zelle gefährdet.

Das Tumorsupressorprotein p53, welches auch als „Wächter des Genoms“ bezeichnet wird [90], ist eines der wichtigsten Bestandteile der zellulären Antwort auf Stress oder DNA-Schäden. Wie in Abbildung 4.1 gezeigt, wird es nach einem DSB durch die Proteinkinasen ATM und Chk2 aktiviert, welche p53 phosphorylieren und auf diese Weise seine Affinität zur E3-Ubiquitin-Ligase Mdm2 reduzieren. Mdm2 sorgt normalerweise in unbeschädigten Zellen dafür, dass p53 ubiquitiniert und durch das Proteasom abgebaut wird [71, 118]. Wird die Bindung von p53 und Mdm2 gehemmt, ist der p53-Abbau verlangsamt und die p53-Konzentration im Zellkern steigt an.
Nachdem p53 aktiviert wurde, induziert es als Transkriptionsfaktor, d.h. indem es an die Promotoren seiner Zielgene bindet, die Expression von Hunderten von Genen [55]. Die von diesen Genen kodierten Proteine sind an den unterschiedlichsten biologischen Prozessen beteiligt, die von der DNA-Reparatur, über die Zellzykluskontrolle bis hin zur Apoptose reichen, wodurch p53 einen maßgeblichen Einfluss auf das Schicksal der Zelle nach einem DNA-Schaden hat [68, 81]. Da einige der von p53 aktivierten Proteine, wie z.B. das bereits erwähnte Mdm2 [14] oder die Phosphatase Wip1 [54], wiederum negative p53-Regulatoren sind, entstehen mehrere Feedbackschleifen im p53-Netzwerk. Diese Feedbackschleifen ermöglichen – vermutlich in Kombination mit weiteren im Netzwerk vorhandenen positiven Rückkopplungen [119] – eine äußerst komplexe p53-Dynamik. Während unter basalen Bedingungen nur vereinzelte, unregelmäßige p53-Pulse beobachtet werden [102], treten nach der Induktion von Doppelstrangbrüchen mittels ionisierender Strahlung mehrere gleichmäßige p53-Pulse bzw. p53-Oszillationen auf, deren Amplitude und Dauer während der gesamten Schadensantwort konstant bleiben und von der Strahlendosis unabhängig sind [89, 17, 102, 127, 56]. Lediglich die Anzahl der p53-Pulse nimmt mit zunehmender Strahlendosis zu, weshalb von einem „digitalen Charakter“ der p53-Antwort nach der Induktion von DSBs durch ionisierende Strahlung gesprochen wird [89].

Eines der wichtigsten Zielgene von p53 ist CDKN1A, welches das Protein p21 kodiert [41]. Bei p21 handelt es sich um einen Cdk-Inhibitor, der nach einer DNA-Schädigung für das Anhalten des Zellzyklus in der G1-Phase entscheidend ist. p21 bindet vor allem Cyclin-Cdk-Komplexe, die für den Beginn und das Durchlaufen der S-Phase nötig sind (u.a. CyclinE-Cdk2-Komplexe), und hemmt deren Kinaseaktivität [43]. Dadurch wird der Eintritt der Zelle in die S-Phase verhindert und sie erhält die notwendige Zeit, um die entstandenen DNA-Schäden zu reparieren. Neben dem Anhalten des Zellzyklus in der G1-Phase kann p21 durch die Hemmung der cyclin-abhängigen Kinase Cdk1 auch zum Zellzyklusstopp in der G2-Phase beitragen [29]. Zudem spielt p21 eine wesentliche Rolle bei der Verhinderung der Endoreduplikation[1] während eines längeren Zellzyklusstopps [155] sowie für die Induktion der zellulären Seneszenz[2] [122]. Schließlich trägt p21 auch zur Inhibierung der DNA-Replikation bei, indem es das Protein PCNA bindet. PCNA ist eine sog. *Gleitklammer (Englisch: sliding clamp)*, d.h. ein Protein, welches während der DNA-Replikation einen gleitenden Ring um die DNA bildet und benötigt wird, um DNA-Polymerasen an der DNA festzuhalten. Durch die Bindung von p21 an PCNA wird die Interaktion von PCNA mit den DNA-Polymerasen verhindert [106], wodurch die DNA-Replikation inhibiert wird.

Die p21-Konzentration in einer Zelle kann während des Zellzyklus auf verschiedene Weisen reguliert werden. Neben der transkriptionellen Regulierung, die insbesondere durch p53 gesteuert wird, stellt der Proteinabbau eine der wichtigsten Regulationsebenen dar.
p21 kann während der gewöhnlichen Zellproliferation je nach aktueller Zellzyklusphase durch drei verschiedene E3-Ubiquitin-Ligase-Komplexe, nämlich SCF^{Skp2}, $CRL4^{Cdt2}$ und APC/C^{Cdc20}, zum Abbau durch das Proteasom markiert werden [1, 146, 143]:
Damit p21 durch SCF^{Skp2} ubiquitiniert werden kann, muss es an CyclinA-Cdk2- oder CyclinE-Cdk2-Komplexe gebunden sein. Somit tritt dieser Abbaumechanismus vorwiegend am G1/S-Übergang und während der S-Phase auf (vgl. Abb. 2.4).
Damit p21 durch $CRL4^{Cdt2}$ zum Abbau durch das Proteasom markiert werden kann, muss es an PCNA gebunden sein, welches wiederum an Chromatin gebunden ist. Da dies in der Regel nur während der DNA-Synthese der Fall ist, tritt dieser Abbaumechanismus nur während der S-Phase auf.
Durch APC/C^{Cdc20} wird p21 dann ubiquitiniert, wenn es an CyclinA-Cdk1- oder CyclinB-Cdk1-Komplexe gebunden ist, d.h. wenn sich die Zelle am Übergang zwischen der G2- und der M-Phase befindet.
Des Weiteren gibt es auch Ubiquitin-unabhängige p21-Abbaumechanismen, welche durch die Proteine Mdm2 und MdmX vermittelt werden. Mdm2 und MdmX fördern die Degradation von p21 in der G1- und frühen S-Phase, indem sie an p21 und das Proteasom binden und diese zusammenbringen [77].

Abbildung 4.2 zeigt nochmal zusammenfassend die Vorgänge bei der Aktivierung von p21 durch p53 und Abbildung 4.3 die verschiedenen p21-Abbaumechanismen.

[1] DNA-Replikation ohne anschließende Zellteilung
[2] Irreversibler Zellzyklusstopp

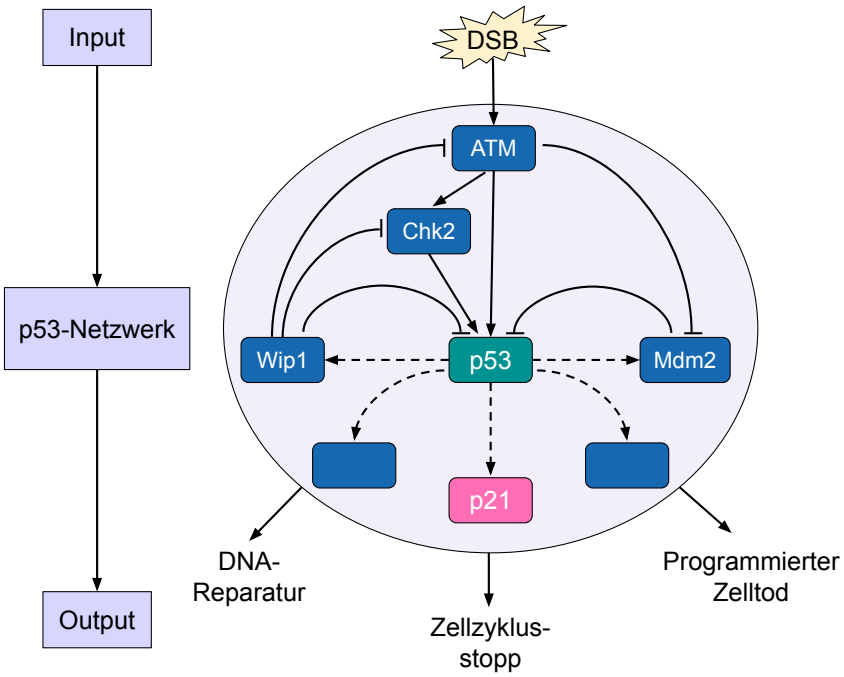


Abbildung 4.1.: Schematische Darstellung der zentralen Elemente des p53-Netzwerks bei der Reaktion von Zellen auf DNA-Doppelstrangbrüche (DSBs) nach [17].
Als Reaktion auf DSBs wird die Proteinkinase ATM aktiviert, die wiederum die Kinase Chk2 aktiviert. Beide Kinasen aktivieren dann das Protein p53, indem sie es phosophorylieren und dadurch seine Interaktion mit der E3-Ubiquitin-Ligase Mdm2 schwächen. p53 aktiviert die Transkription zahlreicher Zielgene, u.a. der Phosphatase Wip1, die das gesamte Netzwerk durch die Dephosphorylierung von ATM, Chk2, p53 und Mdm2[3] negativ reguliert. Durchgezogenen Linien stellen Protein-Protein-Wechselwirkungen dar, während gestrichelte Linien die transkriptionelle Aktivierung repräsentieren.

[3] Diese Verbindung (Wip1⊣ Mdm2) ist aus Platzgründen nicht im Netzwerk eingezeichnet.

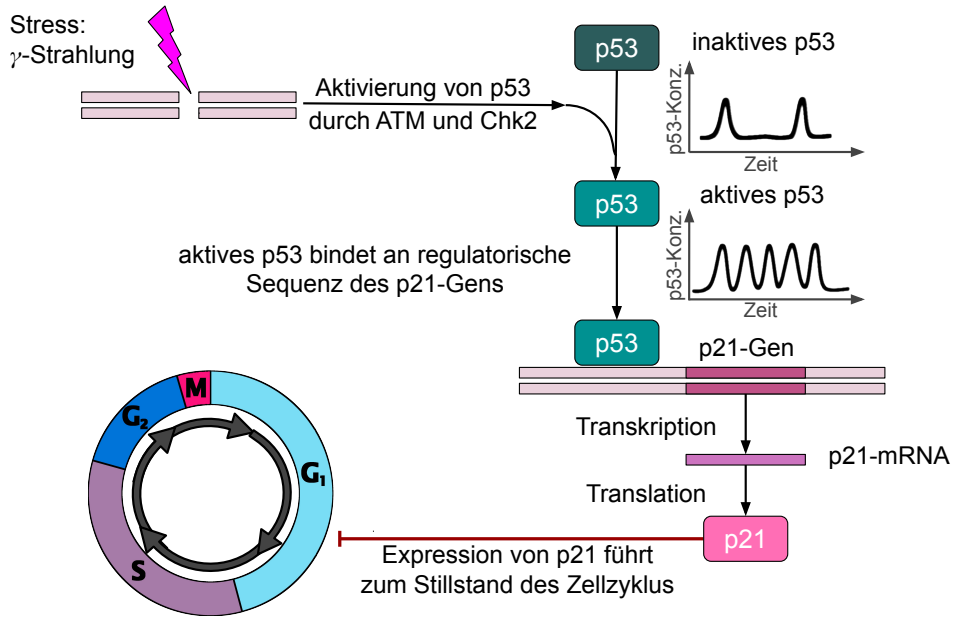


Abbildung 4.2.: Schematische Darstellung der Aktivierung von p21 durch p53.
Infolge von DNA-Doppelstrangbrüchen (DSBs) wird das Protein p53 aktiviert. Während Zellen unter normalen Bedingungen nur vereinzelte, unregelmäßige p53-Pulse aufweisen, treten nach der Induktion von DSBs mehrere gleichmäßige p53-Pulse auf. Aktiviertes p53 stimuliert dann die Transkription des Gens, das für das Cdk-Inhibitorprotein p21 kodiert. Das p21-Protein bindet vor Allem an die Cyclin-Cdk-Komplexe, die für das Eintreten der Zelle in die S-Phase notwendig sind und inaktiviert sie, sodass der Zellzyklus in der G1-Phase zum Stillstand kommt.

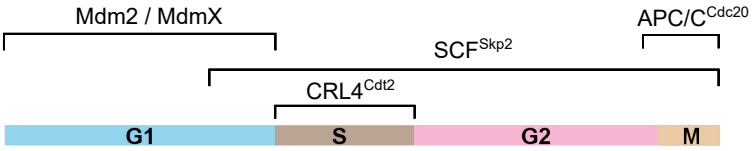


Abbildung 4.3.: Schematische Darstellung der verschiedenen Abbaumechanismen für das Protein p21. Je nach vorliegender Zellzyklusphase kann p21 während der gewöhnlichen Zellproliferation durch drei verschiedene E3-Ubiquitin-Ligase-Komplexe, nämlich SCF^{Skp2}, CRL4^{Cdt2} und APC/C^{Cdc20}, zum Abbau durch das Proteasom markiert werden. Zudem gibt es in der G1- und frühen S-Phase auch Ubiquitin-unabhängige p21-Abbaumechanismen, welche durch die Proteine Mdm2 und MdmX vermittelt werden.

4.2. Experimentelles Vorgehen und Beobachtungen

Die in diesem Kapitel dargestellten Experimente wurden von Caibin Sheng in der Arbeitsgruppe von Prof. Alexander Löwer an der TU Darmstadt durchgeführt und ausgewertet. Teilweise erhielt Caibin dabei Unterstützung von Sara Rieke, Petra Snyder, Marcel Jentsch und Dhana Friedrich (alle AG Löwer). Im Folgenden werden das experimentelle Vorgehen zur Gewinnung der gezeigten Daten, die von Caibin durchgeführte Datenanalyse sowie seine daraus erzielten Erkenntnisse vorgestellt. Hierbei beschränken wir uns auf jene Punkte, die für die Entwicklung und das Verständnis der in Kapitel 4.4 gezeigten Modelle relevant sind. Insbesondere beschränken wir uns auf die Betrachtung der Proteindynamiken nach der Bestrahlung der Zellen mit ionisierender Strahlung. Für eine detaillierte Beschreibung der durchgeführten Experimente und ihrer computergestützten Analyse sei auf die Dissertation von Caibin Sheng [143] sowie auf die zugehörige Publikation [144] verwiesen.

Um die Konzentrationen der Proteine p53 und p21 in einzelnen lebenden menschlichen Brustepithelzellen (Zelllinie MCF10A) im Laufe der Zeit zu messen, wurden zunächst mithilfe eines Verfahrens zur Genom-Editierung, der sog. CRISPR/Cas9-Technologie[4], fluoreszierende Fusionsproteine erzeugt. Das bedeutet, dass das Gen eines fluoreszierenden Proteins in der unmittelbaren Nachbarschaft des für p53 oder p21 kodierenden Gens, nämlich vor das Stopcodon, eingefügt wurde (siehe Abbildung 4.4). Auf diese Weise wird das zu untersuchende Gen gemeinsam mit demjenigen für das fluoreszierende Protein genauso exprimiert als würde es sich nur um ein einziges Gen handeln. Somit werden beide Gene durch denselben Promoter kontrolliert und im Zuge ihrer simultanen Transkription und Translation entsteht ein fluoreszierendes Fusionsprotein.

Abbildung 4.4.: Schematische Darstellung des von Caibin Sheng entwickelten endogenen Reportersystems zur gleichzeitigen Messung von p53 und p21 in einer Zelle. Sequenzen, die für fluoreszierende Proteine (YFP, RFP, CFP) kodieren, wurden mithilfe einer CRISPR/Cas9-vermittelten Genom-Editierung zwischen den kodierenden Sequenzen (CDSs) und den 3' untranslatierten Regionen (3'UTRs) eingefügt. Neben p21 und p53 wurde auch das Protein cbx5 fluoreszierend markiert, um einzelne Zellkerne im Zuge der später durchgeführten automatisierten Bildverarbeitung leichter verfolgen zu können. Quelle: [144].

[4] Genom-Editierungsmethode, für die Emmanuelle Charpentier und Jeniffer Doudna 2020 mit dem Chemie-Nobelpreis ausgezeichnet wurden. Das CRISPR/Cas9-System dient eigentlich Prokaryoten zur Abwehr von Viren, mit denen sie früher einmal infiziert waren. Genauer gesagt, verleiht es ihnen die Fähigkeit, genetische Sequenzen, die zu einem bestimmten Virus gehören, präzise zu erkennen und diese mithilfe spezieller Enzyme, sog. CRISPR-assoziierter Proteine (Cas), zu zerschneiden und somit zu zerstören. Emmanuelle Charpentier und Jeniffer Doudna hatten nach der Entdeckung dieses adaptiven prokaryotischen „Immunsystems" die Idee, dass sich dieser Mechanismus dafür eignen könnte, jede DNA an einer spezifischen Position zu schneiden [78, 135]. Das CRISPR/Cas9-System konnte später von Feng Zhang und seinen Kollegen von Prokaryoten auf Eukaryoten übertragen werden [132] und wird nun zum gezielten Schneiden und Modifizieren von (eukaryotischer) DNA genutzt [76]. Genauere Informationen zur CRISPR/Cas9-Technologie können z.B. [76] oder neueren Biologiebüchern wie beispielsweise [135] entnommen werden.

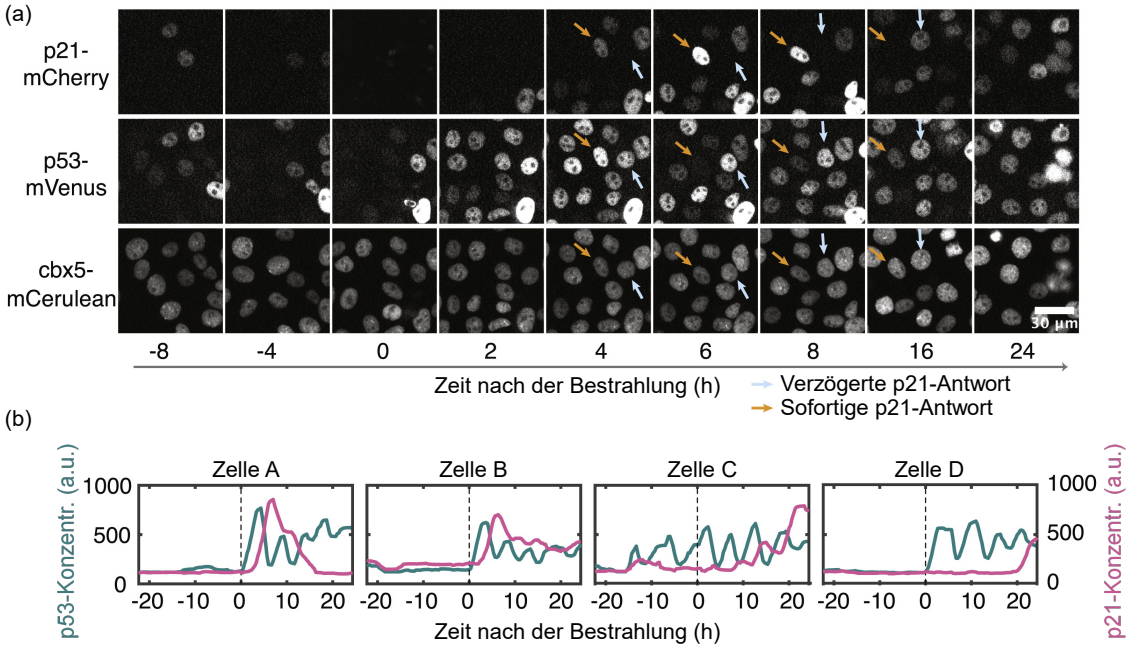


Abbildung 4.5.: Messung der p21- und p53-Konzentration in einzelnen Zellen. Quelle: [144].
(a) Zeitraffermikroskopische Aufnahmen von lebenden MCF10A-Zellen, die die Fusionsproteine p21-mCherry und p53-mVenus exprimieren. In jeder Zeile ist das Signal eines anderen Fluoreszenzreporters gezeigt (oben: p21-mCherry (RFP), Mitte: p53-mVenus (YFP), unten: cbx5-mCerulan (CFP)). Die gezeigten Zellen wurden zunächst unter basalen Bedingungen beobachtet, dann mit ionisierender Strahlung (5 Gy) bestrahlt und weitere 24 Stunden lang aufgenommen. Zwei Beispielzellen mit unterschiedlichen p21-Reaktionen sind durch farbige Pfeile markiert.
(b) Vier beispielhafte Zeitreihen, die die unterschiedliche Dynamik von p53 (grün) und p21 (magenta) in einzelnen Zellen zeigen. Die gestrichelten Linien geben den Zeitpunkt der Bestrahlung an.

