Simulation der Histonmodifikation

Max Hild

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

Universität Leipzig, Institut für medizinische Informatik, Statistik und Epidemiologie
Härtelstraße 16-18, 04107 Leipzig - Germany

Abstract. GWAS (Genome-Wide Association Studies) erklären nur einen kleinen Teil der Heritabilität von genetischen Varianten, die mit komplexen Erkrankungen assoziiert sein könnten. [1] Ein großer Teil der Heritabilität wird heutzutage epigenetischen Prozessen zugeschrieben.
Dazu gehört in Eukaryoten vor allem die reversible chemische Modifikation der Histonproteine. [2] Diese Mechanismen sind dynamisch und können durch Umweltfaktoren beeinflusst werden, was sie zu einem spannenden Forschungsfeld macht. Die vorliegende Simulation soll die Histonmodifikation basierend auf dem komputationalen Modell von Prohaska et al. simulieren. Dabei wurde der deterministische, regelbasierte Ansatz aus der Publikation mit einer stochastischen Markovsimulation verglichen und letztendlich kombiniert. `Coderepository: https://github.com/maxhild/epigen`

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, das 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 Histonen in einem Genom simuliert. Das regulatorische System der Epigenetik wird als ein Schlüssel zur Erklärung 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 Empfehlung 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]

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. Sie beschreiben Chromatin als hierarchisches Informationssystem, in dem Reader- und Writerproteine chemische Markierungen auf Histonen setzen und auslesen. Diese Markierungen modulieren die Zugänglichkeit der DNA für Transkriptions- und Reparaturmaschinen. Das Modell 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 nach heutiger Ansicht darüber, ob und wie die DNA an dieser Stelle für die Transkription verwendet werden kann. Die Ansicht, dass die Zustände des Chromatins binär seien, wurde widerlegt. Filon et al. erkannten durch Principal Component Analysis fünf verschiedene Zustände von Chromatin. [5] In diesem Modell ist zur Vereinfachung die Aktivität des Chromatins binär kodiert. Auch die Histonmodifikationen sind als binäre Zustände kodiert.

3 Prohaska-Modell, (deterministischer Ansatz)

Das auf Prohaska et al. basierende Modell zeigt einen Ansatz, bei dem durch das Festlegen von Regeln eine Vorhersage über die zukünftige Konstellation der Histonmodifikationen getroffen wird. Dieser regelbasierte Ansatz hat den Vorteil, dass hierdurch Erkenntnisse aus der realen Welt in das Modell einfließen können. Die Schwäche dieses Modells ist es jedoch, dass dadurch oft ein Endzustand erreicht wird, bei dem das Modell sich nicht mehr ändert. Die verwendeten Regeln sind: Acetylations-Writer (Aktivierungs-Regel): Diese Regel modelliert, wie ein bereits acetylierter Nachbar (H3K27ac) ein Nukleosom “ak-

tiviert" und selbst acetylieren lässt. Prohaska et al. sprechen hier von einem positiven Rückkopplungsschritt, der aktives Chromatin ausbreitet. Methylations-Writer (Silencing-Regel): Entspricht der Rekrutierung eines Methyltransferase-Komplexes (z. B. DNMT1), der einen Nukleosom re-methyliert und dadurch "stummschaltet". Dies ist in ihrem Framework der Repressive Feedbackschritt, der in Konfliktsituationen immer Vorrang vor Acetylation hat. Zudem wurde zufallsbasiert ein Eraser in das Modell aufgenommen der mit einer wählbahren Wahrscheinlichkeit einzelne Modifikationen löscht. Da das Modell relativ schnell in einen stabilen Zustand konvergierte, wurde es mit einem Markovmodell erweitert, das zusätzliche zufällige Übergangswahrscheinlichkeiten definiert. (Abbildung: 2)

3.1 Markov-Modell (stochastischer Ansatz)

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 Histonstellen abhängen. Jeder Zeitschritt besteht darin, für jede Histonstelle den nächsten Zustand basierend auf den definierten Übergangswahrscheinlichkeiten zufällig zu wählen. (Abbildung: 3)

