Simulation der Histonmodifikation

Max Hild

Abgegeben bei: Dr. Jörg Galle, Prof. Dr. Markus Scholz

Universität Leipzig, Institut für medizinische Informatik, Statistik und Epidemiologie
Neues Augusteum, Augustuspl. 10, 04109 Leipzig - Germany

Abstract. GWAS (Genome-Wide Association Studies) sind eine weit verbreitete Methode zur Identifizierung von genetischen Varianten, die mit komplexen Erkrankungen assoziiert sein könnten. Einer der treibenden Faktoren der Weiterentwicklung der Forschung zum menschlichen Genom war in den letzten Jahren jedoch eine Abkehr von der klassischen Sichtweise, dass die gesamte Heritabilität von Phänotypen lediglich durch die genetische Kodierung erklärt werden kann. In GWAS erreichten die relevanten Maßzahlen für die Heritabilität wiederholt lediglich kleine Werte. [1]

Hier zeigte sich, dass weitere Forschungszweige wie die Transkriptionsanalyse sowie die Epigenetik notwendig sind, um die fehlende Heritabilität der Phänotypen zu erklären. Epigenetische Mechanismen wie DNA-Methylierung und Histonmodifikationen spielen eine entscheidende Rolle bei der Regulation der Genexpression und könnten somit einen Teil der fehlenden Heritabilität erklären. Diese Mechanismen sind dynamisch und können durch Umweltfaktoren beeinflusst werden, was sie zu einem spannenden Forschungsfeld macht.

1 Einleitung

Die Heritabilität von Phänotypen besser zu erforschen lässt Forscherinnen und Forscher leichter nachvollziehen, wie die genetische Information kodiert ist. Prohaska et al. argumentieren, dass die Bedeutung des regulatorischen Systems der Epigenetik für die Vererbung groß ist. Sie skizzieren ein komputationales Modell, dass Reader und Writer Proteine verwendet, die durch eine zugrundeliegende Regulationsfunktion die DNA Replikation sowie Transkription steuern. Dieses System wird in Prohaska et al. mit einem deterministischen System beschrieben, das die Dynamik der epigenetischen Zustände an CpG-Stellen in einem Genom simuliert. Das regulatorische System der Epigenetik wird als ein Schlüssel zur erkennung der fehlenden Heritabilität gesehen. Diese anerkannte Theorie wird durch Evidenz unterstützt. Jedoch ist es noch ein Forschungsgegenstand, wie genau die komplexeren Systeme, die bei Eukaryoten zu finden sind, gesteuert werden. [2] Auch aus der Sicht von Bannister & Kouzarides spielen Histonmodifikationen eine zentrale Rolle im epigenetischen Code. Sie beschreiben die verschiedenen Writer, Eraser und Reader von Histonmarken und deren Einfluss auf Transkription, DNA-Reparatur und Chromosomenkondensation [3]. Transgenerationale Epigenetikstudien zeigen zudem, dass bestimmte epigenetische Markierungen selbst über mehrere Generationen hinweg erhalten bleiben können und so zur fehlenden Heritabilität beitragen [4]. Diese Forschung deckt sich mit

der Einschätzung von McClellan et al. die in ihrer Arbeit ein Plädoyer für die Forschung nach der fehlenden Heritabilität halten.

> Eine der größen Hoffnungen an die GWAS war, dass man - genauso wie eine Vielzahl von mendelschen Erkrankungen auf DNA-Ebene eingegrenzt und das beteiligte Gen samt den Mutationen identifiziert werden konnte - einfach von Einzelgen-Erkrankungen auf komplexe multigenetische Erkrankungen schließen könnte. Das ist jedoch nicht eingetreten. Befürworter werden argumentieren, dass es funktioniert hat und dass allerlei faszinierende Gene identifiziert wurden, die beispielsweise eine Prädisposition zu oder einen Schutz vor Diabetes oder Brustkrebs verleihen, aber die Tatsache bleibt, dass der Großteil der Erblichkeit in diesen Erkrankungen nicht den durch GWAS identifizierten Loci zugeschrieben werden kann, was eindeutig zeigt, dass dies nicht die universelle Lösung sein wird. [1]

Zu verstehen, wie die Modifikation der Histone und DNA-Methylierung funktioniert, ist nach heutiger Sicht ein wesentlicher Schritt, um die Lücke in unserem Verständnis der Heritabilität zu schließen.