Die Fluoreszenzintensität dieser Fusionsproteine, welche auch als *Fluoreszenzreporter* bezeichnet werden, kann dann, wie in Abbildung 4.5 (a) dargestellt, mittels Zeitraffermikroskopie auf der Einzelzellebene aufgenommen und mit Hilfe automatisierter Bildverarbeitungsmethoden ausgewertet werden.

In den von Caibin Sheng durchgeführten Experimenten wurden zunächst die Proteinkonzentrationen ungefähr 20 Stunden lang während der gewöhnlichen Zellproliferation, d.h. unter basalen Bedingungen, beobachtet. Dann folgte zur Induktion von DNA-Doppelstrangbrüchen die Bestrahlung der Zellen mit Gammastrahlung (Dosis: 5 Gy), woraufhin die Proteinkonzentrationen weitere 24 Stunden lang gemessen wurden.

Hierbei zeigte sich, wie in Abbildung 4.5 (b) zu sehen ist, dass die p53-Dynamik nach der Schadensinduktion in verschiedenen Zellen relativ homogen ist, während der Zeitpunkt und die Geschwindigkeit der p21-Induktion heterogen sind. In manchen Zellen steigt die p21-Konzentration bereits unmittelbar nach der p53-Aktivierung an (siehe Abb. 4.5 (b),

Zellen A und B), während andere Zellen erst mit mehrstündiger Verzögerung mit einer Erhöhung der p21-Konzentration auf die Bestrahlung reagieren (siehe Abb. 4.5 (b), Zellen C und D). Die gleiche Beobachtung, dass die p53-Antwort verschiedener Zellen homogen, aber ihre p21-Reaktion heterogen ist, wurde auch bei höheren Bestrahlungsdosen (10 Gy und 20 Gy) gemacht [143, 144]. Daher kann mit hoher Wahrscheinlichkeit ausgeschlossen werden, dass die beobachtete Heterogenität auf unterschiedliche Schädigungsgrade der einzelnen Zellen zurückzuführen ist.

Um zu ermitteln, welcher Mechanismus stattdessen für die Heterogenität der p21-Dynamiken verantwortlich ist, ist es naheliegend zu analysieren, welche Gemeinsamkeiten Zellen mit ähnlichen p21-Verläufen aufweisen. Um solche Zellen mit ähnlichen p21-Verläufen ausfindig zu machen, wurden die p21-Zeitreihen von Caibin Sheng mittels eines automatischen Clusteringalgorithmus, dem 2015 von Paparrizos et al. entwickelten k-Shape-Algorithmus [124], durch eine zweimal hintereinander ausgeführte binäre Klassifizierung, in vier Klassen eingeteilt. Dieses Vorgehen ist in Abbildung 4.6 dargestellt. Als Ähnlichkeitsmaß zwischen zwei Zeitreihen verwendet der k-Shape-Algorithmus dabei die sog. formbezogene Distanz (Englisch: shape-based distance, SBD), welche auf der normierten Kreuzkorrelation der beiden Zeitreihen basiert. Auf diese Weise ermöglicht es der k-Shape-Algorithmus, Zeitreihen mit denselben Charakteristika trotz unterschiedlicher Amplituden und vorliegenden Phasendifferenzen einem gemeinsamen Cluster zuzuordnen. Weitere Details zur durchgeführten Clusteranalyse und dem verwendeten Algorithmus können Caibin Shengs Doktorarbeit [143] und dem Paper von Paparrizos et al. [124] entnommen werden.

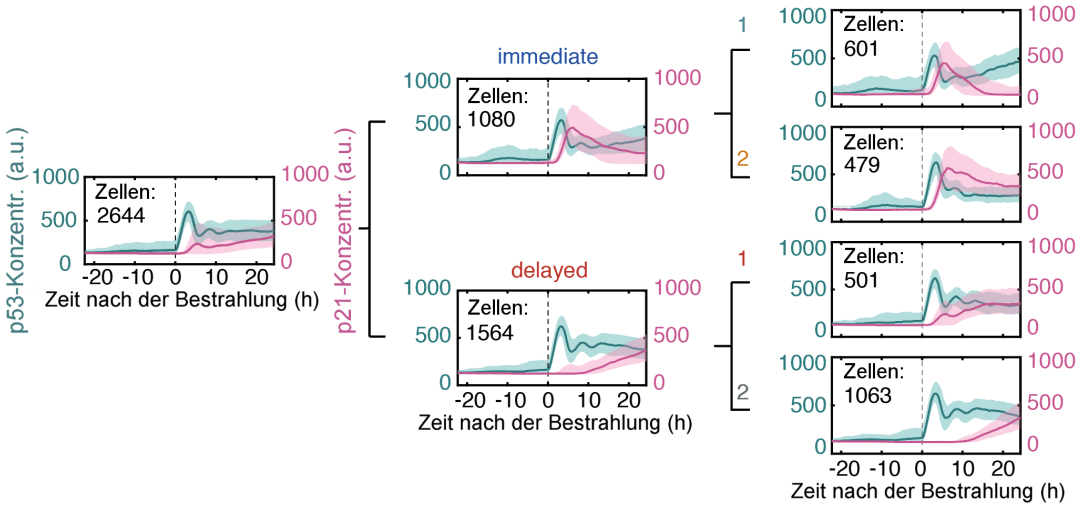


Abbildung 4.6.: Die untersuchten Zellen wurden durch ein zweimal hintereinander ausgeführtes binäres Clustering in vier Untergruppen mit unterschiedlichen p21-Verläufen eingeteilt. Die Anzahl der Zellen in jeder Untergruppe ist angegeben. Die grünen und magentafarbenen Linien geben den Median der p53- bzw. p21-Proteinkonzentrationen in jeder Gruppe im Laufe der Zeit an, während die schattierten Bereiche das 25. bis 75. Perzentil kennzeichnen.

Anhand der Clusteranalyse zeigte sich, dass etwa 40% der Zellen mit einer sofortigen p21-Akkumulation (Gruppe „immediate“) nach der Bestrahlung reagierten, wobei zwei Gruppen unterschieden werden konnten. Bei Zellen, die zur Gruppe „immediate 1“ gehören, nahm die p21-Konzentration vergleichsweise schnell wieder ab und befand sich in der Regel bis zum Ende des Experiments wieder auf basalem Niveau. Im Gegensatz dazu lagen bei Zellen der Gruppe „immediate 2“ auch 24 Stunden nach der Bestrahlung noch erhöhte p21-Konzentrationen vor. Die anderen 60% der Zellen wiesen einen verzögerten Anstieg der p21-Konzentration auf (Gruppe „delayed“), wobei diese Verzögerung bei Zellen der Gruppe „delayed 1“ deutlich geringer ausfiel aus bei Zellen der Gruppe „delayed 2“.

Als nächstes wurde für jede Zelle die Zellzyklusphase zum Zeitpunkt der DNA-Schädigung und am Ende des Beobachtungszeitraums bestimmt. Das Vorgehen hierzu ist in Abbildung 4.7 (a) gezeigt.
Die Zellzyklusphase zum Zeitpunkt der Schadensinduktion wurde anhand der letzten Zellteilung vor der Bestrahlung abgeschätzt. Zellen, die sich wenige Stunden vor der Bestrahlung geteilt hatten, waren mit hoher Wahrscheinlichkeit in der G1-Phase, während Zellen, deren letzte Zellteilung vor einer vergleichsweise langen Zeit stattgefunden hatte, vermutlich während der Bestrahlung in der G2-Phase waren. Zellen, deren letzte Teilung irgendwann dazwischen stattgefunden hatte, waren dementsprechend am wahrscheinlichsten in der S-Phase. Dieser Ansatz zur Bestimmung der Zellzyklusphase anhand des Zellteilungszeitpunkts konnte von Caibin Sheng in einem unabhängigen Experiment, bei welchem die Korrelation zwischen dem Zeitpunkt der Zellteilung und der späteren Zellzyklusphase unter basalen Bedingungen explizit untersucht wurde, validiert werden (siehe Abb. 4.7 (b)).
Um die Phase des Zellzyklus am Ende des Beobachtungszeitraums bestimmen zu können, wurden die Zellen in den letzten 30 Minuten des Experiments mit 5-Ethynyl-2’-deoxyuridine (EdU) markiert. EdU ist ein Thymidinanalogon und wird während der S-Phase, d.h. während der DNA-Synthese, in die DNA eingebaut. Somit lassen sich Zellen, welche am Ende des Beobachtungszeitraums in der S-Phase waren, zuverlässig erkennen. Zudem wurden die Zellkerngröße sowie der DNA-Gehalt mit einem DNA-bindenden Farbstoff (Hoechst 33342) gemessen. Auf Grundlage dieser drei Werte, der EdU-Konzentration, der Zellkerngröße und des DNA-Gehalts, wurde dann ein semi-überwachtes Klassifizierungsverfahren zur Bestimmung der Zellzyklusphase implementiert (siehe. Abb. 4.7 (c)).

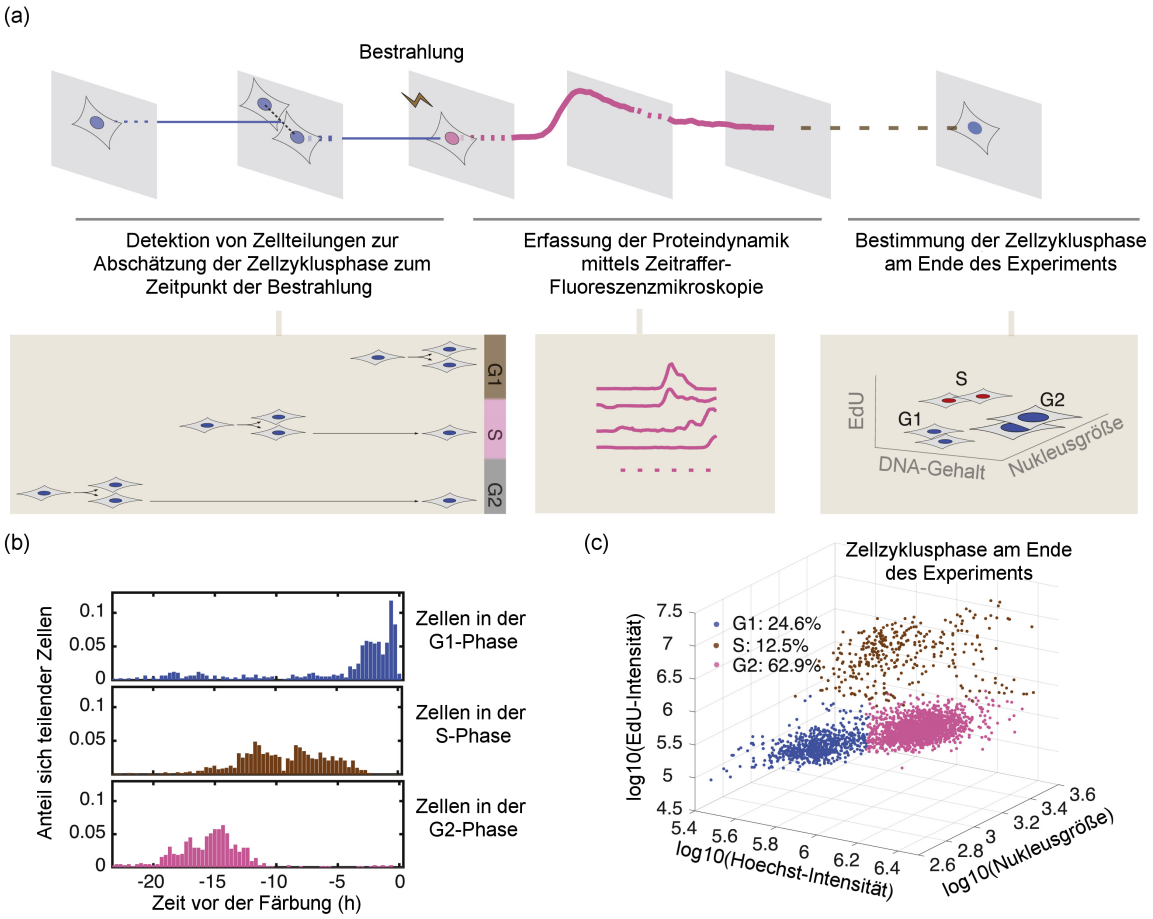


Abbildung 4.7.: (a) Schematische Darstellung des Experiments zur Bestimmung des Zusammenhangs zwischen der p21-Dynamik und des Zellzyklusverlaufs. Der Zellzykluszustand zum Zeitpunkt der Bestrahlung wurde anhand der vor der Schadensinduktion beobachteten Zellteilungen abgeschätzt. Zur Bestimmung der Zellzyklusphase am Ende des Experiments, wurde der EdU-Einbau sowie die Zellkerngröße und der DNA-Gehalt gemessen.
(b) Bestätigung der Korrelation zwischen dem Zeitpunkt der Zellteilung und der Zellzyklusphase unter basalen Bedingungen. MCF10A-Zellen wurden 24 Stunden lang beobachtet und die Zellteilungsereignisse aufgezeichnet. Durch EdU-Markierung (30 Minuten vor dem Ende des Beobachtungszeitraums wurde EdU in das Medium gegeben), welche zur Markierung von Zellen in der S-Phase dient, und Immunfluoreszenzfärbung des Zellzyklusmarkers CyclinB1 (unmittelbar nach der Bildgebung) wurde die Zellzyklusphase am Ende des Beobachtungszeitraums bestimmt.
(c) Bestimmung der Zellzyklusphase am Ende des Experiments bzw. 24 Stunden nach der Bestrahlung durch EdU-Markierung und Messung der Zellkerngröße sowie des DNA-Gehalts.
Quelle: [144], wobei der Text in Abbildung (a) leicht modifiziert wurde.

Anhand dieser Analyse konnte Caibin zeigen, dass die p21-Antwort einer Zelle maßgeblich von der Zellzyklusphase dieser Zelle zum Zeitpunkt der Bestrahlung und dem anschließenden Zellzyklusverlauf während der Schadensantwort abhängt.

Im wesentlichen beobachtete er drei verschiedene Arten der p21-Dynamik, die in Abbildung 4.8 zu sehen sind:

1. Zellen, die in der G1- oder G2-Phase geschädigt wurden und deren Zellzyklus nach der Bestrahlung in dieser Phase zum Erliegen kam, zeigten nach der Bestrahlung eine sofortige p21-Akkumulation, gefolgt von einem langsamen Abfall der p21-Konzentration. Auch 24 Stunden nach der Schädigung wiesen diese Zellen im Vergleich zum Zeitraum vor der Bestrahlung noch eine erhöhte p21-Konzentration auf.

2. Zellen, die in der G1-Phase geschädigt wurden, aber nach der Bestrahlung in die S-Phase übergegangen sind, zeigten einen sofortigen, pulsförmigen p21-Anstieg. Im Gegensatz zur vorherigen Gruppe, war dieser Anstieg allerdings nur vorübergehend und die Zellen wiesen am Ende des Beobachtungszeitraums wieder basale p21-Konzentrationen auf. Falls die Zellen die gesamte S-Phase durchliefen und sich am Ende des Experiments in der G2-Phase befanden, wurden wieder steigende p21-Konzentrationen beobachtet.

3. Zellen, die in der S-Phase bestrahlt wurden und deren Zellzyklus anschließend in der G2-Phase stoppte, zeigten eine verzögerte p21-Antwort. Das Einsetzten der p21-Akkumulation korrelierte dabei mit dem Zeitpunkt der letzten Zellteilung vor der Schädigung.

4.3. Fragestellung

Es stellt sich nun die Frage, welcher zellzyklusabhängige molekulare Mechanismus für die von Caibin Sheng beobachtete Heterogenität der p21-Antwort infolge von DNA-Schäden verantwortlich ist. Um diese Frage zu beantworten, wollen wir im nächsten Abschnitt zunächst die Regulation von p21 durch p53 mathematisch modellieren und hierbei zwei verschiedene Hypothesen für die Entstehung der heterogenen p21-Reaktionen miteinander vergleichen.

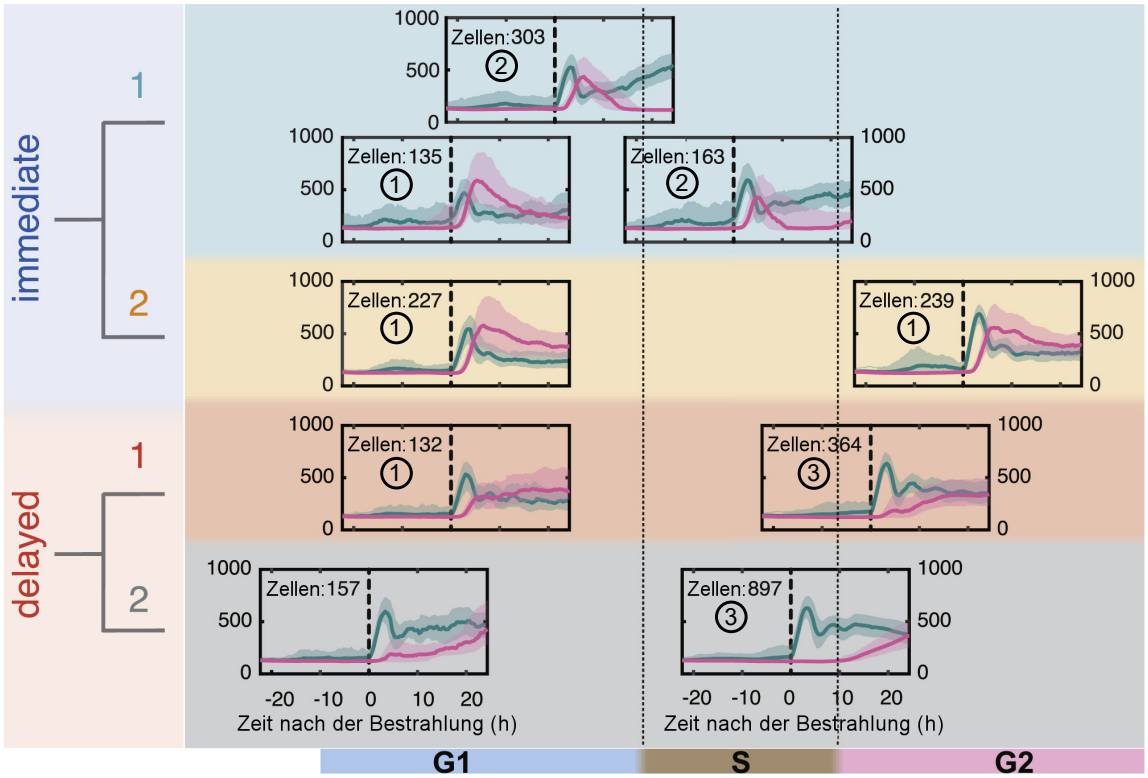


Abbildung 4.8.: Weitere Unterteilung der vier identifizierten Untergruppen von p21-Reaktionen (vgl. Abb. 4.6) gemäß dem ursprünglichen Zellzykluszustand zum Zeitpunkt der Bestrahlung und dem weiteren Zellzyklusverlauf. Quelle:[144]
Die Anordnung jedes einzelnen Plots zeigt die Zellzykluseigenschaften für die entsprechende Untergruppe, wobei die linke Seite des Plots den anfänglichen Zellzykluszustand zum Zeitpunkt der Schadensinduktion und die rechte Seite die Zellzyklusphase am Ende des Beobachtungszeitraums angibt. Ist die Darstellung eines Plots auf eine einzige Zellzyklusphase beschränkt, ist der Zellzyklus der entsprechenden Zellen nach der Schädigung zum Stillstand gekommen. Ein Plot, der mehrere Zellzyklusphasen umfasst, zeigt den Verlauf des Zellzyklus während der Schadensreaktion an. Hierbei ist zu beachten, dass die Zellzyklusphasen nicht maßstabsgetreu dargestellt sind. Die grünen und magentafarbenen Linien zeigen den Median der p53- bzw. p21- Proteinkonzentrationen in jeder Gruppe im Laufe der Zeit. Die schattierten Bereiche geben das 25. bis 75. Perzentil an. Die Anzahl der Zellen in jeder Gruppe ist angegeben. Die rund umrandeten Zahlen in den Plots geben zudem die Art der p21-Antwort an, wie sie im Text links beschrieben ist. Bei 157 Zellen (unten links) konnte die Zellzyklusphase zum Zeitpunkt der Bestrahlung nicht eindeutig bestimmt werden, da die Zellteilungszeiten vor der Schädigung gleichmäßig verteilt waren. Diese Zellen werden nicht weiter betrachtet.