3.2 Kombiniertes Modell

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. 4 In der Analyse der Größe der Cluster sowie der Autokorrelation wird deutlich, dass der Unmethylierte Zustand über die Zeit abnimmt, jedoch auch durch zufällige Ereignisse wieder ansteigen kann. Dieser Effekt wird verstärkt wenn das Markov Model höhere Übergangswahrscheinlichkeiten als Startwerte bekommt und verhindert zu stabile Zustände. Das bietet interessante Möglichkeiten für die Analyse der Auswirkung von Risikofaktoren, die über weitere Übergangswahrscheinlichkeiten modelliert werden könnten. (Abbildung: 6, Abbildung: 5)

3.3 Modellierung

Die Simulation wird in Zeitschritten von 1 Zeiteinheit durchgeführt, wobei die Übergangswahrscheinlichkeiten für jeden Zustand und jede Nachbarschaftskonfiguration vor der Simulation basierend auf aktuellem Kenntnisstand festgelegt wird. Die Simulation wird für eine bestimmte Anzahl von Iterationen durchgeführt, um die zeitliche Entwicklung der Histonestellen zu beobachten. Die Histone (Sites, Orte) wurden auf 10 gesetzt, während die chemischen Histonmodifikationen auf 5 gesetzt wurden (States, Zustände). Die folgenden Zustände wurden definiert:

- **U**nmodifiziert (U): Histon ist nicht modifiziert.
- **M**ethyliert (M): Histon ist methyliert.
- **P**hosphoryliert (P): Histon ist phosphoryliert.
- **A**cetylisiert (A): Histon ist acetylisiert.
- **U**biquitiliert (U): Histon ist ubiquitiliert.

Technische Details: Die `EpigeneticSite`-Klasse speichert den aktuellen Zustand (`ModificationState`) als Enum und stellt Methoden zum Setzen und Abfragen des Zustands bereit. Das `Conditional Model` verwaltet eine Liste von `HistoneSite`-Objekten und ermöglicht den Zugriff auf benachbarte Stellen, was für die konditionalen Übergänge notwendig ist. In der `Main`-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 den Zustand jeder Histonstelle zu jedem Zeitpunkt. 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. Darüberhinaus kann die Autokorrelation sowie die Clusterverteilung visualisiert werden.

4 Ergebnisse

Die Simulation ermöglicht es, die zeitliche Entwicklung der epigenetischen Zustände über mehrere Generationen zu beobachten. Abbildung 4 zeigt die Visualisierung der ersten Version einer Simulation. Die Visualisierung besteht aus zwei Teilen: Die Heatmap wurde verwendet, um den Zustand jeder einzelnen Histonstelle über die Zeit darzustellen. Das Liniendiagramm zeigt die Häufigkeit jedes Zustands über die Zeit.

5 Anwendung des stochastischen 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 [6]:

- **Fehlerrate der Methylierungs-Erhaltung (Maintenance failure)**: 0,024 pro Zellteilung

- **De-novo-Methylierungsrate (Elternstrang)**: Median 0,08 (80 % CI: 0,04-0,13)
- **De-novo-Methylierungsrate (Tochterstrang)**: Median 0,07 (80 % CI: 0,04-0,11)

Es wurde für die Wahrscheinlichkeit einer Methylierung der Wert des Tochterstrangs, also 0,07, verwendet. Die Wahrscheinlichkeit der De-novo-Methylierung wurde auf 0,24 gesetzt. Die Ergebnisse der Simulation mit den Übergangswahrscheinlichkeiten nach Fu et al. (2010) sind in Abbildung 1 dargestellt. 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 die Modifikationen mit den gewählten Wahrscheinlichkeiten über die Zeit entwickeln.

6 Diskussion

Im Vergleich mit der deterministischen Implementierung von Prohaska et al. zeigt das stochastische Modell eine größere Variabilität in der zeitlichen Entwicklung der Histonzustände. Dies könnte darauf hindeuten, dass das stochastische Modell besser geeignet ist, um die biologischen Prozesse der Histonmodifikation durch Umwelteinflüsse und deren Einfluss auf die Genexpression zu erfassen. Die Ergebnisse der Simulationen mit den deterministischen Regeln von Prohaska et al. 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. Wie in der Validierung der Daten mit den Forschungsergebnissen aus Fu et al. kann durch die Simulation ein Erwartungswert für die erwatete Häufigkeit von Modifikationen errechnet werden. Über das stochastische Modell können zusätzlich Umweltfaktoren simuliert werden.