2 Methoden

Zentraler Bestandteil dieser Arbeit ist der Vergleich eines stochastischen Modells, welches auf Basis von Markov Ketten mittels Übergangswahrscheinlichkeiten die Histonmodifikation simuliert, mit einem Modell welches auf der Forschung von Prohaska et al. basiert. Das Modell von Prohaska et al. beschreibt die Histonmodifikation als ein deterministisches System. Es definiert jeden Nukleosom-Knoten durch zwei Flags für Acetylation (H3K27ac) und Methylation (H3K9me3) sowie einen Aktivitätsindikator. In jedem Zeitschritt werden alle Nukleosomen gleichzeitig nach folgenden Regeln aktualisiert: Ein Nukleosom erwirbt Acetylation, falls sein linker Nachbar bereits acetylierte Kennzeichnungen besitzt, andernfalls erhält es Methylation, falls sein rechter Nachbar methyliert ist. Diese strikt logischen Booleschen Operationen werden in der Methode `applyDeterministicProhaskaRules()` umgesetzt. Diese Zustände entscheiden darüber, ob die DNA an dieser Stelle für die Transkription verwendet werden kann. Methylierte Stellen sind gebunden, während CpG-Stellen sind spezifische DNA-Sequenzen, die eine hohe Dichte an Cytosin- und Guanin-Basenpaaren aufweisen. Diese Stellen können durch ihre chemische Zusammensetzung

2.1 Markov-Modell (stochastischer Ansatz)

Zur Modellierung der DNA-Histonmodifikation wurde eine C++ Implementierung entwickelt, die auf dem Modell von Prohaska et al. basiert. Das Modell simuliert die dynamischen Veränderungen der epigenetischen Zustände an CpG-Stellen über die Zeit. Jede CpG-Stelle kann einen von fünf möglichen Zuständen annehmen: Das stochastische Modell basiert auf einem Markov-Prozess, der Übergänge zwischen fünf möglichen Zuständen an jeder CpG-Stelle beschreibt: Unmethyliert

(U), Methyliert (M), Histon-modifiziert (H), Vollständig modifiziert (F, sowohl methyliert als auch Histon-modifiziert) und Komplex-gebunden (C). Diese Übergänge werden mit Wahrscheinlichkeiten gesteuert, die auf empirischen Daten beruhen (z.B. Fu et al., 2010). Jeder Zeitschritt besteht darin, für jede CpG-Stelle den nächsten Zustand basierend auf den definierten Übergangswahrscheinlichkeiten zufällig zu wählen. Zusätzlich zur zufälligen Auswahl wird im Modell eine deterministische Komponente integriert, indem die Prohaska-Regeln vor jedem stochastischen Schritt angewendet werden. Dadurch entsteht ein hybrider Ansatz, der sowohl deterministische Nachbarschaftsregeln als auch stochastische Veränderungen abbildet und somit biologisch realistischer sein könnte.

Die Übergänge zwischen diesen Zuständen werden durch ein stochastisches Modell gesteuert, das die biologischen Prozesse der Modifikation und Demodifikation nachbildet. Die Implementierung verwendet einen Markov-Prozess, bei dem die übergangswahrscheinlichkeiten von den aktuellen Zuständen benachbarter CpG-Stellen abhängen.

2.2 Visualisierung

Die Visualisierung besteht aus zwei Teilen: Die Heatmap wurde verwendet, um den Zustand jeder einzelnen CpG-Stelle über die Zeit darzustellen. Das Liniendiagramm zeigt die Häufigkeit jedes Zustands über die Zeit. Die Simulation wird in Zeitschritten von 1 Zeiteinheit durchgeführt, wobei die Übergangswahrscheinlichkeiten für jeden Zustand und jede Nachbarschaftskonfiguration in einer Matrix gespeichert sind. Die Simulation wird für eine bestimmte Anzahl von Iterationen durchgeführt, um die zeitliche Entwicklung der CpG-Stellen zu beobachten.