4.4. Modellierung

Um den molekularen Mechanismus zu ermitteln, der für die heterogenen p21-Reaktionen verantwortlich ist, wurden zwei einfache mathematische Modelle für die zellzyklusabhängige Regulation von p21 aufgestellt und miteinander verglichen. Beide Modelle basieren dabei auf einer einzigen retardierten Differentialgleichung (DGL), die die Änderungsrate der p21-Konzentration nach einer DNA-Schädigung beschreibt und jeweils einen unterschiedlichen Mechanismus für die Entstehung der heterogenen p21-Reaktionen implementiert. Da laut den experimentellen Beobachtungen die p21-Konzentration in bestrahlten Zellen während der S-Phase immer niedrig ist (vgl. Abbildung 4.8), nehmen wir an, dass in dieser Zellzyklusphase entweder die Produktion von p21 signifikant vermindert (siehe hierzu [18, 66, 110]) oder der Abbau von p21 im Vergleich zu anderen Phasen des Zellzyklus stark erhöht sein muss (siehe [147]).

Im ersten Modell, welches in Abbildung 4.9 links dargestellt ist, gehen wir vereinfachend davon aus, dass die Produktion von p21 während der S-Phase vollständig zum Erliegen kommt. Dies wird durch die folgende DGL für die Änderungsrate der p21-Konzentration implementiert:

$$\frac{\mathrm{d}p21(t)}{\mathrm{d}t} = \begin{cases} \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta \cdot p21(t), & \text{Zelle in der G1- oder G2-Phase} \\ -\delta \cdot p21(t), & \text{Zelle in der S-Phase} \end{cases} \tag{4.1}$$

Im zweiten Modell, welches in Abbildung 4.9 rechts zu sehen ist, nehmen wir hingegen an, dass die Abbaurate von p21 während der S-Phase viel höher ist als in anderen Zellzyklusphasen, was zu der folgenden retardierten DGL führt:

$$\frac{\mathrm{d}p21(t)}{\mathrm{d}t} = \begin{cases} \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta \cdot p21(t), & \text{Zelle in der G1- oder G2-Phase} \\ \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta \cdot D_S \cdot p21(t), & \text{Zelle in der S-Phase} \end{cases} \tag{4.2}$$

Der erste Term auf der rechten Seite der beiden Gleichungen beschreibt dabei die p53-abhängige Produktion von p21. Sie wird durch eine Hill-Funktion mit einer maximalen Produktionsrate m und einer Aktivierungsschwelle θ modelliert, die die Konzentration von p53 angibt, die notwendig ist, um die p21-Expression signifikant zu aktivieren. Der Hill-Koeffizient n beschreibt die Kooperativität des Transkriptionsprozesses und τ repräsentiert die zeitliche Verzögerung der p21-Expression durch die Dauer des Transkriptions- und Translationsprozesses. Die Produktionsrate wird im ersten Modell während der S-Phase auf Null gesetzt.
Der zweite Term auf der rechten Seite der Gleichungen (4.1) und (4.2) beschreibt den proteasomalen Abbau von p21 mit der Abbaurate δ. Im zweiten Modell wird diese Abbaurate während der S-Phase um einen konstanten Faktor $D_S > 1$ erhöht.

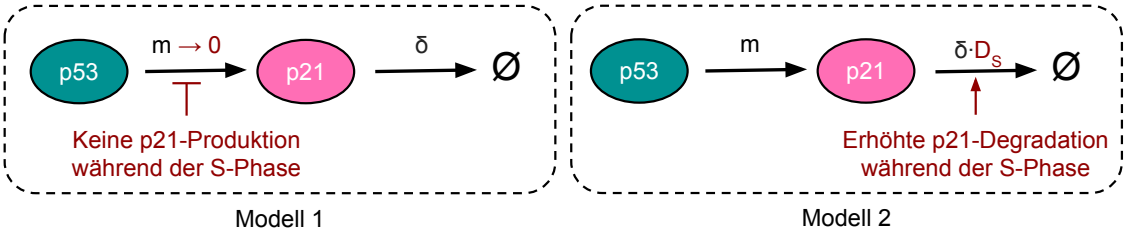


Abbildung 4.9.: Schematische Darstellung der entwickelten Modelle.
Im ersten Modell (siehe Gl. (4.1)) wird angenommen, dass während der S-Phase kein p21-Protein produziert wird. Dies wird durch eine maximale p21-Produktionsrate von $m = 0$ implementiert. Im zweiten Modell (siehe Gl. (4.2)) wird angenommen, dass die Degradation von p21 während der S-Phase signifikant erhöht ist, was durch die Multiplikation der Abbaurate δ mit dem Faktor D_S implementiert wird.

Ein wesentliches Merkmal beider Modelle ist, dass wir für jede untersuchte Zelle die experimentell gemessenen p53-Konzentrationen als Modellinput verwenden. Hierzu wird zunächst eine Interpolation der gemessenen p53-Konzentrationen durchgeführt und der Hintergrund, der jeweils durch den kleinsten gemessenen p21- bzw. p53-Konzentrationswert in der betrachteten Zelle geschätzt wird, von den p21- und p53-Daten jeder Zelle subtrahiert. Dies hat vor allem den Zweck, dass keine zusätzliche konstante, p53-unabhängige p21-Produktionsrate als weiterer Parameter in die Modelle eingefügt werden muss.

Anschließend fitten wir unsere Modelle an die p21-Daten von Tausenden von einzelnen Zellen und berechnen den Median der resultierenden Fits für Zellen mit einer ähnlichen p21-Dynamik, d.h. für Zellen, die laut Caibins Analyse (siehe Abb. 4.8) dieselben Zellzyklus-Charakteristika aufweisen. Hierbei werden die Parameter m, θ, δ und der Zeitpunkt des Zellzyklusphasenwechsels t_S (sofern anhand der experimentellen Daten von einem solchen Wechsel auszugehen ist) als Fit-Parameter verwendet, während die Parameter n, τ und D_S für alle Zellen einen fest definierten Wert erhalten. Für den Hill-Koeffizienten n wählen wir dabei einen Wert von $n = 4$, da p53 als Tetramer an die Promoterregion seiner Zielgene bindet [86, 103, 111]. Die Verzögerungszeit τ wird für jede Zelle auf $\tau = 1.4$ h gesetzt, da wir davon ausgehen können, dass die Verzögerung der p21-Expression durch die Dauer der Transkription und Translation für jede Zelle ähnlich sein sollte. Zudem nehmen wir an, dass die p21-Abbaurate während der S-Phase $D_S = 50$ mal höher ist als in der G1- oder G2-Phase.
Weitere technische Details zur Durchführung der Fits mit dem Programm Mathematica und Erläuterungen zur Eingrenzung der erlaubten Parameterwerte sind im Anhang B.1 beschrieben.

4.4.1. Diskussion des gewählten Modellierungsansatzes

Im Allgemeinen ist es bei der Entwicklung eines Modells sinnvoll, die Anzahl der frei wählbaren Parameter im Verhältnis zur Anzahl der verfügbaren experimentellen Daten möglichst gering zu halten. Modelle mit vielen freien Parametern weisen aufgrund ihrer Komplexität und der hohen Anzahl an Freiheitsgraden oft eine Vielzahl von möglichen Fit-Lösungen auf, die die experimentellen Daten reproduzieren können. Dies limitiert die Aussagekraft solcher Modell bezüglich der Hauptmechanismen, die für das experimentell beobachtete Verhalten verantwortlich sind. Aus diesem Grund wurde in dieser Arbeit versucht, ein minimales Modell mit einer möglichst geringen Anzahl an freien Parametern zu finden, das die experimentellen Daten hinreichend gut beschreiben kann. [96]

Ein zentrales Problem bei der Modellierung der Genexpression ist, dass die betrachteten biologischen Prozesse eine gewisse Zeit benötigen, um abgeschlossen zu werden [87]. So ist beispielsweise (reife) mRNA nicht sofort nach dem Beginn der Transkription verfügbar und muss zunächst zu den Ribosomen im Cytoplasma transportiert werden, bevor sie im Zuge der Translation in ein Protein übersetzt werden kann. Diesem Problem kann man im Wesentlichen auf zwei unterschiedliche Arten begegnen: Entweder man berücksichtigt bei der Modellierung explizit die zahlreichen Zwischenreaktionsschritte, oder man führt eine diskrete Zeitverzögerung ein [87].
Um die Anzahl der benötigten Spezies in den hier entwickelten Modellen zu verringern, bzw. um die Notwendigkeit einer weiteren, mit freien Parametern verbundenen Differentialgleichung (DGL) für die mRNA von p21 zu vermeiden, habe ich mich für die zweite Option entschieden und eine explizite zeitliche Verzögerung in die DGL für p21 eingeführt. Dadurch handelt es sich allerdings nicht mehr um eine gewöhnliche DGL, sondern um eine sog. *retardierte DGL*. Ein Nachteil von solchen retardierten DGLs ist, dass sie aufgrund eines unendlich dimensionalen Phasenraums deutlich schwerer zu handhaben und in der Regel nur numerisch lösbar sind [134]. Da uns jedoch keine experimentellen Daten zum zeitlichen Verlauf der p21-mRNA-Konzentration vorlagen, wurde die Einführung einer konstanten zeitlichen Verzögerung gegenüber der Einführung mehrerer freier Modellparameter präferiert.

Eine Besonderheit der hier verwendeten Modelle ist zudem, dass die gemessenen p53-Konzentrationen direkt in die Modelle eingehen und die p53-Dynamik somit nicht durch eine explizite DGL beschrieben wird. Dadurch lassen sich theoretische Analysemethoden wie beispielsweise die Bifurkationsanalyse in diesem Fall leider nicht anwenden. Gleichzeitig können auf diese Weise zahlreiche Unsicherheiten und viele unbekannte Parameter vermieden werden, die bei einer expliziten Modellierung von p53, wie sie z.B. von Chong et al. [33] durchgeführt wurde, auftreten. Dieses Vorgehen ermöglicht es uns somit – trotz des damit verbundenen Nachteils bezüglich der Analyse des Modellverhaltens – den direkten Einfluss von p53 auf p21 zu betrachten.

4.5. Ergebnisse

Die Ergebnisse des Vergleichs beider Modell mit den experimentellen Daten sind in den Abbildungen 4.10 und 4.11 gezeigt. Sowohl auf der Einzellzellebene als auch beim Vergleich der Mediane der Fits mit den experimentell ermittelten Medianen der p21-Konzentrationen stellen wir fest, dass beide Modelle die wesentlichen Merkmale des zellzyklusabhängigen dynamischen Verhaltens von p21 nach der Induktion von DNA-Schäden reproduzieren können. So reproduzieren beide Modelle die unmittelbare, pulsartige p21-Antwort von Zellen, die in der G1-Phase bestrahlt wurden und anschließend in die S-Phase übergegangen sind („G1 $\rightarrow$ S“), – und zwar unabhängig vom genauen Wert von D_S in Modell 2 (siehe Appendix B.2) – sowie die Verzögerung der p21-Antwort in Zellen, die den Schaden in der S-Phase erfahren haben („S $\rightarrow$ G2“).
Quantitativ lassen sich beide Modelle vergleichen, indem man die Summe der quadratischen Abweichungen zwischen den experimentellen Daten ($p21_{\text{Exp}}$) und den durch das jeweilige Modell vorhergesagten p21-Konzentrationen ($p21_{\text{Modell}}$), die sog. Residuenquadratsumme $\chi^2 = \sum_i \left|p21_{\text{Exp}}(i) - p21_{\text{Modell}}(i)\right|^2$, betrachtet. Dieser direkte Vergleich anhand der Residuenquadratsumme ist möglich, da beide Modelle (insbesondere durch die Wahl eines festen Wertes für D_S in Modell 2) gleich viele freie Parameter besitzen. Wie man in Abbildung 4.10 (f) anhand der Verteilung der Residuenquadratsummen für die durchgeführten Fits sieht, können beide Modelle die Daten ähnlich gut beschreiben und keines der Modelle ist eindeutig überlegen. Allerdings fällt beim Vergleich der beiden Modelle am Übergang von der G1- zur S-Phase auf (siehe Abbildung 4.10 (e)), dass Modell 1, welches von einem Stopp der p21-Produktion während der S-Phase ausgeht, im Allgemeinen zu einer langsameren Abnahme der p21-Konzentration führt als es in den experimentellen Daten beobachtet wird und Modell 2, das einen verstärkten p21-Abbau in der S-Phase annimmt, tendenziell zu einer schnelleren Abnahme. Während bei Modell 2 die Ergebnisse jedoch durch die Implementierung einer biologisch plausiblen, allmählich zunehmenden p21-Abbaurate (siehe hierzu z.B. [37]) weiter verbessert werden könnten, gibt es bei Modell 1 konzeptionell keine Möglichkeit, um einen steileren Abfall der p21-Konzentration zu erreichen (da bereits ein kompletter Stopp der p21-Produktion angenommen wurde).
Dies deutet darauf hin, dass die heterogene p21-Antwort eher durch einen erhöhten p21-Abbau, der einen schnellen p21-Abfall am G1-S-Übergang ermöglicht, als durch eine verminderte Produktion des Proteins während der S-Phase verursacht wird.

Für Zellen, deren Zellzyklus entweder in der G1- oder G2-Phase zum Erliegen kam, beobachten wir beim Vergleich der Mediane der Fits mit den Medianen der experimentell ermittelten p21-Konzentrationen, dass unsere einfachen Modelle – die in diesem Fall beide identisch sind – den Zeitpunkt und die Amplitude des ersten p21-Maximums nicht perfekt reproduzieren können, während spätere p21-Konzentrationen gut wiedergegeben werden (siehe Abbildung 4.11 (a)-(b)). Um zu untersuchen, ob dieses Ergebnis verbessert werden kann, wurden weitere Simulationen durchgeführt, bei denen die Dynamik von p21 nur für etwa zehn Stunden nach der Bestrahlung gefittet wurde. Wie in Abbildung 4.11 (c)-(d) zu sehen ist, kann auf diese Weise zwar das Maximum der p21-Konzentration gut reproduziert werden, allerdings hat dies zu späteren Zeitpunkten größere Abweichungen zwischen den Fits und den gemessenen Daten zur Folge.

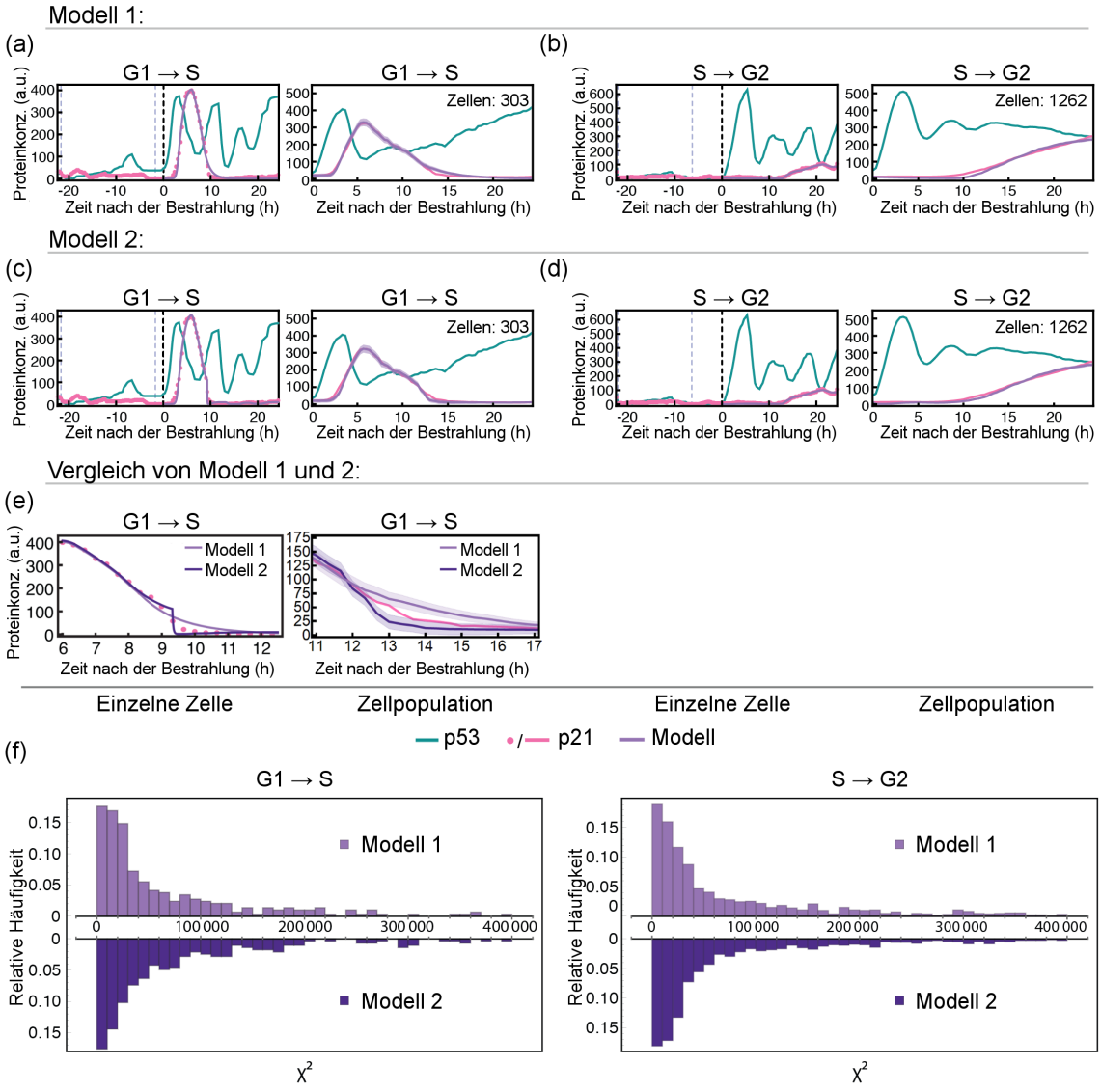


Abbildung 4.10.: Vergleich der beiden Modelle mit den experimentellen Daten.
In Abb. (a-b) sind die Ergebnisse für Modell 1 und in Abb. (c-d) die Ergebnisse für Modell 2 dargestellt. In Abb. (e) ist ein Vergleich beider Modelle am Übergang von der G1- zur S-Phase zu sehen und Abb. (f) zeigt die Verteilung der Residuenquadratsummen $\chi^2 = \sum_i |p21_{\text{Exp}}(i) - p21_{\text{Modell}}(i)|^2$ für beide Modelle. In Abb. (a-e) zeigen die linken Abbildungen jeweils die Proteindynamik in einzelnen Zellen und die rechten Abbildungen zeigen die Medianwerte für mehrere Zellen mit denselben Zellzyklus-Charakteristika. Die Anzahl der analysierten Zellen in jeder Kategorie ist an der entsprechenden Stelle angegeben. Schwarz gestrichelte Linien kennzeichnen den Zeitpunkt der Bestrahlung (5 Gy) und blau gestrichelte Linien markieren Zellteilungen. Die lila schattierten Bereiche stellen den Standardfehler des Medians dar, der gemäß $\sqrt{\pi/2} \cdot \sigma/\sqrt{N}$ [72] berechnet wurde, wobei σ die Standardabweichung ist und N die Anzahl der betrachteten Zellen. Folgende Parameter wurden zur Modellierung der p21-Konzentration in den hier gezeigten, einzelnen Zellen (jeweils links) verwendet:
(a) $m = 418.65, \theta = 196.18, \delta = 0.85, t_S = 8.33$; (b) $m = 27.88, \theta = 151.02, \delta = 0.2, t_S = 12.74$; (c) $m = 399.73, \theta = 204.90, \delta = 0.77, D_S = 50, t_S = 9.32$; (d) $m = 28.08, \theta = 152.11, \delta = 0.2, D_S = 50, t_S = 12.84$.