6.1 Weitere Nutzung

Im Modul main.cpp kann das Modell vielfältig parametrisiert werden sowie Simulationen mit und ohne die Regeln von Prohaska et al. und dem Markov Modell durchgeführt werden. Die Übergangswahrscheinlichkeiten können nach Forschungsergebnissen gesetzt werden um Umweltfaktoren zu simulieren. 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 Histonstellen eingestellt werden.

Das ermöglicht eine vielfältige Nutzung des Modells um verschiedene Szenarien zu simulieren und deren Auswirkungen auf die Histonzustände zu analysieren. Durch die Übergangswahrscheinlichkeiten könnten verschiedene Risikofaktoren, die das regulatorische System beeinflussen könnten, mit der regelbasierten Simulation kombiniert werden.

6.2 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. Risikofaktoren, Messungen aus Studien Interaktion mit Transkriptionsfaktoren) könnten einfach integriert werden, um das Modell zu erweitern.
- Die Annahmen von Prohaska et al. müssten aktualisert und erweitert werden, um die Dynamik der Histonmodifikation nach heutigem Forschungsstand bestmöglich über neue Regeln und Klassen abzubilden.

7 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.

8 Abbildungen

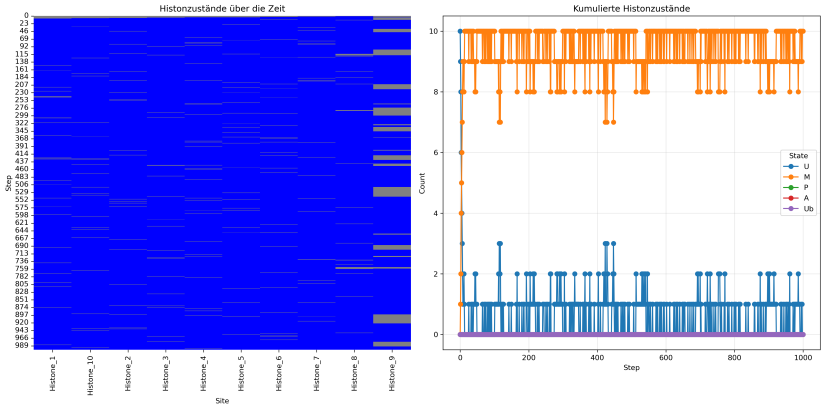

Abbildung: 1: Visualisierung des stochastischen Modells mit den Wahrscheinlichkeiten für de-novo-methylierung im Tochterstrang und Fehlerrate der Methylierungserhaltung aus Fu et al. (2010) und 1000 Iterationen. [6]

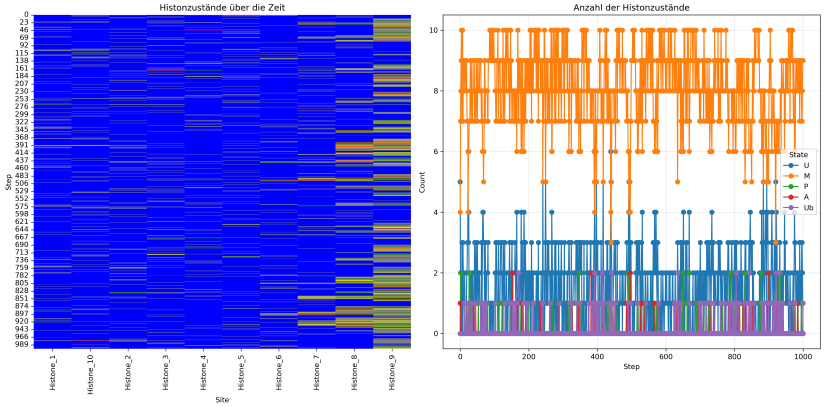

Abbildung: 2: Visualisierung des deterministischen Modells: links Heatmap der Zustände über die Zeit, rechts Häufigkeiten der Zustände.