Technische Details: Die `CpGSite`-Klasse speichert den aktuellen Zustand als Enum und stellt Methoden zum Setzen und Abfragen des Zustands bereit. Die `Genome`-Klasse verwaltet eine Liste von `CpGSite`-Objekten und ermöglicht den Zugriff auf benachbarte Stellen, was für die konditionalen Übergänge notwendig ist. In der `Simulator`-Klasse wird die Simulationslogik implementiert. Sie verwendet einen Zufallszahlengenerator, um stochastische Übergänge gemäßden spezifizierten Wahrscheinlichkeiten durchzuführen. Die Übergangswahrscheinlichkeiten sind in einer Matrix abgelegt, die für jeden Zustand und jede Nachbarschaftskonfiguration die jeweilige Wahrscheinlichkeit enthält. Die Ergebnisse werden nach jedem Zeitschritt in eine CSV-Datei geschrieben. Diese Datei enthält sowohl den Zustand jeder CpG-Stelle zu jedem Zeitpunkt als auch aggregierte Statistiken (z.B. Häufigkeit der Zustände). Für die Visualisierung wird ein separates Python-Skript verwendet, das die CSV-Datei einliest und sowohl eine Heatmap als auch eine Zeitreihe der Zustandsverteilungen erzeugt.

3 Ergebnisse

Die Simulation ermöglicht es, die zeitliche Entwicklung der epigenetischen Zustände über mehrere Generationen zu beobachten. Abbildung 1 zeigt die Visualisierung der ersten Version einer Simulation.

4 Validierung des Modells mit realistischen Daten

Im nächsten Schritt wurde eine Visualisierung mit Übergangswahrscheinlichkeiten aus reellen Daten durchgeführt. Die für diese Simulation gewählten Übergangswahrscheinlichkeiten basieren auf quantitativen Schätzungen aus der Literatur. Hier wurden die Ergebnisse von Fu et al. (2010) genutzt, die mittels eines Bayes'schen Modells an humanen FMR1-Daten folgende Raten ermittelt haben [5]:

- **Fehlerrate der Methylierungs-Erhaltung (Maintenance failure)**: 0,024 pro Zellteilung
- **De-noo-Methylierungsrate (Elternstrang)**: Median 0,08 (80 % CI: 0,04-0,13)
- **De-novo-Methylierungsrate (Tochterstrang)**: Median 0,07 (80 % CI: 0,04-0,11)

Hier zeigte sich, dass bei einer Anwendung der Werte aus der Fu et al. viele Iterationen notwendig sind, um Effekte zu sehen. Aus diesem Grund wurde die Länge der Simulation auf 1000 und dann 100000 Iterationen gesetzt. Es wurde für die Wahrscheinlichkeit einer Methylierung der Wert des Tochterstrangs, also 0,07, verwendet. Die Wahrscheinlichkeit der De-novo-Methylierung wurde auf 0,1 gesetzt, um die Simulation zu beschleunigen. Die Ergebnisse der Simulation mit den Übergangswahrscheinlichkeiten nach Fu et al. (2010) sind in Abbildung 2 und 3 dargestellt. Hier ist zu beachten, dass diese Ergebnisse zur besseren Visualisierung einige Übergangswahrscheinlichkeiten sowie einige Zustände gar nicht abbilden.