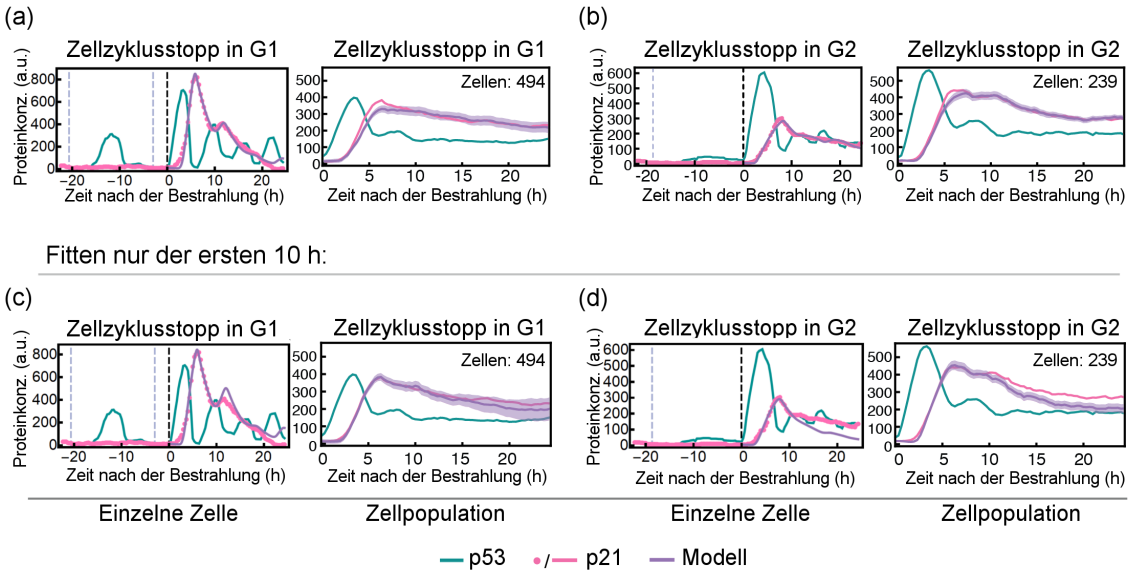


Abbildung 4.11.: Vergleich von simulierten (lila) und gemessenen (magenta) p21-Proteinkonzentrationen für Zellen, deren Zellzyklus in der G1- (a,c) oder G2-Phase (b,d) zum Erliegen kam. Die linken Abbildungen zeigen jeweils die Proteindynamik in einzelnen Zellen, die rechten Abbildungen zeigen die Medianwerte für mehrere Zellen (Anzahl ist jeweils angegeben) mit denselben Zellzyklus-Charakteristika. In den oberen Abbildungen (a) und (b) wurde das Modell (in diesem Fall sind beide betrachteten Modelle identisch) wie zuvor an alle p21-Daten nach der Bestrahlung der Zellen gefittet. Dies führte allerdings dazu, dass der Zeitpunkt und die Amplitude des ersten p21-Maximums auf der Zellpopulationsebene nicht richtig wiedergegeben wurde. Aus diesem Grund wurden in den unteren Abbildungen (c) und (d) nur die p21-Konzentrationen der ersten 10 Stunden nach der Bestrahlung zum Fitten des Modells genutzt. Dies führte zwar im Bereich des p21-Maximums zu einer deutlich besseren Übereinstimmung des Modells mit den experimentellen Daten, allerdings ergaben sich hierdurch zu späteren Zeitpunkten deutlich größere Abweichungen zwischen den Fits und den gemessenen Daten.
Folgende Parameter wurden zur Modellierung der p21-Konzentration in den hier gezeigten, einzelnen Zellen (jeweils links) verwendet:
(a) $m = 639.30, \theta = 519.18, \delta = 0.25$; (b) $m = 82.83, \theta = 232.46, \delta = 0.16$;
(c) $m = 505.68, \theta = 414.74, \delta = 0.27$; (d) $m = 111.49, \theta = 343.99, \delta = 0.18$;
Für weitere Details zur Abbildung siehe auch Abb. 4.10.

Für Zellen, deren Zellzyklus in der G2-Phase stoppte, sind in diesem Fall die zu späteren Zeitpunkten vorhergesagten p21-Konzentrationen deutlich niedriger als die experimentellen Daten (siehe Abb. 4.11(d)).
Für Zellen, deren Zellzyklus in der G1-Phase zum Stillstand kam, zeigt eine detailliertere Betrachtung der unterschiedlichen p21-Dynamik-Untergruppen, die in Abbildung 4.12 zu sehen ist, dass die Abweichung zwischen den simulierten und den experimentellen p21-Verläufen für Zellen, die zur Untergruppe „immediate 1“ gehören, besonders groß ist. Für diese Zellen scheint es nicht möglich zu sein, das hohe p21-Maximum kurz nach der Bestrahlung und die relativ niedrigen p21-Konzentrationen am Ende der Messung gleichzeitig mit demselben Satz von Modellparametern zu reproduzieren.
Insgesamt könnten diese Ergebnisse darauf hinweisen, dass die untersuchten Zellen auf den ersten p53-Puls anders reagieren als auf spätere p53-Pulse (vgl. [32]).

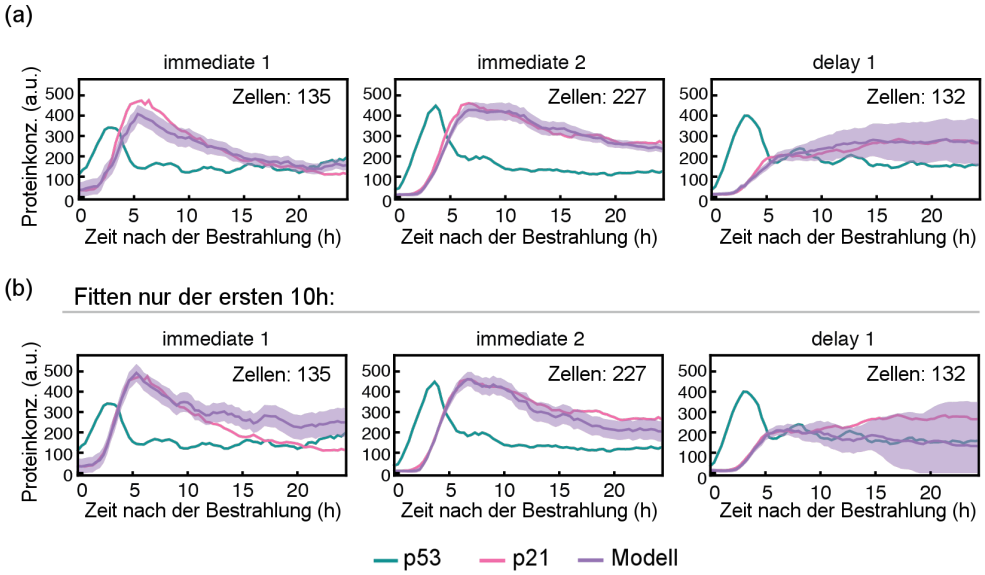


Abbildung 4.12.: Vergleich von simulierten (lila) und gemessenen (magenta) p21-Proteinkonzentrationen für Zellen, deren Zellzyklus in der G1-Phase zum Erliegen kam. In Abb. (a) wurde das Modell (in diesem Fall sind beide betrachteten Modelle identisch) an alle p21-Daten nach der Bestrahlung der Zellen gefittet, während in Abb. (b) nur die p21-Konzentrationen der ersten 10 Stunden nach der Bestrahlung zum Fitten des Modells genutzt wurden.
Die Anzahl der analysierten Zellen in jeder Kategorie ist angegeben. Die lila schattierten Bereiche stellen den Standardfehler des Medians dar, dessen Berechnung in Abb. 4.10 beschrieben ist.
Wie man sieht, ist es vor allem für Zellen, die zur Gruppe „immediate 1“ gehören, nicht möglich, den hohen p21-Peak kurz nach der Bestrahlung und die relativ niedrigen p21-Konzentrationen am Ende der Messung gleichzeitig, d.h. mit demselben Satz von Modellparametern, zu reproduzieren.

4.6. Weitere experimentelle Untersuchungen

Da die Modellierung keine eindeutige Unterscheidung zwischen den beiden potenziellen Mechanismen für die heterogenen p21-Reaktionen ermöglichte, wurden beide Mechanismen von Caibin Sheng und seinen Kollegen experimentell überprüft.

Zunächst untersuchte Caibin Sheng hierbei die p21-Produktionsrate mittels des in Abbildung 4.13 (a) gezeigten Reportersystems. Diese Untersuchung ergab, wie in Abbildung 4.13 (b) zu sehen ist, dass sich die p21-Produktion für Zellen, die in der S-Phase bestrahlt wurden, und für Zellen, die in der G1- oder G2-Phase geschädigt wurden, nach der Bestrahlung quantitativ kaum unterschied. Zudem war auch die Menge der p21-mRNAs pro Zelle für beide Zellgruppen annähernd gleich (siehe Abb. 4.13 (c)). Es ist daher mit großer Sicherheit davon auszugehen, dass zellzyklusspezifische p21-Produktionsraten nicht der Grund für die heterogenen p21-Reaktionen sind.

Wie bereits in Kapitel 4.1 erwähnt wurde, gibt es mehrere zellzyklusabhängige p21-Abbaumechanismen. Unter basalen Bedingungen ist dabei der PCNA/CRL4^{Cdt2}-vermittelte p21-Abbau für die niedrigen p21-Konzentrationen während der S-Phase verantwortlich. Um zu ermitteln, ob auch nach der Bestrahlung dieser Mechanismus für die niedrigen p21-Konzentrationen während der S-Phase verantwortlich ist und somit die beobachtete Heterogenität der p21-Dynamik verursacht, wurde die Interaktion zwischen PCNA und p21 durch

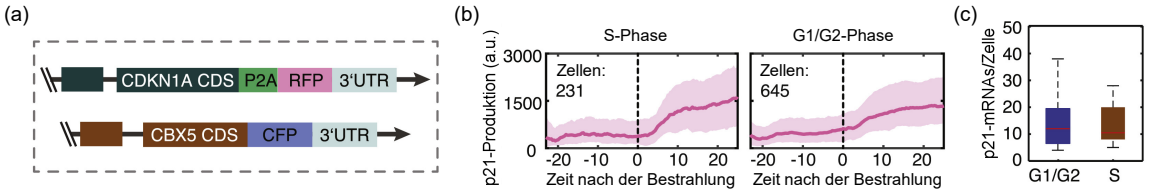


Abbildung 4.13.: (a) Endogenes Reportersystem zur Messung der p21-Produktionsrate. Ein selbstspaltendes Peptid (P2A) wurde zwischen die kodierenden Sequenzen von p21 (CDKN1A CDS) und dem rot fluoreszierenden Protein (RFP) mCherry eingefügt. Da die P2A-Sequenz zur Trennung der beiden Polypeptide während der Translation führt, dient das Signal des rot fluoreszierenden Proteins (RFP) als Reporter für die p21-Produktion.
(b) Vergleich der Produktionsrate von p21 in Zellen die während der S- oder G1/G2-Phase bestrahlt wurden.
Der in Abb. (a) gezeigte Transkriptionsreporter wurde 24 Stunden lang aufgenommen, 30 Minuten lang in EdU-haltigem Medium inkubiert und einer 5 Gy-Bestrahlung ausgesetzt, gefolgt von einer weiteren 24-stündigen Aufnahme und EdU-Nachweis. Die erhaltenen EdU-Signale wurden zur Unterscheidung von Zellen in der S- und G1/G2-Phase verwendet. Basale Signale von mCherry, die die basale Transkription von p21 widerspiegeln, wurden subtrahiert. Die resultierenden Werte zeigen die durch die Bestrahlung induzierte p21-Produktionsrate an. Die Anzahl der jeweils untersuchten Zellen ist angegeben.
(c) Messung der RNA-Konzentrationen in Zellen der S- und G1/G2-Phase 4 Stunden nach der Bestrahlung mittels smFISH und EdU-Markierung.
Quelle:[144]

CRISPR-Cas9 basierte Genom-Editierung verhindert. Dies wurde, wie in Abbildung 4.14 (a) gezeigt, durch eine Mutation der sog. PCNA-interagierenden Peptidbox (PIP-Box) von p21 erreicht.
Im Gegensatz zu den p21-Wildtyp-Zellen ($p21^{wt}$) zeigten die $p21^{PIPmut}$-Zellen eine homogene p21-Reaktion nach der Schadensinduktion (siehe Abb. 4.14 (b)) und auch in Zellen, welche in der S-Phase bestrahlt wurden, stieg die p21-Konzentration ohne nennenswerte Verzögerung an (siehe Abb. 4.14 (c)). Eine anschließende Untersuchung der Schadensreaktion von $p21^{PIPmut}$-Zellen in Abhängigkeit des Zellzyklusverlaufs der untersuchten Zellen bestätigte zudem, dass die Dynamik von $p21^{PIPmut}$ tatsächlich unabhängig von der Zellzyklusphase zum Zeitpunkt der Schädigung und vom anschließenden Fortschreiten des Zellzyklus ist (siehe Abb. 4.14 (d)). Daraus lässt sich zusammenfassend schließen, dass der $PCNA/CRL4^{Cdt2}$-vermittelte p21-Abbau der Hauptgrund für die heterogenen p21-Reaktionen nach der Bestrahlung der Zellen ist.

4.7. Neue Fragestellung und Modellierung der $p21^{PIPmut}$-Daten

Eine neue Fragestellung, die sich durch die in Abbildung 4.14 (d) gezeigten $p21^{PIPmut}$-Verläufe ergibt, ist, wieso die $p21^{PIPmut}$-Konzentrationen nach dem Erreichen ihres Höhepunkts ca. vier Stunden nach der Bestrahlung wieder abfallen: Ist dies lediglich eine Folge der p53-Dynamik oder ein Hinweis auf den Abbau von p21 durch alternative Mechanismen, wie z.B. den SCF^{Skp2}-vermittelten Proteinabbau?

Um diese Frage zu beantworten, habe ich das in Kapitel 4.4 entwickelte Modell (4.2) an die entsprechenden $p21^{PIPmut}$-Einzelzelldaten gefittet und ermittelt, ob dieses einfache Modell ausreicht, um die experimentellen Beobachtungen zu reproduzieren oder ob weitere p21-Abbaumechanismen berücksichtigt werden müssen.
Da wir bei den $p21^{PIPmut}$-Zellen im Gegensatz zu den $p21^{wt}$-Zellen davon ausgehen können, dass die p21-Abbaurate während der S-Phase unverändert bleibt, wurde die p21-Dynamik für alle Zellzyklusphasen durch die gleiche DGL

$$\frac{\mathrm{d}p21(t)}{\mathrm{d}t} = \underbrace{\frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n}}_{\text{p21-Produktion}} - \underbrace{\delta_{p21} \cdot p21(t)}_{\text{p21-Abbau}} \tag{4.3}$$

modelliert. Bis auf die Tatsache, dass in diesem Fall lediglich die ersten 10 Stunden nach der Bestrahlung gefittet wurden (analog zum Vorgehen bei den $p21^{wt}$-Zellen, deren Zellzyklus in der G1- oder G2-Phase zum Erliegen kam) und hier eine etwas geringere zeitliche Verzögerung von $\tau = 1.2$ h angenommen wurde, wurden die Fits auf die gleiche Weise durchgeführt, wie es in Abschnitt 4.4 beschrieben ist.

[5] BrdU ist wie EdU ein Thymidinanalogon, das während der S-Phase in die DNA eingebaut werden kann.

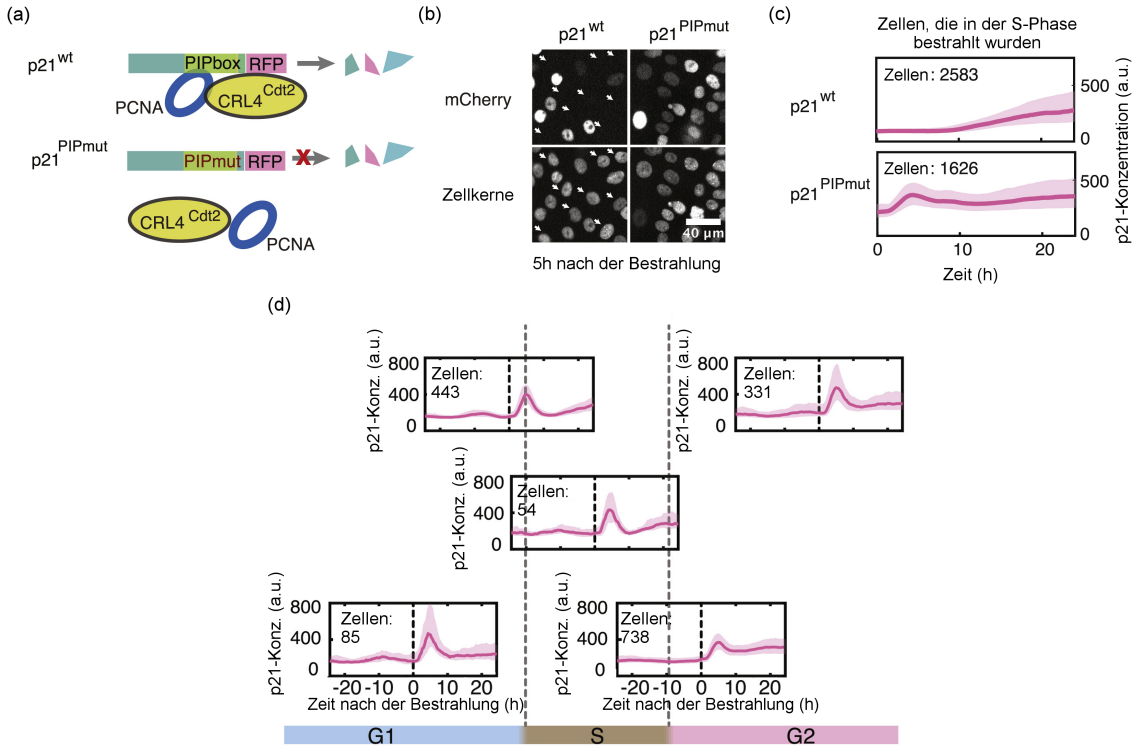


Abbildung 4.14.: (a) Schematische Darstellung des PCNA-vermittelten p21-Abbaus und eines Ansatzes zur gezielten Unterbindung dieses Abbauprozesses.
(b) Vergleich der p21-Akkumulation in $p21^{PIPmut}$- und Kontrollzellen ($p21^{wt}$). Hierzu wurden $p21^{PIPmut}$- und $p21^{wt}$-Zellen mit 5 Gy ionisierender Strahlung bestrahlt und nach 5 Stunden untersucht. Die Pfeile markieren $p21^{wt}$-Zellen mit einer verzögerten Reaktion. $p21^{PIPmut}$ hingegen reichert sich nach der Schädigung in allen Phasen des Zellzyklus an.
(c) Vergleich der p21-Dynamik in $p21^{PIPmut}$- und $p21^{wt}$-Zellen, die während der S-Phase bestrahlt wurden. $p21^{PIPmut}$-Zellen und p21-Wildtyp-Zellen ($p21^{wt}$) wurden für dieses Experiment mit EdU markiert, einer Bestrahlung mit 5 Gy ausgesetzt und 24 Stunden lang beobachtet. Die Linien zeigen den Median der Proteinkonzentrationen der EdU-positiven Zellen in jeder Gruppe im Laufe der Zeit und die schattierten Bereiche kennzeichnen das 25. bis 75. Perzentil. Die Anzahl der untersuchten Zellen ist in der Abbildung angegeben. Wie man sieht, weisen $p21^{PIPmut}$-Zellen, die in der S-Phase geschädigt wurden, im Gegensatz zu $p21^{wt}$-Zellen keine verzögerte p21-Reaktion auf.
(d) Untersuchung der p21-Dynamik in $p21^{PIPmut}$-Zellen in Abhängigkeit des ursprünglichen Zellzykluszustands zum Zeitpunkt der Bestrahlung und dem weiteren Zellzyklusverlauf. Zur Bestimmung des Zellzyklusverlaufs wurden die $p21^{PIPmut}$-Zellen zunächst 24 Stunden lang aufgenommen, 30 Minuten lang in einem BrdU-haltigen[5] Medium inkubiert, dann einer ionisierenden 5-Gy-Strahlung ausgesetzt und weitere 24 Stunden lang aufgenommen, gefolgt von einer 30-minütigen EdU-Inkorporation, Fluoreszenzbleichung und Einzelzell-Immunfluoreszenz zur Detektion von BrdU- und EdU-Signalen. Die jeweilige Anzahl der analysierten Zellen ist angegeben. Man erkennt, dass $p21^{PIPmut}$ im Gegensatz zu $p21^{wt}$ eine homogene Dynamik zeigt, die unabhängig von der Zellzyklusphase zum Zeitpunkt der Bestrahlung und dem weiteren Zellzyklusverlauf ist.
Quelle:[144]

4.8. Ergebnisse zu den p21^{PIPmut}-Daten

Wie in Abbildung 4.15 zu sehen ist, kann mithilfe unseres einfachen Modells (4.3) sowohl die p21^{PIPmut}-Dynamik von Zellen, deren Zellzyklus in der G1- oder G2-Phase stoppte, als auch von Zellen, die in der S-Phase bestrahlt wurden, mit der gleichen Gleichung und überlappenden Parameterverteilungen gefittet werden. Da das Modell nur die p53-vermittelte p21-Produktion sowie den unregulierten Proteinabbau erster Ordnung einschließt, können wir daraus schließen, dass infolge einer Bestrahlung der Beitrag anderer zellzyklusspezifischer p21-Abbaumechanismen in den von Caibin untersuchten MCF10A-Zellen in den ersten 10 Stunden nach der Bestrahlung vernachlässigbar ist.

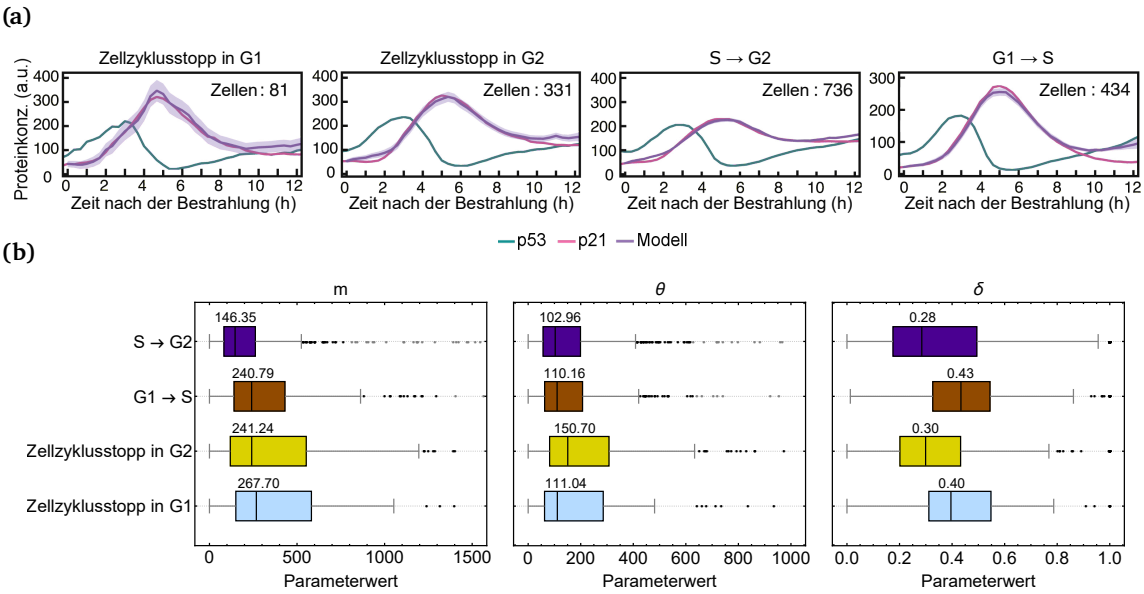


Abbildung 4.15.: (a) Vergleich von simulierten (lila) und gemessenen (magenta) p21^{PIPmut}-Proteinkonzentrationen während der ersten 12 Stunden nach der Bestrahlung. Die Anzahl der analysierten Zellen in jeder Kategorie ist angegeben. Die durchgezogenen Linien stellen stets den Median dar und die lila schattierten Bereiche geben den Standardfehler des Medians an.
(b) Boxplots für die Verteilung der Werte der Modellparameter m (maximale p21-Produktionsrate), θ (Aktivierungsschwelle) und δ (Abbaurate) für die vier verschiedenen zellulären Untergruppen in Abb.(a). Die schwarzen Linien geben die Mediane der Verteilungen an, während die Kästen jeweils das 25. bis 75. Perzentil kennzeichnen. Die Whisker erstrecken sich bis zum Maximalwert innerhalb des 1,5-fachen Interquartilsbereichs. Ausreißer werden als schwarze Punkte dargestellt und entfernte Ausreißer als graue Punkte.

4.9. Fazit und Diskussion

Das primäre Ziel der in diesem Kapitel durchgeführten Modellierung war es in enger Zusammenarbeit mit der experimentellen Gruppe von Prof. Dr. Alexander Löwer im Rahmen des Graduiertenkollegs 1657, den Grund für die Heterogenität der p21-Antwort nach der Induktion von DNA-Doppelstrangbrüchen näher zu bestimmen. Dazu wurden zwei minimalistische Modelle für die Regulierung von p21 durch p53 entwickelt und untersucht. Beide Modelle beruhen jeweils auf einer einzigen nichtlinearen retardierten Differentialgleichung für die Dynamik der p21-Konzentration in einer Zelle und implementieren jeweils unterschiedliche potentielle, in der Literatur vorgeschlagene molekulare Mechanismen zur Erklärung der experimentell beobachteten Heterogenität. Während das eine Modell von einer zellzyklusabhängigen p21-Produktionsrate ausgeht, genauer gesagt von einem kompletten Stopp der p21-Produktion während der S-Phase, beinhaltet das zweite Modell als zentrales Element eine zellzyklusabhängige p21-Abbaurate, bzw. einen erhöhten p21-Abbau während der S-Phase.
Beim quantitativen Vergleich beider Modelle mit den experimentell gemessenen p21-Konzentrationsverläufen stellte sich unerwarteterweise heraus, dass beide Modelle annähernd gleich gut mit den experimentellen Daten vereinbar sind. So reproduzieren beide Modelle die unmittelbare, pulsartige p21-Antwort von Zellen, die in der G1-Phase bestrahlt wurden und anschließend in die S-Phase übergegangen sind, sowie die Verzögerung der p21-Antwort in Zellen, die den Schaden in der S-Phase erfahren haben. Lediglich die Beobachtung, dass der Abfall der p21-Konzertrationen am Übergang von der G1- in die S-Phase durch einen Stopp der p21-Produktion nicht so steil ausfällt wie in den experimentellen Daten und es in diesem Modell konzeptionell auch keine weitere Möglichkeit gibt, die Steilheit des p21-Abfalls zu erhöhen, deutete darauf hin, dass ein erhöhter p21-Abbau während der S-Phase die experimentellen Beobachtungen besser erklären kann.
Diese Vermutung konnte letztendlich durch weitere Experimente, bei denen der in der S-Phase entscheidende, durch PCNA/$CRL4^{Cdt2}$-vermittelte p21-Abbaumechanismus blockiert wurde, bestätigt werden.

In weiteren von Caibin durchgeführten Experimenten stellte sich zudem heraus, dass die Degradation von p21 zu Beginn der S-Phase notwendig ist, um eine zuverlässige Reparatur und Replikation des Genoms zu gewährleisten. So führte die Unterbindung des PCNA/$CRL4^{Cdt2}$-vermittelten p21-Abbaus zu einer erhöhten genomischen Instabilität, die durch eine erhöhte Häufikeit von Chromosomenaberrationen 24 Stunden nach der Bestrahlung gekennzeichnet war [144].

In einer später von Hafner et al. durchgeführten Untersuchung [67] an MCF7-Zellen konnte zusätzlich zur p53- und p21-Proteinkonzentration auch die p21-mRNA-Konzentration mittels fluoreszenzbasierter Lebendzellmikroskopie gemessen werden. In Übereinstimmung mit unseren Experimenten wurde dort ebenfalls festgestellt, dass die p21-Transkription unabhängig von der Zellzyklusphase zum Zeitpunkt der Bestrahlung ist. Somit bestätigt auch diese experimentelle Studie, dass die Hypothese einer zellzyklusabhängigen p21-Transkriptionsrate ausgeschlossen werden kann.

Im Gegensatz zu Caibin Shengs Untersuchung beobachteten Hafner et al. jedoch bei Zellen, die in der G1-Phase bestrahlt wurden und dann in die S-Phase übergingen, keinen pulsförmigen p21-Verlauf, sondern eine zu den in der S-Phase geschädigten Zellen ähnliche, verzögerte p21-Reaktion. Ein weiterer Unterschied zu den in dieser Arbeit vorgestellten experimentellen Beobachtungen war auch die beobachtete p21-Proteindynamik für Zellen, deren Zellzyklus nach der Schädigung in der G1-Phase zum Stillstand kam. Dort nahmen die p21-Konzentrationen während des gesamten Beobachtungszeitraums (der wie in Caibins Experimenten etwa 24 Stunden nach der Induktion von DSBs betrug) stufenförmig zu und fielen nicht wieder ab.
Ein möglicher Grund für diese Unterschiede könnte die Tatsache sein, dass Hafner et al. eine andere Zelllinie, nämlich MCF7-Brustkrebszellen anstatt (gesunder) MCF10A-Brustepithelzellen verwendet haben und ein Teil ihrer Experimente nicht mit ionisierender Strahlung sondern dem radiomimetischen Stoff Neocarzinostatin durchgeführt wurde. Trotzdem bleiben diese Unterschiede überraschend und zeigen, wie schwierig es ist, die komplexe Proteindynamik in einzelnen Zellen tiefgreifend zu verstehen.

Obwohl es im Rahmen der Modellierung leider nicht möglich war, den Mechanismus für die heterogenen p21-Verläufe nach der Induktion von DNA-Doppelstrangbrüchen eindeutig zu identifizieren, so konnten die entwickelten Modelle doch wertvolle Beiträge zur experimentellen Arbeit leisten:

(1) Wir konnten zeigen, dass eine zellzyklusabhängige p21-Abbaurate in Kombination mit der Aktivierung von p21 durch p53 ausreichend ist, um die beobachtete heterogene p21-Antwort nach dem Auftreten von DNA-Doppelstrangbrüchen zu reproduzieren.

(2) Zudem konnte durch die erfolgreiche Reproduktion der $p21^{PIPmut}$-Daten mittels unseres einfachen Modells mit lediglich drei freien Parametern gezeigt werden, dass bis auf den PCNA/$CRL4^{Cdt2}$-vermittelten p21-Abbau keine weiteren zellzyklusspezifischen Abbaumechanismen zur Erklärung der beobachteten p21-Verläufe notwendig sind.

Ein besonderes Kennzeichen der im Rahmen dieses Kapitels entwickelten Modelle ist, dass reale p53-Messdaten als Modellinput dienten und die p53-Dynamik somit nicht aufwändig durch ein System verschiedener Differentialgleichungen, das die Dynamik verschiedener p53-Regulatoren zur Erzeugung der typischen p53-Pulse umfasst, simuliert werden musste. Dies war nur aufgrund der engen interdisziplinären Zusammenarbeit mit der Arbeitsgruppe von Prof. Alexander Löwer und der großen Menge an zeitlich aufgelösten Daten zur Proteindynamik in einzelnen Zellen möglich.

5. Fazit

Die vorliegende Arbeit hat sich mit der Untersuchung zweier medizinisch relevanter, biologischer Systeme beschäftigt, die im Kontext nichtübertragbarer Krankheiten von außerordentlicher Bedeutung sind: Zum einen mit dem menschlichen Mikrobiom und zum anderen mit dem p53-Netzwerk.

Zur Untersuchung des menschlichen Mikrobioms wurde im ersten Teil dieser Dissertation eine neue evolutionsbasierte Methode zur Inferenz mikrobieller Interaktionsnetzwerke, die ESABO-gestützte Evolution, eingeführt. Diese Inferenzmethode beruht auf der ursprünglich von Claussen et al. [34] entwickelten ESABO-Methode und verfolgt den Ansatz, dass ein korrekt inferiertes (Boolesches) Netzwerk die ursprünglichen, zu seiner Inferenz genutzten binären Mikrobiomdaten – welche von uns als Fixpunkte der Dynamik interpretiert werden – als seine Attraktoren reproduzieren sollte.
Im Rahmen der Entwicklung der ESABO-gestützten Evolution konnte zunächst die ursprüngliche ESABO-Methode durch zwei Modifikationen erheblich verbessert werden: Zum einen konnte durch die Einführung einer Formel zur analytischen Berechnung der ESABO-Scores die ESABO-Methode beschleunigt werden, und es konnte sichergestellt werden, dass sie reproduzierbare Ergebnisse liefert. Zum anderen konnte mithilfe der analytischen Berechnung der Grund für das Auftreten von großen negativen ESABO-Scores für positive Interaktionen ermittelt und durch das Vertauschen von Nullen und Einsen in Abundanzvektorpaaren mit einem hohen relativen Anteil von Einsen behoben werden. Diese zweite Modifikation stellt bereits an sich eine deutliche Verbesserung der ursprünglichen ESABO-Methode dar, die es selbst bei Anwendung dieser einfacheren Methode ermöglicht, mutualistische Interaktionen zuverlässig anhand ihres ESABO-Scores zu erkennen.
Anschließend folgte die in dieser Arbeit zentrale Erweiterung der ESABO-Methode um einen evolutionären Algorithmus, mit der das Ziel verfolgt wurde, den Überlapp zwischen der Menge der Attraktoren des inferierten Netzwerks und der Menge der ursprünglichen (beobachteten) binären Abundanzmuster zu maximieren. Wie bei der Untersuchung der Güte der ESABO-gestützten Evolution anhand von zufällig generierten Booleschen Netzwerken mit 15 Knoten gezeigt werden konnte, wurde dieses Ziel erreicht. Netzwerke, die ausgehend von allen Attraktoren mithilfe der ESABO-gestützten Evolution inferiert wurden, wiesen in der Regel wieder die gleichen Attraktoren auf, die zu ihrer Inferenz verwendet wurden (d.h. sie besaßen im Median eine Fitness von $F = 1$) und hatten eine hohe topologische Ähnlichkeit zu den ursprünglichen Netzwerken (Median-Jaccard-Index von $J = 1$) . Zudem war die ESABO-gestützte Evolution deutlich schneller als eine zufällige Evolution, und sie war der einfachen (modifizierten) ESABO-Methode stets überlegen. Dies galt auch in dem Fall, in dem die zur Inferenz genutzten binären Daten nicht alle Attraktoren des Netzwerks umfassten. Selbst wenn nur 50% der Attraktoren eines Netzwerks zu seiner Inferenz verwendet wurden, konnte mit der ESABO-gestützten Evolution noch eine gute

topologische Übereinstimmung des inferierten Netzwerks mit dem Originalnetzwerk erreicht werden (nach 10.000 Evolutionsschritten wurde im Median ein Jaccard-Index von $J = 0.87$ für die Netzwerkkanten erreicht). Diese Untersuchung der Methode bei unvollständiger Kenntnis der Attraktoren ermöglichte es uns außerdem, eine Beziehung zwischen dem Prozentsatz der bekannten Attraktoren und der durchschnittlichen Fitness, die durch unsere Inferenzmethode erreicht wurde, zu finden.
Mithilfe dieser Beziehung konnte schließlich bei der Anwendung der ESABO-gestützten Evolution auf echte, empirische Speichelmikrobiomdaten abgeschätzt werden, dass die verwendeten Daten weniger als 50% der Attraktoren des Systems abdeckten.

Im zweiten Teil dieser Arbeit wurde der Frage nachgegangen, welcher zellzyklusabhängige molekulare Mechanismus dafür verantwortlich ist, dass Zellen nach dem Auftreten von DNA-Doppelstrangbrüchen eine heterogene p21-Dynamik aufweisen, obwohl die Dynamik des p21 regulierenden Transkriptionsfaktors p53 vergleichsweise homogen ist. Zur Beantwortung dieser Frage wurden zwei minimalistische Modelle für die Regulation von p21 durch p53 entwickelt, die jeweils einen unterschiedlichen Mechanismus für die Entstehung der heterogenen p21-Reaktionen implementierten: Während das eine Modell von einer zellzyklusabhängigen p21-Produktionsrate ausging, nahm das zweite Modell eine zellzyklusabhängige Proteinabbaurate als Grund für die Heterogenität an.
Im Rahmen der hier durchgeführten Untersuchung stellte sich heraus, dass beide Modelle quantitativ gleich gut mit den experimentellen Daten vereinbar waren und beide Modelle, die charakteristischen Merkmale des zellzyklusabhängigen dynamischen Verhaltens von p21 nach der Induktion von DNA-Doppelstrangbrüchen reproduzieren konnten. Lediglich die Tatsache, dass eine in der S-Phase erhöhte p21-Abbaurate einen schnelleren Abfall der p21-Konzentration am G1-S-Übergang ermöglichte als eine verminderte Proteinproduktionsrate, wies auf eine höhere Plausibilität der zellzyklusabhängigen Proteinabbaurate als Ursache für die beobachtete Heterogenität hin. Diese Vermutung konnte schließlich durch weitere Experimente bestätigt werden, bei denen eine zellzyklusabhängige p21-Produktionsrate ausgeschlossen und der PCNA/CRL4^{Cdt2}-vermittelte p21-Abbau als für die Heterogenität verantwortlicher Mechanismus identifiziert werden konnte.
Zusammenfassend konnte mithilfe der entwickelten Modelle und den von Caibin Sheng durchgeführten Experimenten also gezeigt werden, dass eine zellzyklusabhängige p21-Abbaurate in Kombination mit der Aktivierung von p21 durch p53 ausreichend ist, um die beobachtete heterogene p21-Antwort nach dem Auftreten von DNA-Doppelstrangbrüchen zu reproduzieren. Die entwickelten Modelle ergänzten die experimentelle Arbeit insbesondere dadurch, dass sie zeigten, dass bis auf den PCNA/CRL4^{Cdt2}-vermittelten p21-Abbau keine weiteren zellzyklusspezifischen Abbaumechanismen zur Erklärung der beobachteten p21-Verläufe notwendig sind.

Eine Gemeinsamkeit der beiden im Rahmen dieser Arbeit durchgeführten Projekte war, dass die Dynamik des jeweilig betrachteten biologischen Systems für seine Untersuchung eine entscheidende Rolle spielte. So berücksichtigt beispielsweise die hier entwickelte ESABO-gestützte-Evolution, dass es sich beim Mikrobiom um ein dynamisches Netzwerk handelt und dass die gemessenen Daten stationäre Zustände eines solchen Netzwerks darstellen sollten.

Eine weitere Gemeinsamkeit war, dass wir uns in beiden Teilen der Arbeit lediglich mit der deterministischen Modellierung der betrachteten biologischen Systeme beschäftigt haben. Gerade im Kontext von Genregulationsnetzwerken können aber auch stochastische Modelle, die den Einfluss von (intrinsischem) Rauschen berücksichtigen, wertvolle Einsichten liefern. Es ist beispielweise bekannt, dass die Transkription von Genen in mRNAs nicht kontinuierlich erfolgt, sondern in sog. Bursts oder Schüben. Wie Falk et al. 2017 anhand eines einfachen Modells zeigen konnten, kann solches bursthaftes Rauschen unter anderem ein monostabiles System bistabil werden lassen [49].
Auch mittels Boolescher Netzwerke kann Rauschen bzw. die Stochastizität biologischer Prozesse berücksichtigt werden. So könnten beispielsweise einzelne Netzwerkknoten nach ihrer deterministischen Aktualisierung mit einer gewissen Wahrscheinlichkeit einen neuen zufälligen Wert zugewiesen bekommen, um zu simulieren, dass eine Spezies durch externe, nicht explizit modellierte Einflüsse in das System eingefügt bzw. entfernt wird.

In dieser Arbeit haben wir uns auf die Untersuchung der betrachteten biologischen Systeme, d.h. des menschlichen Mikrobioms und des p53-Netzwerks, in gesunden Individuen bzw. Zellen beschränkt. Dies stellt selbstverständlich nur einen ersten Schritt auf dem Weg zu einem besseren Verständnis dieser beiden Systeme im Kontext nichtübertragbarer Krankheiten dar, und insbesondere die hier entwickelte Methode zur Inferenz mikrobieller Interaktionsnetzwerke lässt sich auch problemlos auf Mikrobiomdaten von beispielsweise an Diabetes erkrankten Personen anwenden. Zudem wäre es interessant, die Gültigkeit des in dieser Arbeit entwickelten Modells für die zellzyklusabhängige Regulation von p21 durch p53 nach der Induktion von DNA-Doppelstrangbrüchen auch in Krebszelllinien, die Mutationen in dem für p53 oder p21 kodierenden Gen aufweisen, zu überprüfen und eventuelle Unterschiede näher zu untersuchen.