Die Ergebnisse zeigen deutliche Muster in der räumlichen und zeitlichen Verteilung der epigenetischen Zustände. Insbesondere kann man beobachten, wie sich die Methylierungszustände in bestimmten Regionen zusammenhäufen und wie sich diese Cluster über die Zeit entwickeln. Als Erweiterung wurde nun das stochastische Modell mit dem regulatorischen System nach Prohaska et al. erweitert. Die Simulationsergebnisse zeigen deutliche Unterschiede zwischen den beiden Modellierungsansätzen. Im deterministischen Modell nach Prohaska bilden sich schnell stabile Muster heraus. Insbesondere die Clusterstatistik zeigte bei unmethylierten Stellen (Zustand U) eine auffällige Häufigkeit von etwa 0,5 bei Clustergrößen von 2 und 6. Im Gegensatz dazu präsentierte das stochastische Modell ein dynamischeres und weniger stabiles Verhalten. Die Clusteranalyse ergab hierbei bei unmethylierten Stellen folgende Verteilung der Häufigkeiten: etwa 0,67 für Clustergröße 1, 0,23 für Clustergröße 2, 0,06 für Größe 3 und etwa 0,01 für Größe 4. Das erklärt ebenfalls das unterschiedliche Verhalten bei der Autokorrelation. Hier weist das deterministische Modell logischerweise 0 auf, da dieses Modell schnell stabile Zustände erreicht. Das Stochastische Modell erreicht je nach Konfiguration unterschiedlich schnell stabile Zustände.

Die beigefügten Abbildungen illustrieren den Verlauf der CpG-Zustände im deterministischen (Abb. 4) und im stochastischen Modell (Abb. 5).

Eine weitere Statistik, die zur Beurteilung der Modelle verwendet wurde, ist die Autokorrelation. Hierbei wurde die Autokorrelation der CpG-Stellen

in den verschiedenen Zuständen über die Zeit analysiert. Die Autokorrelation misst die Korrelation eines Signals mit sich selbst zu verschiedenen Zeitpunkten. Ein positiver Wert deutet darauf hin, dass ähnliche Zustände in der Zeitreihe auftreten, während ein negativer Wert auf eine Abnahme der Ähnlichkeit hinweist. Sie wurde für den Zustand U berechnet, da dieser Zustand "Unmethyliert" repräsentiert und somit eine interessante Basislinie für die Analyse der Zustände, in denen die Transkription möglich ist, innerhalb des Modells darstellt.

Im letzten Schritt wurden beide Modelle kombiniert, um die Vorteile beider Ansätze zu nutzen. Hierbei wurde in jedem durchlauf das stochastische Modell mit den Übergangswahrscheinlichkeiten aus Fu et al. (2010) verwendet, bevor die Regeln nach Prohaska et al. angewendet wurden. Die Ergebnisse zeigen, dass die Kombination der beiden Ansätze zu einer stabileren und biologisch realistischeren Simulation führt. Die Clusterbildung ist weniger ausgeprägt als im deterministischen Modell, aber stabiler als im stochastischen Modell. Dies deutet darauf hin, dass die Kombination der beiden Ansätze eine vielversprechende Basis für zukünftige Simulationen darstellt.

5 Diskussion

Im Vergleich mit der deterministischen Implementierung von Prohaska et al. zeigt das stochastische Modell eine größere Variabilität in der zeitlichen Entwicklung der CpG-Zustände. Dies könnte darauf hindeuten, dass das stochastische Modell besser geeignet ist, um die biologischen Prozesse der Histonmodifikation und deren Einfluss auf die Genexpression zu erfassen. Die Ergebnisse der Simulationen zeigen, dass das stochastische Modell in der Lage ist, komplexe Muster zu erzeugen, die in der Natur beobachtet werden können. Das deterministische Modell zeigt eine größere Stabilität, die sich in der Bildung stabiler Cluster widerspiegelt. Diese Stabilität könnte in biologischen Systemen von Bedeutung sein, da hier das Ziel eine robuste Regulation der Genexpression wäre. Die Kombination des stochastischen und des deterministischen Modells weist als Erweiterung den realistischsten Verlauf auf, und nutzt die Vorteile beider Ansätze. Das kann als Äquivalent der biologisch gewollten Anpassung des regulatorischen Systems und der Einflüsse durch die Umwelt betrachtet werden.