A. Anhang zu Kapitel 3

A.1. Korrelation zwischen Fitness und Jaccard-Index für die Kanten

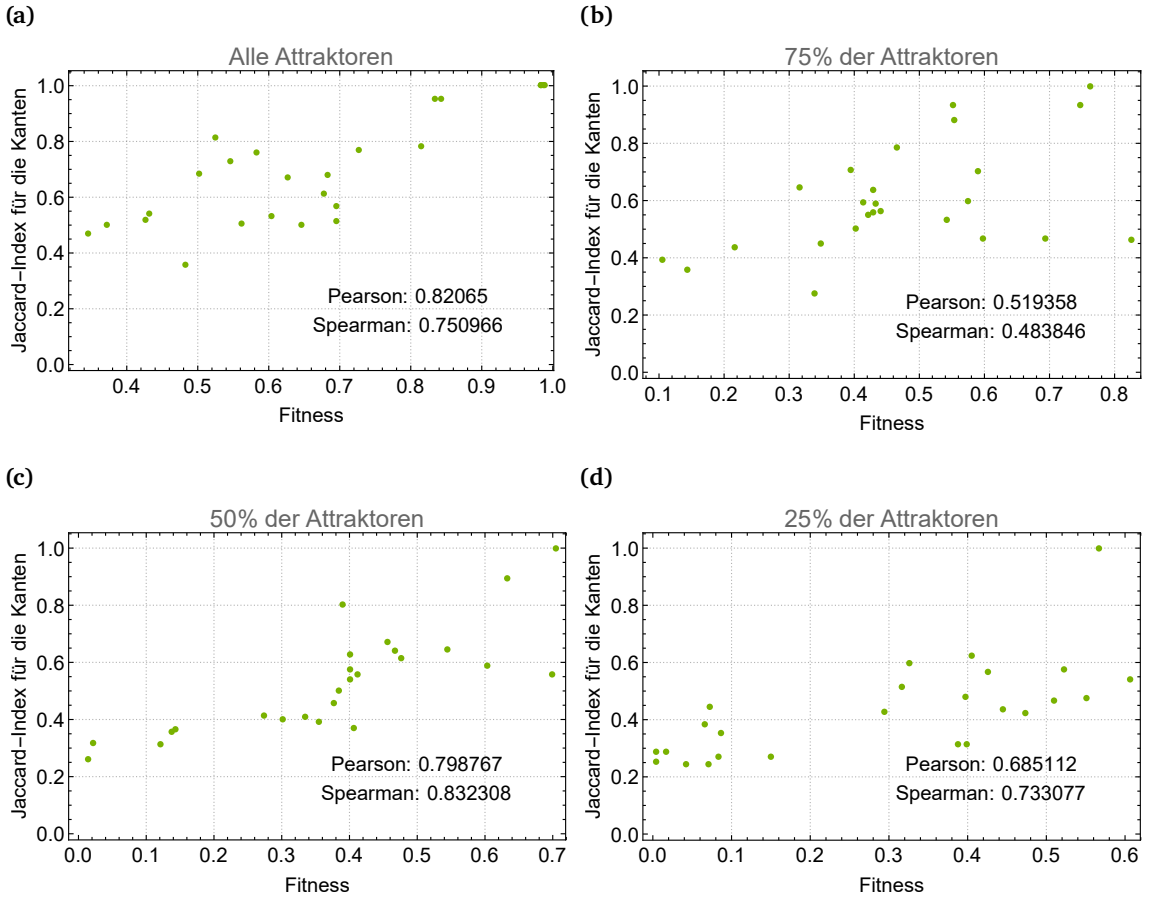


Abbildung A.1.: Korrelation zwischen der Fitness der evolvierten Netzwerke (fitteste Netzwerke nach einer Evolution von 2000 Generationen) aus Abb. 3.15 und dem Jaccard-Index zwischen ihren Kanten und denen der jeweiligen ursprünglichen Netzwerke (Abb. 3.11).

A.2. Analyse des menschlichen Speichelmikrobioms

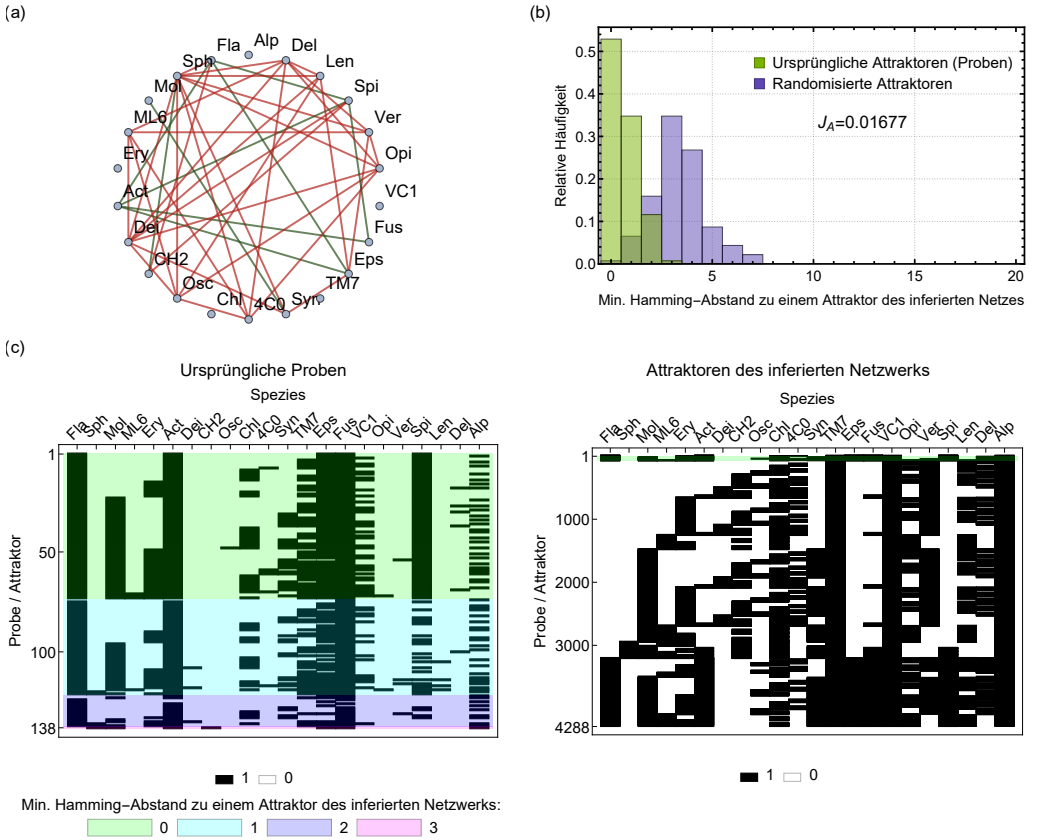


Abbildung A.2.: (a) Vollständiges rekonstruiertes Speichel-Netzwerk (fittestes Netzwerk, das nach einer Evolution von 3000 Generationen gefunden wurde). Abkürzungen: Fla: *Flavobacteria*, Sph: *Sphingobacteria*, Mol: *Mollicutes*, ML6: *ML615J-28*, Ery: *Erysipelotrichi*, Act: *Actinobacteria*, Dei: *Deinococci*, CH2: *CH2*, Osc: *Oscillatoriophycideae*, Chl: *Chloroplast*, 4CO: *4C0d-2*, Syn: *Synergistia*, TM7: *TM7-3*, Eps: *Epsilonproteobacteria*, Fus: *Fusobacteria*, VC1: *VC12-cl04*, Opi: *Opitutae*, Ver: *Verrucomicrobiae*, Spi: *Spirochaetes*, Len: *Lentisphaerae*, Del: *Deltaproteobacteria*, Alp: *Alphaproteobacteria*.
(b) Histogramm, das die minimalen Hamming-Distanzen der ursprünglichen Attraktoren (Proben) oder randomisierten Versionen dieser Attraktoren zu den Attraktoren des rekonstruierten Netzwerks zeigt. Die Randomisierung eines Attraktors wurde durch eine Permutation seiner Einträge durchgeführt. J_A gibt den Jaccard-Index zwischen den Attraktoren des rekonstruierten Netzwerks und den ursprünglichen Proben an.
(c) Vergleich der ursprünglichen, binarisierten Abundanzdaten (Proben) mit den Attraktoren des rekonstruierten Netzwerks. Für jede Probe ist die minimale Hamming-Distanz zu einem Attraktor des rekonstruierten Netzwerks durch die entsprechende Farbe angegeben. Grün markierte Proben werden als Attraktoren des rekonstruierten Netzwerks reproduziert. Das gesamte rekonstruierte Netzwerk hat $2^5 = 32$ mal mehr Attraktoren als seine größte zusammenhängende Komponente, da 5 Knoten völlig unverbunden sind.

A.3. Wahl alternativer Boolescher Aktualisierungsfunktionen

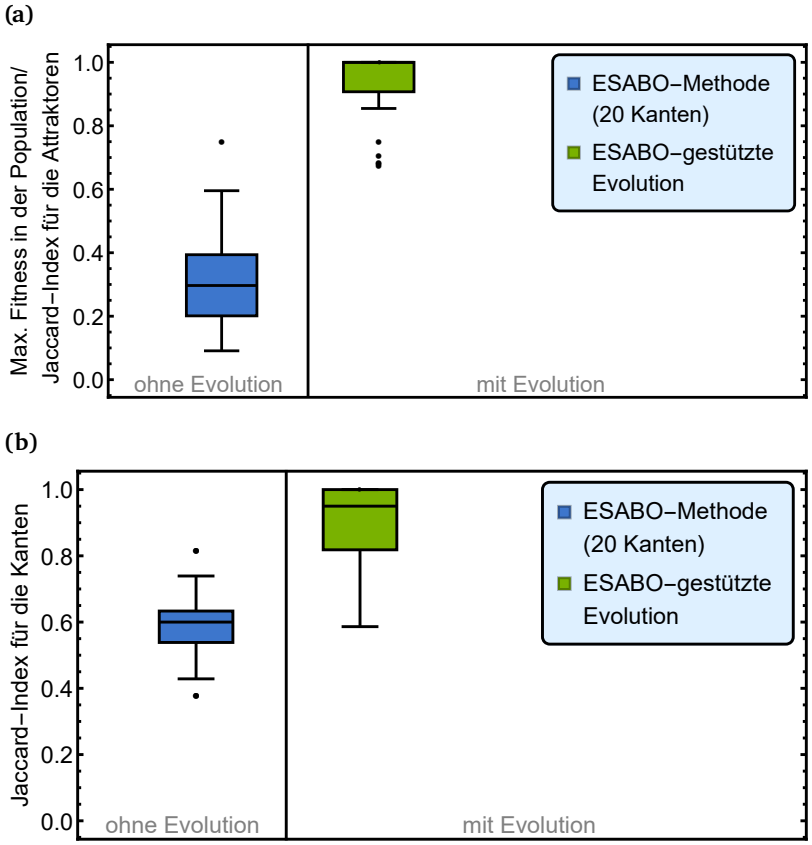


Abbildung A.3.: Ergebnisse mit der durch Gleichung (3.10) beschriebenen alternativen Aktualisierungsfunktion. Hierbei wurden die gleichen 40 Netzwerke mit $N = 15$ Knoten und $L_+ = L_- = 10$ positiven und negativen Kanten wie in Abbildung 3.7 bzw. 3.9 untersucht. Zudem wurde die Evolution auch hier für 10000 Generationen mit $M = 50$, $\nu = 0.25$ und $L_{\min} = 10$ durchgeführt.
(a) Boxplots, die den Jaccard-Index zwischen den Attraktoren des ursprünglichen Netzwerks und den Attraktoren des inferierten oder fittesten evolvierten Netzwerks zeigen, das zu irgendeinem Zeitpunkt während der Evolution gefunden wurde.
(b) Boxplots, die den Jaccard-Index zwischen den Kanten des ursprünglichen Netzwerks und den Kanten des inferierten oder fittesten evolvierten Netzwerks zeigen, das zu irgendeinem Zeitpunkt während der Evolution gefunden wurde.

B. Anhang zu Kapitel 4

B.1. Details zum Fitten der Modelle an die p21-Daten einzelner Zellen

Beide Modelle wurden in Wolfram Mathematica 11 durch unterschiedliche retardierte Differentialgleichungen (DGLen) implementiert.
Für Zellen, die in der G1-Phase bestrahlt wurden und anschließend in die S-Phase übergegangen sind, wurde die p21-Dynamik durch die folgenden DGLen modelliert:

$$\frac{\mathrm{d}p21(t)}{\mathrm{d}t} = \begin{cases} \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta_{p21} \cdot p21(t), & t < t_S \\ -\delta_{p21} \cdot p21(t), & t \geq t_S \end{cases} \tag{B.1}$$

$$\frac{\mathrm{d}p21(t)}{\mathrm{d}t} = \begin{cases} \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta_{p21} \cdot p21(t), & t < t_S \\ \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta_{p21} \cdot D_S \cdot p21(t), & t \geq t_S \end{cases} \tag{B.2}$$

Hierbei gibt t_S den Beginn der S-Phase an.
Für Zellen, die in der S-Phase bestrahlt wurden und in die G2-Phase übergegangen sind, wurde die p21-Dynamik mittels folgender Gleichungen modelliert:

$$\frac{\mathrm{d}p21(t)}{\mathrm{d}t} = \begin{cases} -\delta_{p21} \cdot p21(t), & t < t_S \\ \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta_{p21} \cdot p21(t), & t \geq t_S \end{cases} \tag{B.3}$$

$$\frac{\mathrm{d}p21(t)}{\mathrm{d}t} = \begin{cases} \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta_{p21} \cdot D_S \cdot p21(t), & t < t_S \\ \frac{m \cdot p53(t-\tau)^n}{\theta^n + p53(t-\tau)^n} - \delta_{p21} \cdot p21(t), & t \geq t_S \end{cases} \tag{B.4}$$

In diesem Fall gibt t_S das Ende der S-Phase an.

Mathematica 11 wurde auch verwendet, um die Modelle an die p21-Daten einzelner Zellen zu fitten. Zur Durchführung der Fits wurde die Funktion „NonlinearModelFit“ mit der Methode „NMinimize“ benutzt. Dabei wurde zur Minimierung der Größe $\chi^2 = \sum_i |r_i|^2$ der Nelder-Mead-Algorithmus („NelderMead“) verwendet. Die r_i stellen hierbei Residuen dar, die die Differenz zwischen jedem ursprünglichen Datenpunkt und seinem gefitteten Wert angeben. Die maximale Anzahl der Iterationen („MaxIterations“) wurde zudem auf den Wert 2000 und der „AccuracyGoal“ sowie der „PrecisionGoal“ auf den Wert 50 gesetzt. Die Bedingungen und die Startwerte, die für die Schätzung der Modellparameter verwendet wurden, sind in der nachfolgenden Tabelle aufgeführt:

Parameter	Einheit	Einschränkungen für den Fit	Startwert für den Fit wird zufällig gewählt aus dem Intervall
m	$[m] = \mathrm{C_{a.u.}h^{-1}}$	$0 < m < 100\,\mathrm{Max(p21\text{-}Daten)}$	$[100, 2000]$
θ	$[\theta] = \mathrm{C_{a.u.}}$	$0 < \theta < \mathrm{Max(p53\text{-}Daten)}$	$[0.25\,\mathrm{Max(p53\text{-}Daten)}, \mathrm{Max(p53\text{-}Daten)}]$
δ_{p21}	$[\delta_{p21}] = \mathrm{h^{-1}}$	$0 < \delta_{p21} < 1.0$	$[0, 1.0]$
t_S	$[t_S] = \mathrm{h}$	$23 < t_S < 46$	$[23, 46]$

Tabelle B.1.: Einschränkungen und Startwerte für die Bestimmung der Modellparameter. $\mathrm{C_{a.u.}}$ sind willkürliche Konzentrationseinheiten und t_S bezeichnet den Beginn (für Zellen, die in der G1-Phase bestrahlt wurden und dann in die S-Phase übergegangen sind) oder das Ende (für Zellen, die in der S-Phase bestrahlt wurden und in die G2-Phase übergegangen sind) der S-Phase. Die Zeit wird dabei vom Beginn des Experiments an gezählt.

Für Zellen, die in der S-Phase bestrahlt wurden und in die G2-Phase übergegangen sind, habe ich den Wert der Degradationsrate auf $\delta = 0.2$ gesetzt und nicht gefittet, da ansonsten das Fit-Verfahren sehr lange dauerte und der Algorithmus häufig Lösungen mit $\delta \approx 0$ fand, was aus biologischer Sicht unrealistisch ist. Außerdem wurde in Modell 2 ein fester Wert von $D_S = 50$ angenommen, damit beide Modelle die gleiche Anzahl an freien Parametern aufweisen. Die Verzögerungszeit wurde immer auf $\tau = 1.4$ h gesetzt (sofern nicht anders angegeben) und für den Hill-Koeffizienten ein Wert von $n = 4$ für die p53-abhängige p21-Aktivierung gewählt, da p53 ein Tetramer ist. Um eine gute Fit-Qualität zu gewährleisten und um sicherzustellen, dass die Ergebnisse nicht von den vorgegebenen Anfangswerten für die Fit-Parameter abhängen, wurden für jede betrachtete Zelle 20 Fits mit verschiedenen zufällig gewählten Anfangswerten der Fit-Parameter durchgeführt und der besten Fit für die Mittelwertbildung ausgewählt.

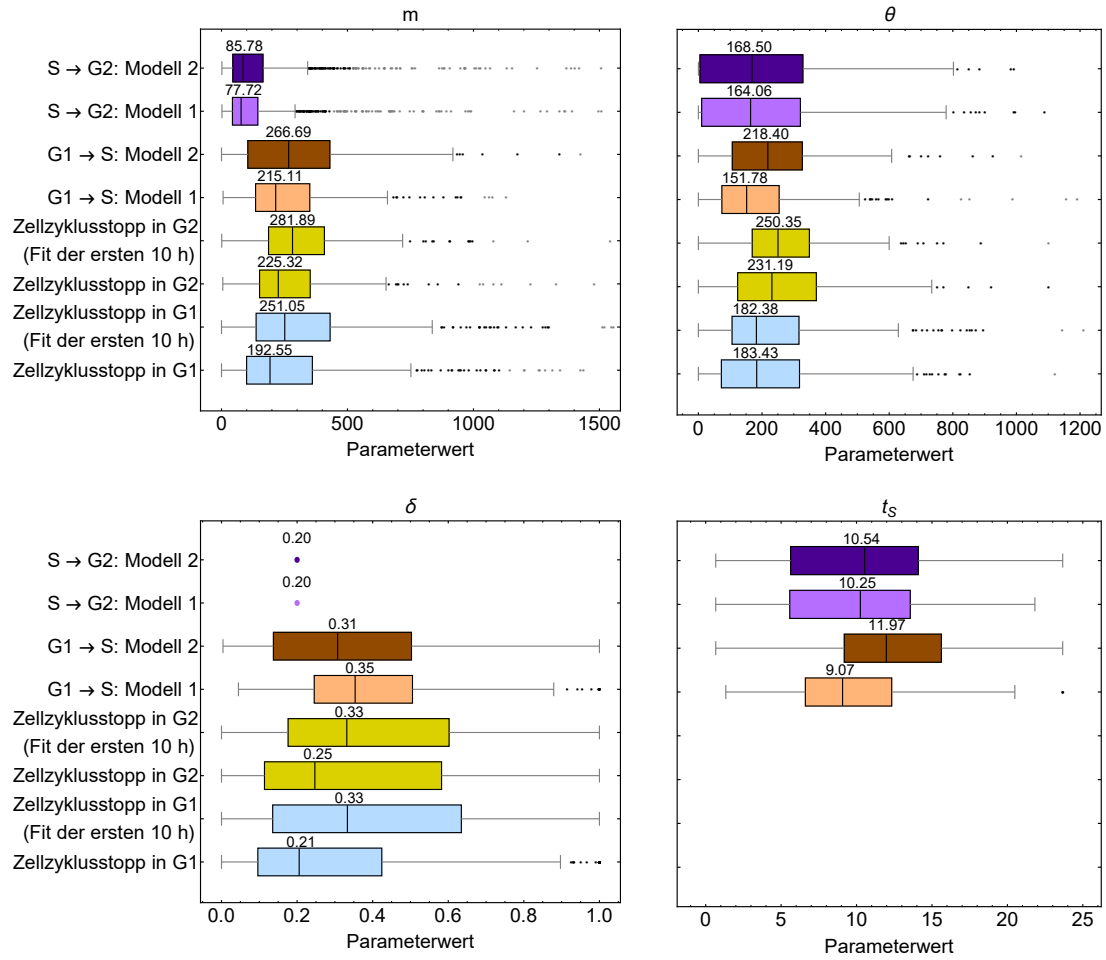


Abbildung B.1.: Boxplots für die Verteilung der Werte der Modellparameter m (maximale p21-Produktionsrate), θ (Aktivierungsschwelle), δ (Abbaurate) und t_S (Ende bzw. Beginn der S-Phase; $t_S = 0$ entspricht dem Zeitpunkt der Bestrahlung) für die verschiedenen zellulären Untergruppen und Modelle. Die schwarzen Linien geben die Mediane der Verteilungen an, während die Kästen jeweils das 25. bis 75. Perzentil kennzeichnen. Die Whisker erstrecken sich bis zum Maximalwert innerhalb des 1,5-fachen Interquartilsbereichs. Ausreißer werden als schwarze Punkte dargestellt und entfernte Ausreißer als graue Punkte.

B.2. Wahl alternativer Werte für D_S

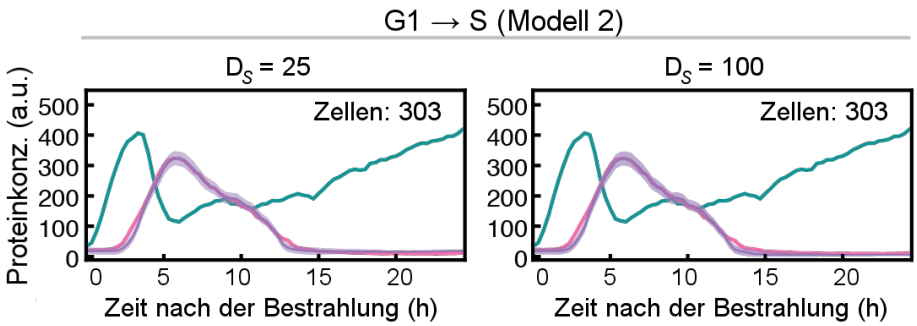


Abbildung B.2.: Vergleich von simulierten (lila) und gemessenen (magenta) p21-Proteinkonzentrationen für unterschiedliche Werte des Parameters D_S in Modell 2. Wie man sieht, spielt der exakte Wert des Parameters keine Rolle für das Ergebnis.

Literatur

[1] Tarek Abbas und Anindya Dutta. „p21 in cancer: intricate networks and multiple activities“. Englisch. In: *Nature Reviews Cancer* 9 (2009), S. 400–414. DOI: https://doi.org/10.1038/nrc2657.

[2] Eva Ackermann. „Vergleich von Boolescher und kontinuierlicher Dynamik auf Genregulationsnetzwerken“. Diss. Technische Universität Darmstadt, 2012.

[3] Jiyoung Ahn und Richard B. Hayes. „Environmental Influences on the Human Microbiome and Implications for Noncommunicable Disease“. In: *Annual Review of Public Health* 42.1 (2021). PMID: 33798404, S. 277–292. DOI: 10.1146/annurev-publhealth-012420-105020.

[4] J. Aitchison. „The Statistical Analysis of Compositional Data“. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2 (1982), S. 139–160. DOI: https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.

[5] Réka Albert. „Network Inference, Analysis, and Modeling in Systems Biology“. In: *The Plant Cell* 19.11 (Dez. 2007), S. 3327–3338. DOI: 10.1105/tpc.107.054700.

[6] Bruce Alberts u. a. *Essential Cell Biology*. Englisch. 4. Aufl. Garland Science, 2013.

[7] Bruce Alberts u. a. *Molecular Biology of the Cell*. 6. Aufl. Garland Science, 2015.

[8] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. Chapman und Hall/CRC, 2006.

[9] Uri Alon. „Network motifs: theory and experimental approaches“. In: *Nature Reviews Genetics* 8.6 (2007), S. 450–461. DOI: 10.1038/nrg2102.

[10] Manimozhiyan Arumugam u. a. „Enterotypes of the human gut microbiome“. In: *nature* 473.7346 (2011), S. 174–180.

[11] Yaneer Bar-Yam. *Dynamics Of Complex Systems*. 1997. URL: https://necsi.edu/dynamics-of-complex-systems (besucht am 17. 02. 2023).

[12] Albert-László Barabási. „The network takeover“. In: *Nature Physics* 8.1 (2012), S. 14–16. DOI: 10.1038/nphys2188.

[13] Albert-László Barabási und Zoltán N. Oltvai. „Network biology: understanding the cell's functional organization“. In: *Nature Reviews Genetics* 5.2 (2004), S. 101–113. DOI: 10.1038/nrg1272.

[14] Y. Barak u. a. „mdm2 expression is induced by wild type p53 activity.“ In: *The EMBO Journal* 12.2 (1993), S. 461–468. DOI: https://doi.org/10.1002/j.1460-2075.1993.tb05678.x.

[15] Thomas Bartz-Beielstein u. a. „Evolutionary Algorithms“. In: *WIREs Data Mining and Knowledge Discovery* 4.3 (2014), S. 178–195. DOI: https://doi.org/10.1002/widm.1124.

[16] Eric Batchelor und Alexander Loewer. „Recent progress and open challenges in modeling p53 dynamics in single cells.“ In: *Current opinion in systems biology* 3 (2017), S. 54–59. DOI: 10.1016/j.coisb.2017.04.007.

[17] Eric Batchelor, Alexander Löwer und Galit Lahav. „The ups and downs of p53: understanding protein dynamics in single cells“. In: *Nature Reviews Cancer* 9 (2009), S. 371–377. DOI: https://doi.org/10.1038/nrc2604.

[18] Rachel Beckerman u. a. „A role for Chk1 in blocking transcriptional elongation of p21 RNA during the S-phase checkpoint“. In: *Genes and Development* 23.11 (2009), S. 1364–1377. DOI: 10.1101/gad.1795709.

[19] Daniel Belstrøm u. a. „Temporal Stability of the Salivary Microbiota in Oral Health“. In: *PLOS ONE* 11.1 (Jan. 2016), S. 1–9. DOI: 10.1371/journal.pone.0147472.

[20] David R. Bentley u. a. „Accurate whole human genome sequencing using reversible terminator chemistry“. In: *Nature* 456.7218 (2008), S. 53–59. DOI: 10.1038/nature07517.

[21] J. Gordon Betts u. a. *Anatomy and Physiology*. Houston, Texas: OpenStax, Nov. 2021. URL: https://openstax.org/books/anatomy-and-physiology/pages/1-introduction (besucht am 22. 11. 2021).

[22] Eva Bianconi u. a. „An estimation of the number of cells in the human body“. In: *Annals of Human Biology* 40.6 (2013). PMID: 23829164, S. 463–471. DOI: 10.3109/03014460.2013.807878.

[23] F Guillaume Blanchet, Kevin Cazelles und Dominique Gravel. „Co-occurrence is not evidence of ecological interactions“. In: *Ecology Letters* 23.7 (2020), S. 1050–1063.

[24] Isabella-Hilda Bodea. „Evolution einer Population Boolescher Netzwerke“. Masterarbeit. TU Darmstadt, 2016.

[25] Stefan Bornholdt. „Less Is More in Modeling Large Genetic Networks“. In: *Science* 310.5747 (2005), S. 449–451. DOI: 10.1126/science.1119959.

[26] Andrew P. Bradley. „The use of the area under the ROC curve in the evaluation of machine learning algorithms“. In: *Pattern Recognition* 30.7 (1997), S. 1145–1159. DOI: https://doi.org/10.1016/S0031-3203(96)00142-2.

[27] I.N. Bronstein u. a. *Taschenbuch der Mathematik*. 8. Aufl. Harri Deutsch Verlag, 2012.

[28] James Brown u. a. „Translating the human microbiome“. In: *Nature biotechnology* 31.4 (2013), S. 304–308.

[29] Fred Bunz u. a. „Requirement for p53 and p21 to sustain G2 arrest after DNA damage“. In: *Science* 282.5393 (1998), S. 1497–1501.

[30] J. Gregory Caporaso u. a. „QIIME allows analysis of high-throughput community sequencing data“. In: *Nature Methods* 7.5 (2010), S. 335–336. DOI: 10.1038/nmeth.f.303.

[31] Marco Cappellato u. a. „Modeling Microbial Community Networks: Methods and Tools.“ eng. In: *Current genomics* 22 (4 Dez. 2021), S. 267–290. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8822226/.

[32] Sheng-hong Chen, William Forrester und Galit Lahav. „Schedule-dependent interaction between anticancer treatments“. In: *Science* 351.6278 (2016), S. 1204–1208. DOI: 10.1126/science.aac5610.

[33] Ket Hing Chong u. a. „Mathematical modelling of p53 basal Dynamics and DNA damage response“. In: *Mathematical Biosciences* 259.14728 (Mai 2015), S. 27–42. DOI: 10.1038/ncomms14728.

[34] Jens Christian Claussen u. a. „Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome“. In: *PLoS computational biology* 13.6 (2017), e1005361. DOI: https://doi.org/10.1371/journal.pcbi.1005361.

[35] Jose C Clemente u. a. „The impact of the gut microbiota on human health: an integrative view“. In: *Cell* 148.6 (2012), S. 1258–1270.

[36] James R. Cole u. a. „Ribosomal Database Project: data and tools for high throughput rRNA analysis“. In: *Nucleic Acids Research* 42.D1 (Nov. 2013), S. D633–D642. DOI: 10.1093/nar/gkt1244.

[37] Kate E. Coleman u. a. „Sequential replication-coupled destruction at G1/S ensures genome stability“. In: *Genes and Development* 29.16 (2015), S. 1734–1746. DOI: 10.1101/gad.263731.115.

[38] *Country profile for Germany generated from the World Health Organization NCD Data Portal*. URL: https://ncdportal.org/CountryProfile/GHE110/DEU#mor2 (besucht am 11. 05. 2023).

[39] Maria I. Davidich und Stefan Bornholdt. „Boolean Network Model Predicts Cell Cycle Sequence of Fission Yeast“. In: *PLoS ONE* 3(2) (2008), e1672.

[40] Hidde De Jong. „Modeling and Simulation of Genetic Regulatory Systems: A Literature Review“. In: *Journal of Computational Biology* 9.1 (2002), S. 67–103.

[41] Wafik S. El-Deiry u. a. „WAF1, a potential mediator of p53 tumor suppression“. In: *Cell* 75.4 (1993), S. 817–825. DOI: 10.1016/0092-8674(93)90500-P.

[42] Eva Delmas u. a. „Analysing ecological networks of species interactions“. In: *Biological Reviews* 94.1 (2019), S. 16–36. DOI: https://doi.org/10.1111/brv.12433.

[43] Chuxia Deng u. a. „Mice lacking p21CIP1/WAF1 undergo normal development, but are defective in G1 checkpoint control“. In: *Cell* 82.4 (1995), S. 675–684.

[44] T. Z. DeSantis u. a. „Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB“. In: *Applied and Environmental Microbiology* 72.7 (2006), S. 5069–5072. DOI: 10.1128/AEM.03006-05.

[45] Peter Deuflhard und Susanna Röblitz. *A Guide to Numerical Modelling in Systems Biology*. 1. Aufl. Springer Cham, 2015, S. 12. URL: https://aproxy.ulb.tu-darmstadt.de:2062/10.1007/978-3-319-20059-0.

[46] Barbara Drossel. *Komplexe Dynamische Systeme*. Unveröffentlichtes Skript WiSe 2014/15. Institut für Festkörperphysik, Technische Universität Darmstadt. 2014.

[47] Barbara Drossel. „Random Boolean Networks“. In: *Reviews of Nonlinear Dynamics and Complexity*. Bd. 1. Wiley, 2007.

[48] Juliana Durack und Susan V Lynch. „The gut microbiome: relationships with disease and opportunities for therapy“. In: *Journal of Experimental Medicine* 216.1 (2019), S. 20–40.

[49] Johannes Falk, Marc Mendler und Barbara Drossel. „A minimal model of burst-noise induced bistability“. In: *PLOS ONE* 12.4 (Apr. 2017), S. 1–15. DOI: 10.1371/journal.pone.0176410.

[50] Karoline Faust. „Open challenges for microbial network construction and analysis“. In: *The ISME Journal* 15 (11 2021), S. 3111–3118. DOI: 10.1038/s41396-021-01027-4.

[51] Karoline Faust und Jeroen Raes. „Microbial interactions: from networks to models“. In: *Nature Reviews Microbiology* 10 (2012), S. 538–550. DOI: https://doi.org/10.1038/nrmicro2832.

[52] Tom Fawcett. „An introduction to ROC analysis“. In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, S. 861–874. DOI: https://doi.org/10.1016/j.patrec.2005.10.010.

[53] Peter Filzmoser, Karel Hron und Matthias Templ. *Applied Compositional Data Analysis. With Worked Examples in R*. Springer Cham, 2018. URL: https://aproxy.ulb.tu-darmstadt.de:2062/10.1007/978-3-319-96422-5.

[54] Michele Fiscella u. a. „Wip1, a novel human protein phosphatase that is induced in response to ionizing radiation in a p53-dependent manner“. In: *Proceedings of the National Academy of Sciences* 94.12 (1997), S. 6048–6053. DOI: 10.1073/pnas.94.12.6048.

[55] M. Fischer. „Census and evaluation of p53 target genes“. In: *Oncogene* 36.28 (2017), S. 3943–3956. DOI: 10.1038/onc.2016.502.

[56] Laura Friedel und Alexander Loewer. „The guardian's choice: how p53 enables context-specific decision-making in individual cells“. In: *The FEBS Journal* 289.1 (2022), S. 40–52. DOI: https://doi.org/10.1111/febs.15767.

[57] Jonathan Friedman und Eric J. Alm. „Inferring Correlation Networks from Genomic Survey Data“. In: *PLOS Computational Biology* 8.9 (Sep. 2012), S. 1–11. DOI: 10.1371/journal.pcbi.1002687.

[58] Olaf Fritsche. *Mikrobiologie*. Springer Spektrum Berlin, Heidelberg, 2016. DOI: https://aproxy.ulb.tu-darmstadt.de:2062/10.1007/978-3-662-49729-6.

[59] Kei E. Fujimura und Susan V. Lynch. „Microbiota in Allergy and Asthma and the Emerging Relationship with the Gut Microbiome“. In: *Cell Host and Microbe* 17.5 (2015), S. 592–602. DOI: https://doi.org/10.1016/j.chom.2015.04.007.

[60] Joshua Garcia und Jenny Kao-Kniffin. „Can dynamic network modelling be used to identify adaptive microbiomes?“ In: *Functional Ecology* 34.10 (2020), S. 2065–2074.

[61] Beatriz García-Jiménez und Mark D Wilkinson. „Robust and automatic definition of microbiome states“. In: *PeerJ* 7 (2019), e6657.

[62] Jordi Garcia-Ojalvo. „Physical approaches to the dynamics of genetic circuits: a tutorial“. In: *Contemporary Physics* 52.5 (2011), S. 439–464. DOI: 10.1080/00107514.2011.588432.

[63] Melanie Ghoul und Sara Mitri. „The Ecology and Evolution of Microbial Competition“. In: *Trends in Microbiology* 24.10 (2016), S. 833–845. DOI: https://doi.org/10.1016/j.tim.2016.06.011.

[64] Jack A. Gilbert u. a. „Current understanding of the human microbiome.“ In: *Nature medicine* 24 (4 2018), S. 392–400. DOI: 10.1038/nm.4517. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7043356/.

[65] Klaus Gotthardt. „Grundlagen der Informationstechnik“. In: 2. Aufl. Münster: LIT Verlag, 2005, S. 40.

[66] Vanesa Gottifredi u. a. „p53 accumulates but is functionally impaired when DNA synthesis is blocked“. In: *Proceedings of the National Academy of Sciences* 98.3 (2001), S. 1036–1041. DOI: 10.1073/pnas.98.3.1036.

[67] Antonina Hafner u. a. „Quantifying the Central Dogma in the p53 Pathway in Live Single Cells“. In: *Cell Systems* 10.6 (Juni 2020), 495–505.e4. DOI: 10.1016/j.cels.2020.05.001.

[68] Antonina Hafner u. a. „The multiple mechanisms that regulate p53 activity and cell fate“. In: *Nature Reviews Molecular Cell Biology* 20.4 (2019), S. 199–210. DOI: 10.1038/s41580-019-0110-x.

[69] Richard W Hamming. *Coding and information theory*. Prentice-Hall, Inc., 1986.

[70] Annick V. Hartstra u. a. „Insights Into the Role of the Microbiome in Obesity and Type 2 Diabetes“. In: *Diabetes Care* 38.1 (Dez. 2014), S. 159–165. DOI: 10.2337/dc14-0769.

[71] Ygal Haupt u. a. „Mdm2 promotes the rapid degradation of p53“. In: *Nature* 387.6630 (Mai 1997), S. 296–299. URL: https://doi.org/10.1038/387296a0.

[72] Jürgen Hedderich und Lothar Sachs. *Angewandte Statistik. Methodensammlung mit R*. 17. Aufl. Springer Spektrum Berlin, Heidelberg, 2020. DOI: https://doi.org/10.1007/978-3-662-62294-0.

[73] Beth A. Helmink u. a. „The microbiome, cancer, and cancer therapy“. In: *Nature Medicine* 25.3 (2019), S. 377–388. DOI: 10.1038/s41591-019-0377-7.

[74] Michael E. Hibbing u. a. „Bacterial competition: surviving and thriving in the microbial jungle“. In: *Nature Reviews Microbiology* 8.1 (2010), S. 15–25. DOI: 10.1038/nrmicro2259.

[75] Martin Hopfensitz u. a. „Attractors in Boolean networks: a tutorial“. In: *Computational Statistics* 28 (1. Feb. 2013), S. 19–36. DOI: 10.1007/s00180-012-0324-2.

[76] Patrick D. Hsu, Eric S. Lander und Feng Zhang. „Development and Applications of CRISPR-Cas9 for Genome Engineering“. In: *Cell* 157 (6 2014), S. 1262–1278. DOI: https://doi.org/10.1016/j.cell.2014.05.010.

[77] Yetao Jin u. a. „MDMX Promotes Proteasomal Turnover of p21 at G₁ and Early S Phases Independently of, but in Cooperation with, MDM2“. In: *Molecular and Cellular Biology* 28.4 (2008), S. 1218–1229. DOI: 10.1128/MCB.01198-07.

[78] Martin Jinek u. a. „A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity“. In: *Science* 337.6096 (2012), S. 816–821. DOI: 10.1126/science.1225829.

[79] Guy Karlebach und Ron Shamir. „Modelling and analysis of gene regulatory networks“. In: *Nature Reviews Molecular Cell Biology* 9 (Okt. 2008), S. 770–780. DOI: https://doi.org/10.1038/nrm2503.

[80] Fredrik H. Karlsson u. a. „Gut metagenome in European women with normal, impaired and diabetic glucose control“. In: *Nature* 498.7452 (2013), S. 99–103. DOI: 10.1038/nature12198.

[81] Edward R. Kastenhuber und Scott W. Lowe. „Putting p53 in Context“. In: *Cell* 170.6 (Sep. 2017), S. 1062–1078. DOI: 10.1016/j.cell.2017.08.028.

[82] Stuart A. Kauffman. „Metabolic stability and epigenesis in randomly constructed genetic nets.“ In: *Journal of Theoretical Biology* 22 (1969), S. 437–467.

[83] D P Kelly. „Autotrophy: Concepts of Lithotrophic Bacteria and their Organic Metabolism“. In: *Annual Review of Microbiology* 25.1 (1971). PMID: 4342704, S. 177–210. DOI: 10.1146/annurev.mi.25.100171.001141.

[84] Donovan P. Kelly und Ann P. Wood. „The Chemolithotrophic Prokaryotes“. In: *The Prokaryotes: Volume 2: Ecophysiology and Biochemistry*. Hrsg. von Martin Dworkin u. a. New York, NY: Springer New York, 2006, S. 441–456. ISBN: 978-0-387-30742-8. DOI: 10.1007/0-387-30742-7_15.

[85] Hiroaki Kitano. „Systems Biology: A Brief Overview“. In: *Science* 295.5560 (2002), S. 1662–1664. DOI: 10.1126/science.1069492.

[86] Malka Kitayner u. a. „Structural Basis of DNA Recognition by p53 Tetramers“. In: *Molecular Cell* 22.6 (2006), S. 741–753. DOI: https://doi.org/10.1016/j.molcel.2006.05.015.

[87] Edda Klipp u. a. *Systems Biology in Practice*. WILEY-VCH Verlag, 2005.

[88] Zachary D. Kurtz u. a. „Sparse and Compositionally Robust Inference of Microbial Ecological Networks“. In: *PLOS Computational Biology* 11.5 (Mai 2015), S. 1–25. DOI: 10.1371/journal.pcbi.1004226.

[89] Galit Lahav u. a. „Dynamics of the p53-Mdm2 feedback loop in individual cells“. In: *Nature Genetics* 36.2 (Feb. 2004), S. 147–150. URL: https://doi.org/10.1038/ng1293.

[90] D. P. Lane. „p53, guardian of the genome“. In: *Nature* 358.6381 (1992), S. 15–16. DOI: 10.1038/358015a0.

[91] David Lane und Arnold Levine. „p53 Research: the past thirty years and the next thirty years“. In: *Cold Spring Harbor perspectives in biology* 2.12 (2010), a000893. URL: https://cshperspectives.cshlp.org/content/2/12/a000893.full.