5.1 Weitere Nutzung

Im Modul main.cpp kann das Modell vielfältig parametrisiert werden. Neben der Auswahl der verschiedenen Methoden (Markov sowie Prohaska-Regelbasiert) können auch aus der Forschung geschätzte Übergangswahrscheinlichkeiten sowie die Anzahl der Iterationen und die Anzahl der CpG-Stellen eingestellt werden. Das ermöglicht eine vielfältige Nutzung des Modells um verschiedene Szenarien zu simulieren und deren Auswirkungen auf die CpG-Zustände zu analysieren. Durch die Übergangswahrscheinlichkeiten könnten verschiedene Risikofaktoren, die das regulatorische System beeinflussen könnten, mit der regelbasierten Simulation kombiniert werden.

5.2 Limitationen und Ausblick

Obwohl das Modell viele Aspekte der epigenetischen Regulation abbildet, gibt es Einschränkungen:

- Die Übergangswahrscheinlichkeiten sind aktuell statisch und basieren auf Literaturwerten oder Annahmen. Eine Kalibrierung mit experimentellen Daten sowie ein Training dieser könnte die Aussagekraft weiter erhöhen indem das epigenetische neuronale Netz erweitert wird.
- Die Validierung erfolgte bisher nur qualitativ. Eine quantitative Validierung gegen experimentelle Zeitreihen ist ein nächster Schritt.
- Weitere epigenetische Mechanismen (z.B. DNA-Acetylierung, Interaktion mit Transkriptionsfaktoren) könnten integriert werden, um das Modell zu erweitern.
- Die Annahmen von Prohaska et al. müssten aktualisert und erweitert werden, um die Dynamik der Histonmodifikation bestmöglich abzubilden.

6 Fazit

Die C++ Implementierung der DNA-Histonmodifikation ermöglicht ein tieferes Verständnis der dynamischen Prozesse, die der epigenetischen Regulation zugrunde liegen. Durch die Kombination von effizienter Simulation und anschaulicher Visualisierung können komplexe epigenetische Muster simuliert und analysiert werden. Diese Art der Modellierung kann dazu beitragen, die fehlende Heritabilität besser zu verstehen, indem sie aufzeigt, wie epigenetische Mechanismen zur Vererbung phänotypischer Merkmale beitragen können. Zukünftige Erweiterungen der Implementierung könnten eine parallele Simulation von Eltern und Kind Strängen beinhalten, um die transgenerationale Vererbung von epigenetischen Markierungen zu untersuchen. Zudem könnte ein Messfehlerparameter eingeführt werden. Die Nutzung von weiteren experimentellen Daten kann zur Validierung des Modells sowie zur Simulation spezifischer genetischer Erkrankungen genutzt werden, um potenzielle epigenetische Therapieansätze zu identifizieren. Weiterhin können mit dieser Implementierung Erwartungswerte basierend auf vergangenen Beobachtungen errechnet werden. So kann über längere Zeit beobachtet werden, wie sich das epigenetische regulatorische System verhalten würde, wenn bestimmte Annahmen über die Histonmodifikation der Wahrheit entsprechen.

7 Abbildungen

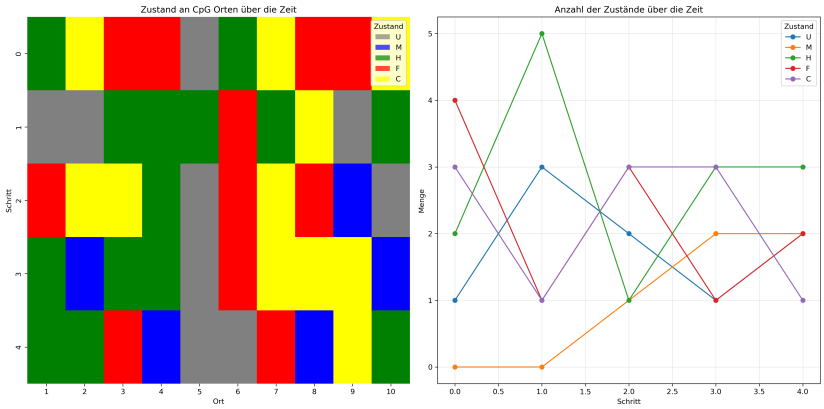

Fig. 1: Visualisierung der CpG-Zustände über die Zeit. Links: Heatmap der Zustände an verschiedenen CpG-Stellen über die Zeit (U: grau, M: blau, H: grün, F: rot, C: gelb). Rechts: Anzahl der CpG-Stellen in jedem Zustand über die Zeit.

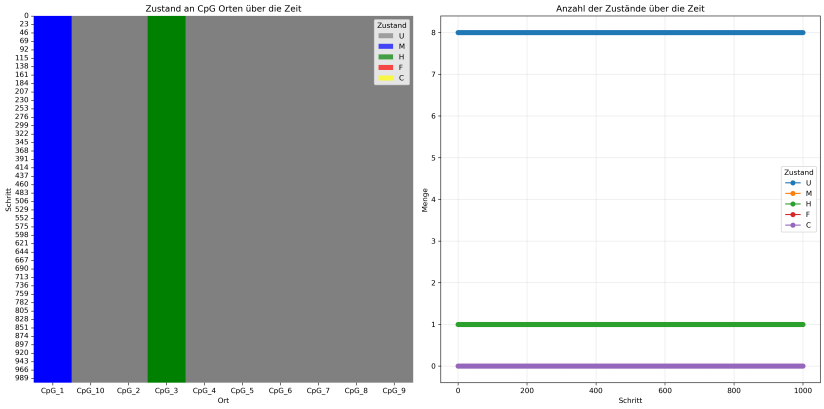

Fig. 2: Visualisierung mit den Wahrscheinlichkeiten aus Fu et al. (2010) und 1000 Iterationen. [5]

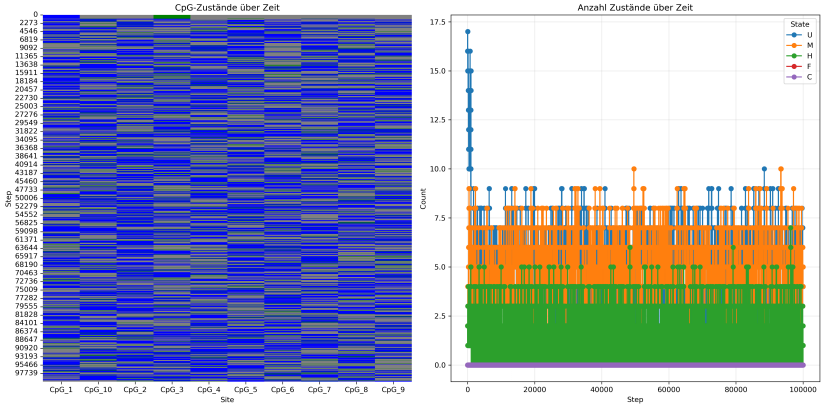

Fig. 3: Visualisierung mit den Wahrscheinlichkeiten aus Fu et al. (2010) und 100000 Iterationen. [5]

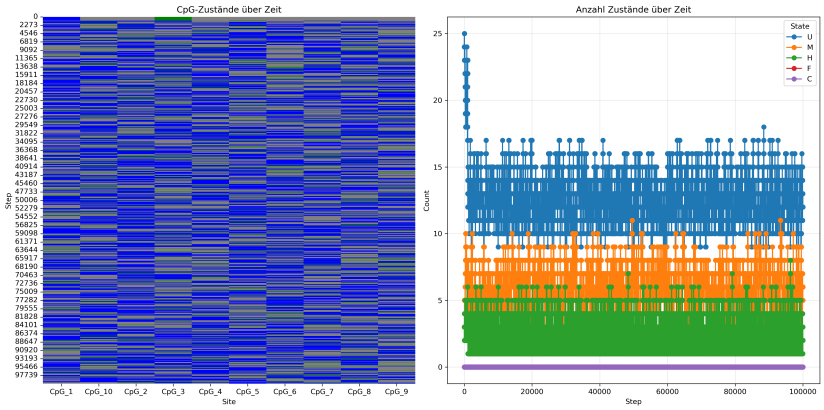

Fig. 4: Visualisierung des deterministischen Modells: links Heatmap der Zustände über die Zeit, rechts Häufigkeiten der Zustände.