[92] Nadja Larsen u. a. „Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults“. In: *PLOS ONE* 5.2 (Feb. 2010), S. 1–10. DOI: 10.1371/journal.pone.0009085.

[93] Peter E. Larsen, Sean M. Gibbons und Jack A. Gilbert. „Modeling microbial community structure and functional diversity across time and space“. In: *FEMS Microbiology Letters* 332.2 (Juli 2012), S. 91–98. DOI: 10.1111/j.1574-6968.2012.02588.x.

[94] Mehdi Layeghifard, David M. Hwang und David S. Guttman. „Disentangling Interactions in the Microbiome: A Network Perspective“. In: *Trends in Microbiology* 25.3 (2017), S. 217–228. DOI: https://doi.org/10.1016/j.tim.2016.11.008.

[95] Vladimir Lazarevic, Katrine Whiteson, David Hernandez u. a. „Study of inter- and intra-individual variations in the salivary microbiota“. In: *BMC Genomics* (2010). DOI: https://doi.org/10.1186/1471-2164-11-523.

[96] Roman Nicor Lengert. „Regulation von zellulären Proteinnetzwerken nach Stresssignalen“. Diss. Technische Universität Darmstadt, 2017. URL: http://tuprints.ulb.tu-darmstadt.de/6966/.

[97] Arnold J. Levine. „Exploring the future of research in the Tp53 field“. In: *Cell Death and Differentiation* 29.5 (2022), S. 893–894. DOI: 10.1038/s41418-022-00986-1.

[98] Arnold J. Levine. „Targeting the P53 Protein for Cancer Therapies: The Translational Impact of P53 Research“. In: *Cancer Research* 82.3 (Feb. 2022), S. 362–364. DOI: 10.1158/0008-5472.CAN-21-2709.

[99] Arnold J. Levine und Moshe Oren. „The first 30 years of p53: growing ever more complex“. In: *Nature Reviews Cancer* 9.10 (2009), S. 749–758. DOI: 10.1038/nrc2723.

[100] Fangting Li u. a. „The yeast cell-cycle network is robustly designed“. In: *Proceedings of the National Academy of Sciences of the United States of America* 101(14) (2004), S. 4781–4786.

[101] Shoudan Liang, Stefanie Fuhrman, Roland Somogyi u. a. „Reveal, a general reverse engineering algorithm for inference of genetic network architectures“. In: *Pacific symposium on biocomputing*. Bd. 3. 3. Citeseer. 1998, S. 18–29.

[102] Alexander Löwer u. a. „Basal Dynamics of p53 Reveal Transcriptionally Attenuated Pulses in Cycling Cells“. In: *Cell* 142.14728 (Juli 2010), S. 89–100. DOI: 10.1016/j.cell.2010.05.031.

[103] Elina Ly, Jennifer F. Kugel und James A. Goodrich. „Single molecule studies reveal that p53 tetramers dynamically bind response elements containing one or two half sites“. In: *Scientific Reports* 10.1 (2020), S. 16176. DOI: 10.1038/s41598-020-73234-6.

[104] Avi Ma'ayan. „Complex systems biology“. In: *Journal of The Royal Society Interface* 14.134 (2017), S. 20170391. DOI: 10.1098/rsif.2017.0391.

[105] Ohad Manor u. a. „Health and disease markers correlate with gut microbiome composition across thousands of people“. In: *Nature communications* 11.1 (2020), S. 1–12.

[106] Sabrina F Mansilla u. a. „UV-triggered p21 degradation facilitates damaged-DNA replication and preserves genomic stability“. In: *Nucleic acids research* 41.14 (2013), S. 6942–6951.

[107] Marcel Margulies u. a. „Genome sequencing in microfabricated high-density picolitre reactors“. In: *Nature* 437.7057 (2005), S. 376–380. DOI: 10.1038/nature03959.

[108] Denis Mariat u. a. „The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age“. In: *BMC Microbiology* 9.1 (2009), S. 1–6.

[109] Monica Steffi Matchado u. a. „Network analysis methods for studying microbial communities: A mini review“. In: *Computational and structural biotechnology journal* (2021).

[110] Melissa Mattia u. a. „p53-Dependent p21 mRNA Elongation Is Impaired when DNA Replication Is Stalled“. In: *Molecular and Cellular Biology* 27.4 (2007), S. 1309–1320. DOI: 10.1128/MCB.01520-06.

[111] Kevin G. McLure und Patrick W.K. Lee. „How p53 binds DNA as a tetramer“. In: *The EMBO Journal* 17.12 (1998), S. 3342–3350. DOI: https://doi.org/10.1093/emboj/17.12.3342.

[112] Paul J. McMurdie und Susan Holmes. „phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data“. In: *PLOS ONE* 8.4 (Apr. 2013), S. 1–11. DOI: 10.1371/journal.pone.0061217.

[113] Isabella-Hilda Mendler, Barbara Drossel und Marc-Thorsten Hütt. *Microbiome abundance patterns as attractors and the implications for the inference of microbial interaction networks*. 2023. arXiv: 2306.02100 [q-bio.PE].

[114] Nancy Merino u. a. „Living at the Extremes: Extremophiles and the Limits of Life in a Planetary Context“. In: *Frontiers in Microbiology* 10 (2019). DOI: 10.3389/fmicb.2019.00780.

[115] Tamara Mihaljev. „Dynamics and evolution of random Boolean networks“. Diss. Technische Universität Darmstadt, 2008.

[116] Melanie Mitchell. „Complex systems: Network thinking“. In: *Artificial Intelligence* 170.18 (2006). Special Review Issue, S. 1194–1212. DOI: https://doi.org/10.1016/j.artint.2006.10.002.

[117] Melanie Mitchell. *Complexity: A guided tour*. Oxford university press, 2009.

[118] Ute M. Moll und Oleksi Petrenko. „The MDM2-p53 Interaction“. In: *Molecular Cancer Research* 1.14 (Dez. 2003), S. 1001–1008.

[119] Gregor Mönke u. a. „Excitability in the p53 network mediates robust signaling with tunable activation thresholds in single cells“. In: *Scientific Reports* 7.1 (2017), S. 46571. DOI: 10.1038/srep46571.

[120] Xochitl C. Morgan und Curtis Huttenhower. „Chapter 12: Human Microbiome Analysis“. In: *PLOS Computational Biology* 8.12 (Dez. 2012), S. 1–14. DOI: 10.1371/journal.pcbi.1002808.

[121] Patricia A. J. Muller und Karen H. Vousden. „p53 mutations in cancer“. In: *Nature Cell Biology* 15.1 (2013), S. 2–8. DOI: 10.1038/ncb2641.

[122] Daniel Muñoz-Espín und Manuel Serrano. „Cellular senescence: from physiology to pathology“. In: *Nature reviews Molecular cell biology* 15.7 (2014), S. 482–496.

[123] Sunil Nagpal u. a. „MetagenoNets: comprehensive inference and meta-insights for microbial correlation networks“. In: *Nucleic Acids Research* 48.W1 (2020), W572–W579.

[124] John Paparrizos und Luis Gravano. „K-Shape: Efficient and Accurate Clustering of Time Series“. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: Association for Computing Machinery, 2015, S. 1855–1870. ISBN: 9781450327589. DOI: 10.1145/2723372.2737793. URL: https://doi.org/10.1145/2723372.2737793.

[125] Stephen R. Proulx, Daniel E.L. Promislow und Patrick C. Phillips. „Network thinking in ecology and evolution“. In: *Trends in Ecology and Evolution* 20.6 (2005). SPECIAL ISSUE: BUMPER BOOK REVIEW, S. 345–353. DOI: https://doi.org/10.1016/j.tree.2005.04.004.

[126] Foster Provost und Tom Fawcett. „Robust Classification for Imprecise Environments“. In: *Machine Learning* 42.3 (2001), S. 203–231. DOI: 10.1023/A:1007601015854.

[127] Jeremy E Purvis u. a. „p53 dynamics control cell fate“. In: *Science* 336.6087 (2012), S. 1440–1444. DOI: 10.1126/science.1218351.

[128] Junjie Qin, Ruiqiang Li, Jeroen Raes u. a. „A human gut microbial gene catalogue established by metagenomic sequencing“. In: *Nature* 464.7285 (2010), S. 59–65.

[129] Junjie Qin u. a. „A metagenome-wide association study of gut microbiota in type 2 diabetes“. In: *Nature* 490.7418 (2012), S. 55–60. DOI: 10.1038/nature11450.

[130] Christian Quast u. a. „The SILVA ribosomal RNA gene database project: improved data processing and web-based tools“. In: *Nucleic Acids Research* 41.D1 (Nov. 2012), S. D590–D596. DOI: 10.1093/nar/gks1219.

[131] Nicole E. Radde und Marc-Thorsten Hütt. „The Physics behind Systems Biology“. In: *EPJ Nonlinear Biomedical Physics* 4 (7 2016). DOI: https://doi.org/10.1140/epjnbp/s40366-016-0034-8.

[132] F Ann Ran u. a. „Genome engineering using the CRISPR-Cas9 system“. In: *Nature Protocols* 8 (2013), S. 2281–2308. DOI: https://doi.org/10.1038/nprot.2013.143.

[133] Lisa Röttjers und Karoline Faust. „From hairballs to hypotheses–biological insights from microbial networks“. In: *FEMS microbiology reviews* 42.6 (2018), S. 761–780.

[134] Marc R Roussel. „Delay-differential equations“. In: *Nonlinear Dynamics*. 2053-2571. Morgan & Claypool Publishers, 2019, 12-1 to 12–14. DOI: 10.1088/2053-2571/ab0281ch12.

[135] David Sadava u. a. *Purves Biologie*. Hrsg. von Jürgen Markl. 10. Aufl. Springer Spektrum Berlin, Heidelberg, 2019. DOI: https://doi.org/10.1007/978-3-662-58172-8.

[136] Sreedevi Sarsan u. a. „Synergistic Interactions Among Microbial Communities“. In: *Microbes in Microbial Communities: Ecological and Applied Perspectives* (2021), S. 1–37.

[137] Robert F. Schwabe und Christian Jobin. „The microbiome and cancer“. In: *Nature Reviews Cancer* 13.11 (2013), S. 800–812. DOI: 10.1038/nrc3610.

[138] Caitlin A. Selway, Jaya Sudarpa und Laura S. Weyrich. „Moving beyond the gut microbiome: Combining systems biology and multi-site microbiome analyses to combat non-communicable diseases“. In: *Medicine in Microecology* 12 (2022), S. 100052. DOI: https://doi.org/10.1016/j.medmic.2022.100052.

[139] Ron Sender, Shai Fuchs und Ron Milo. „Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans“. In: *Cell* 164.3 (2016), S. 337–340. DOI: https://doi.org/10.1016/j.cell.2016.01.013.

[140] Shai S. Shen-Orr u. a. „Network motifs in the transcriptional regulation network of Escherichia coli“. In: *Nature Genetics* 31.1 (2002), S. 64–68. DOI: 10.1038/ng881. URL: https://doi.org/10.1038/ng881.

[141] Jay Shendure und Hanlee Ji. „Next-generation DNA sequencing“. In: *Nature Biotechnology* 26.10 (2008), S. 1135–1145. DOI: 10.1038/nbt1486.

[142] Jay Shendure u. a. „DNA sequencing at 40: past, present and future“. In: *Nature* 550.7676 (2017), S. 345–353. DOI: 10.1038/nature24286.

[143] Caibin Sheng. „Cellular heterogeneity in the DNA damage response is determined by cell cycle specific p21 degradation“. Diss. Humboldt-Universität zu Berlin, Lebenswissenschaftliche Fakultät, 2018. DOI: http://dx.doi.org/10.18452/18730.

[144] Caibin Sheng u. a. „PCNA-Mediated Degradation of p21 Coordinates the DNA Damage Response and Cell Cycle Regulation in Individual Cells“. In: *Cell Reports* 27.1 (2019), 48–58.e7. DOI: https://doi.org/10.1016/j.celrep.2019.03.031.

[145] Feiyu Shi u. a. „Altered gut microbiome composition by appendectomy contributes to colorectal cancer“. In: *Oncogene* 42.7 (2023), S. 530–540. DOI: 10.1038/s41388-022-02569-3.

[146] Natalia G. Starostina und Edward T. Kipreos. „Multiple degradation pathways regulate versatile CIP/KIP CDK inhibitors“. In: *Trends in Cell Biology* 22.1 (2012), S. 33–41. DOI: https://doi.org/10.1016/j.tcb.2011.10.004.

[147] Jacob Stewart-Ornstein und Galit Lahav. „Dynamics of CDKN1A in Single Cells Defined by an Endogenous Fluorescent Tagging Toolkit“. In: *Cell Reports* 14 (2016), S. 1800–1811. DOI: https://doi.org/10.1016/j.celrep.2016.01.045.

[148] Jakob Stokholm u. a. „Maturation of the gut microbiome and risk of asthma in childhood“. In: *Nature Communications* 9.1 (2018), S. 141. DOI: 10.1038/s41467-017-02573-2.

[149] Steven H. Strogatz. *Nonlinear Dynamics and Chaos*. Perseus Books, 1994.

[150] John A. Swets. „Measuring the Accuracy of Diagnostic Systems“. In: *Science* 240.4857 (1988), S. 1285–1293. DOI: 10.1126/science.3287615.

[151] John A. Swets, Robyn M. Dawes und John Monahan. „Better Decisions through Science“. In: *Scientific American* 283.4 (2000), S. 82–87. URL: http://www.jstor.org/stable/26058901 (besucht am 20. 10. 2022).

[152] Agnes Szejka. „Evolution Boolescher Netzwerke“. Diss. Technische Universität Darmstadt, Feb. 2010.

[153] The Human Microbiome Project Consortium. „A framework for human microbiome research“. In: *Nature* 486 (7402 2012), S. 215–221. DOI: 10.1038/nature11209.

[154] The Integrative HMP (iHMP) Research Network Consortium. „The Integrative Human Microbiome Project“. In: *Nature* 569.7758 (2019), S. 641–648. DOI: 10.1038/s41586-019-1238-8.

[155] Jared E Toettcher u. a. „Distinct mechanisms act in concert to mediate cell cycle arrest“. In: *Proceedings of the National Academy of Sciences* 106.3 (2009), S. 785–790.

[156] Takumi Toya u. a. „Coronary artery disease is associated with an altered gut microbiome composition“. In: *PLOS ONE* 15.1 (2020), S. 1–13. DOI: 10.1371/journal.pone.0227147.

[157] Marius Trøseid u. a. „The gut microbiome in coronary artery disease and heart failure: Current knowledge and future directions“. In: *EBioMedicine* 52 (2020), S. 102649. DOI: https://doi.org/10.1016/j.ebiom.2020.102649.

[158] Rajith Vidanaarachchi u. a. „IMPARO: inferring microbial interactions through parameter optimisation“. In: *BMC Molecular and Cell Biology* 21.1 (2020), S. 1–11.

[159] Bert Vogelstein, David Lane und Arnold J. Levine. „Surfing the p53 network“. In: *Nature* 408.6810 (2000), S. 307–310. DOI: 10.1038/35042675.

[160] WHO. *Fact sheet on noncommunicable diseases*. 16. Sep. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases (besucht am 11. 05. 2023).

[161] Yandong Xiao u. a. „Mapping the ecological networks of microbial communities“. In: *Nature communications* 8.1 (2017), S. 1–12.

[162] Wenyu Zhou u. a. „Longitudinal multi-omics of host-microbe dynamics in prediabetes“. In: *Nature* 569.7758 (2019), S. 663–671. DOI: 10.1038/s41586-019-1236-x.

[163] Kelly H. Zou, A. James O'Malley und Laura Mauri. „Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models“. In: *Circulation* 115.5 (2007), S. 654–657. DOI: 10.1161/CIRCULATIONAHA.105.594929.

Wissenschaftlicher Werdegang

Lebenslauf

Seit 06/2016	Interdisziplinäre Promotion in Physik und Biologie, AG Drossel, TU Darmstadt
04/2014 - 05/2016	Master of Science, Physik, TU Darmstadt Master-Thesis: *Evolution einer Population Boolescher Netzwerke*
10/2010 - 04/2014	Bachelor of Science, Physik, TU Darmstadt Bachelor-Thesis: *Untersuchung der Kontaminationsschicht auf Metalloberflächen mittels Femtosekunden-Laserdesorption*
09/2001 - 06/2010	Lessing-Gymnasium Lampertheim, Abschluss Abitur

Publikationen

- Caibin Sheng, Isabella-Hilda Mendler, Sara Rieke, Petra Snyder, Marcel Jentsch, Dhana Friedrich, Barbara Drossel und Alexander Loewer. „PCNA-Mediated Degradation of p21 Coordinates the DNA Damage Response and Cell Cycle Regulation in Individual Cells“. In: Cell Reports (2019)

- Isabella-Hilda Mendler, Barbara Drossel und Marc-Thorsten Hütt. „Microbiome abundance patterns as attractors and the implications for the inference of microbial interaction networks“. arXiv:2306.02100. Eingereicht bei Physica A.

In Vorbereitung

- Laura Friedel, Raphael Löffler, Isabella-Hilda Mendler, Theresa Ingenhaag, Barbara Drossel und Alexander Löwer. „Differential activation by upstream kinases enables p53 to distinguish acute and sustained DNA damage“.

Konferenzbeiträge

03/2018	„Modelling the cell-cycle dependent regulation of p21 after DNA damage“, DPG Frühjahrstagung, Berlin (Poster)
03/2017	„Modelling the regulation of p21 by p53 after DNA damage“, DPG Frühjahrstagung, Dresden (Poster)
03/2016	„Evolution of a population of Boolean threshold networks for a targeted expression pattern“, DPG Frühjahrstagung, Regensburg (Poster)

Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei allen Menschen bedanken, die zum Gelingen dieser Arbeit beigetragen haben.

Allen voran möchte ich mich bei meiner Doktormutter Barbara Drossel bedanken, die mir diese interdisziplinäre Arbeit erst ermöglicht hat. Vielen Dank, dass ich schon seit meiner Masterarbeit Teil deiner unglaublich tollen Arbeitsgruppe sein durfte und du dich nie darüber beschwert hast, dass die Arbeit nach Merles Geburt nur langsam voranging.

Ein ganz großes Dankeschön geht auch an Marc-Thorsten Hütt, der das Zweitgutachten für diese Arbeit übernommen hat. Vielen Dank für die zahlreichen Anregungen und die konkreten Vorschläge zur Durchführung der numerischen Experimente. Unsere gemeinsamen Skype- bzw. Zoom-Gespräche und insbesondere Ihre Begeisterung für die ESABO-Methode haben mich immer sehr motiviert.

Ganz herzlich möchte ich mich auch bei Alexander Löwer, Caibin Sheng und Laura Friedel bedanken, durch die ich viel über das p53-Netzwerk lernen konnte und die stets ein offenes Ohr für alle biologischen Fragen hatten. Die Zusammenarbeit mit euch und unsere gemeinsamen GRK-Retreats haben mir viel Freude bereitet.

Ganz herzlich danke ich auch allen Mitgliedern der AG Drossel, die ich im Laufe der Jahre kennenlernen durfte. Insbesondere die Zeit vor der Corona-Pandemie mit den gemeinsamen Fahrten zur DPG-Frühjahrstagung, langen Wanderungen, Spieleabenden, dem Scheunenfest, legendären Doktorfeiern oder unserer Hochzeitsfeier wird mir für immer in guter Erinnerung bleiben.

Ein großes Dankeschön geht natürlich auch an meine Familie:
Meinen Eltern und Schwiegereltern, danke ich dafür, dass sie mich auf diesem Weg stets unterstützt und insbesondere im Hinblick auf die Kinderbetreuung entlastet haben.
Meinem Mann Marc, der nun schon seit unserer gemeinsamen Zeit im Physik-Leistungskurs Teil meines Lebens ist und mich seit jeher bei allem begleitet, danke ich für seinen festen Glauben daran, dass diese Promotion früher oder später gelingen wird. Vielen Dank für deine fachliche und mentale Unterstützung in all den Jahren, ohne die ich es wohl nie geschafft hätte, diese Arbeit fertigzustellen. Ich danke dir außerdem ganz besonders für das Korrekturlesen der Arbeit, da ich genau weiß, wie wenig Lust du dazu hattest.
Abschließend möchte ich mich noch bei meiner Tochter Merle bedanken, die in letzter Zeit viel Geduld mit mir haben musste, da ich öfter mal an der Arbeit saß, anstatt mit ihr zu spielen. Danke, dass du mein Leben durch deine Neugier und dein riesiges Interesse an allem, was uns umgibt, so unglaublich bereicherst!