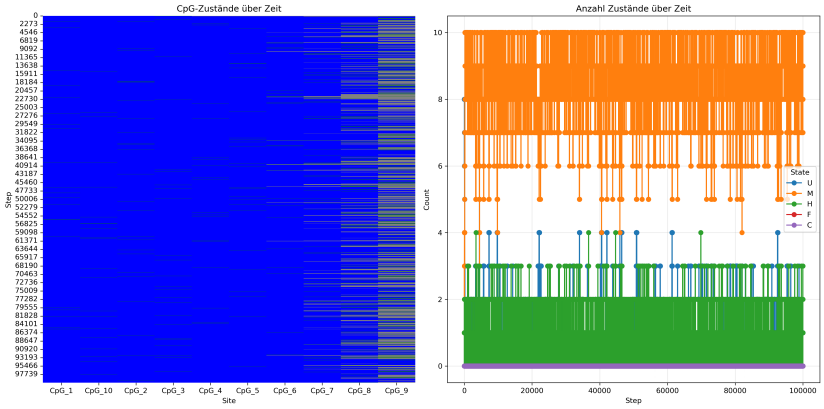

Fig. 5: Visualisierung des stochastischen Modells: links Heatmap der Zustände über die Zeit, rechts Häufigkeiten der Zustände.

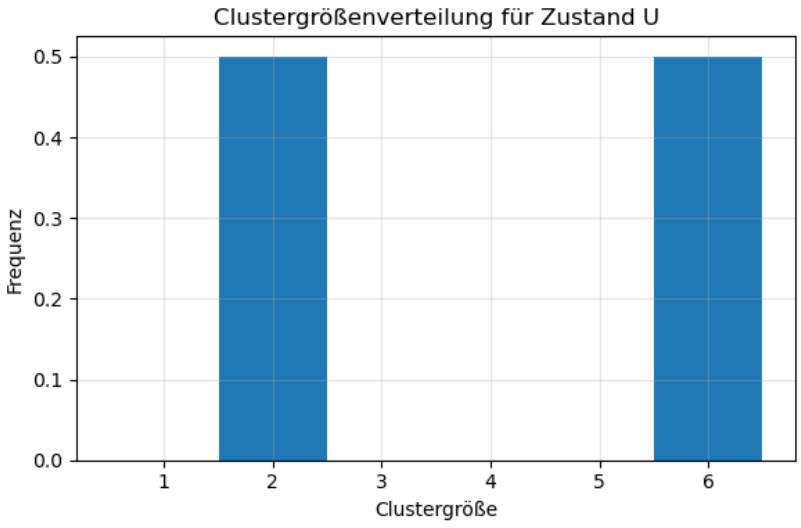

Fig. 6: Clustergröße im deterministischen Modell

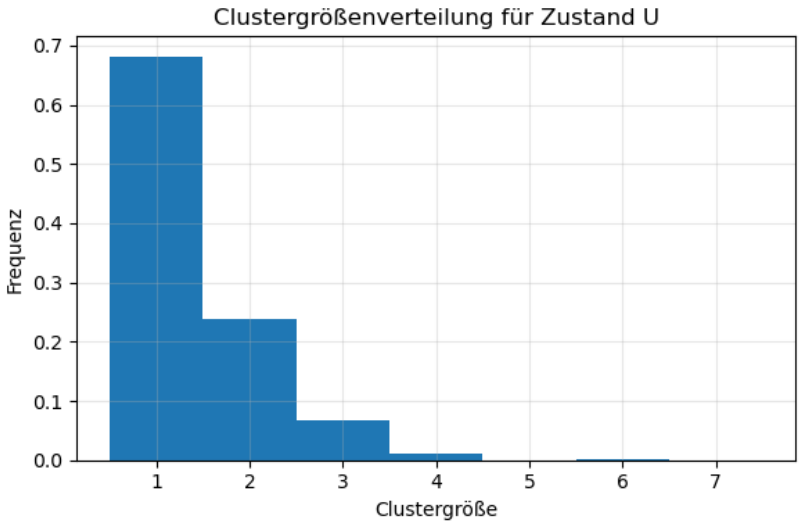

Fig. 7: Clustergröße im stochastischen Modell

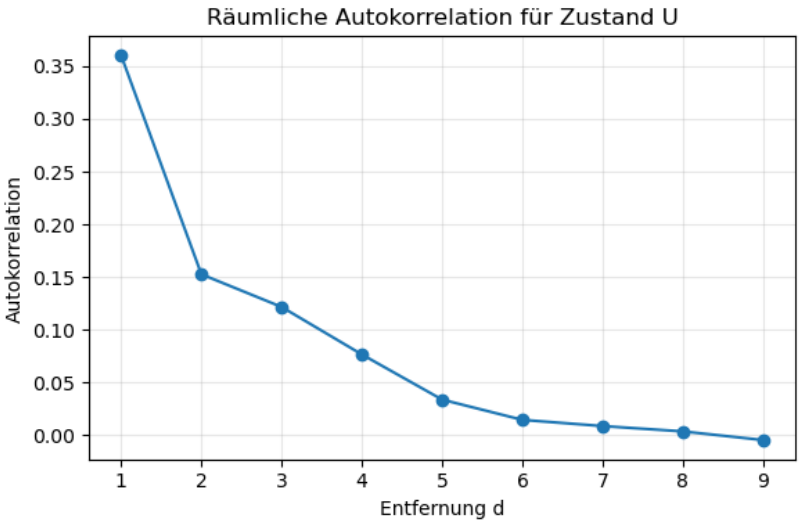

Fig. 8: Autokorrelation im stochastischen Modell

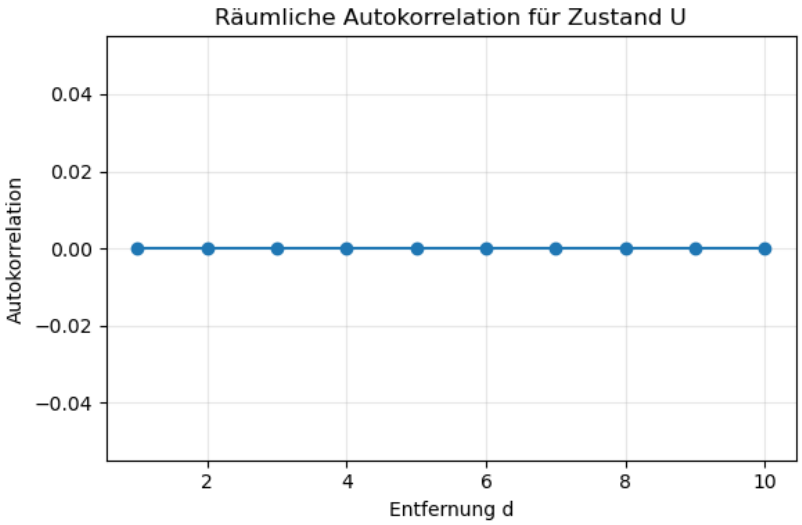

Fig. 9: Autokorrelation im deterministischen Modell

References

[1] Jon McClellan and Mary-Claire King. Genetic heterogeneity in human disease. *Cell*, 141(2):210–217, 4 2010.

[2] Sonja J. Prohaska, Peter F. Stadler, and David C. Krakauer. Innovation in gene regulation: The case of chromatin computation. *Journal of Theoretical Biology*, 265(1):27–44, 3 2010.

[3] Antony J. Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395, 2011.

[4] Trygve Tollefsbol. *Transgenerational epigenetics*. Elsevier, 2014.

[5] Audrey Qiuyan Fu, Diane P. Genereux, Reinhard Stöger, Charles D. Laird, and Matthew Stephens. Statistical inference of transmission fidelity of DNA methylation patterns over somatic cell divisions in mammals. *arXiv preprint arXiv:1011.2025*, 2010.