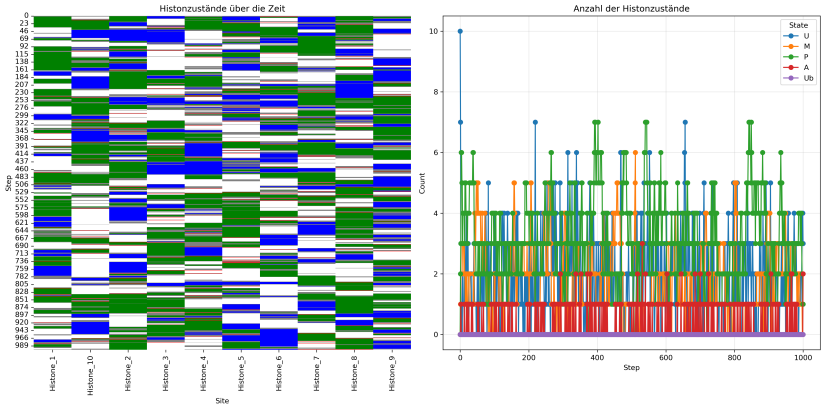

Abbildung: 3: Visualisierung des stochastischen Modells: links Heatmap der Zustände über die Zeit, rechts Häufigkeiten der Zustände.

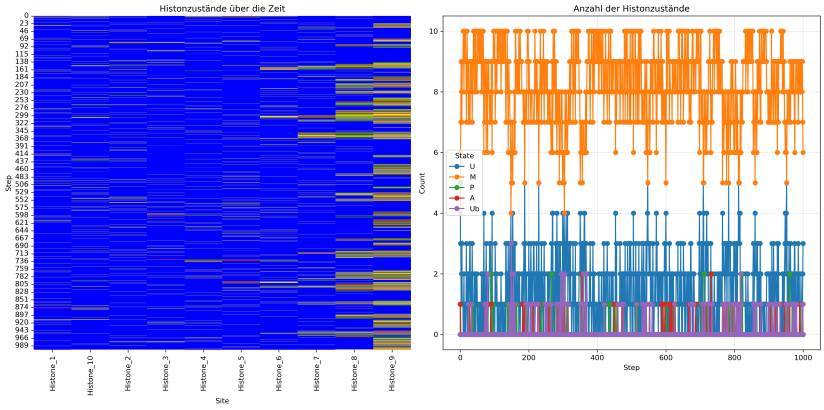

Abbildung: 4: Visualisierung des gemischten Modells: links Heatmap der Zustände über die Zeit, rechts Häufigkeiten der Zustände.

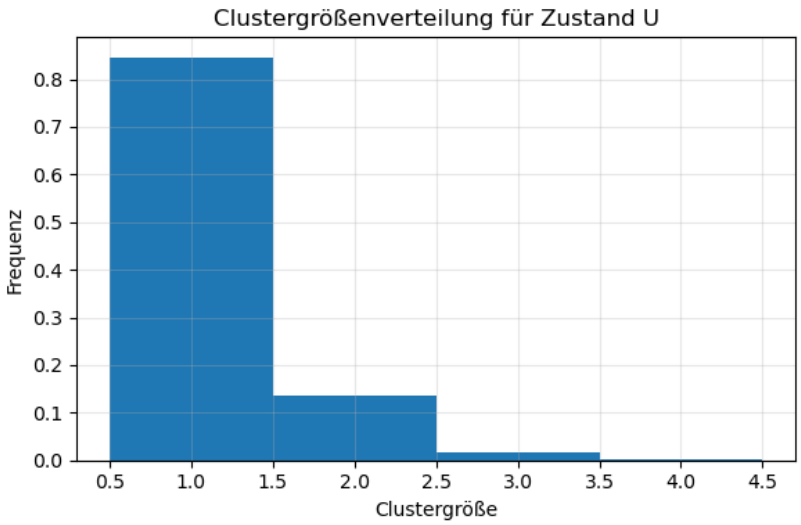

Abbildung: 5: Clustergröße von Unmethylierten Zuständen im Prohaska und Markov Modell

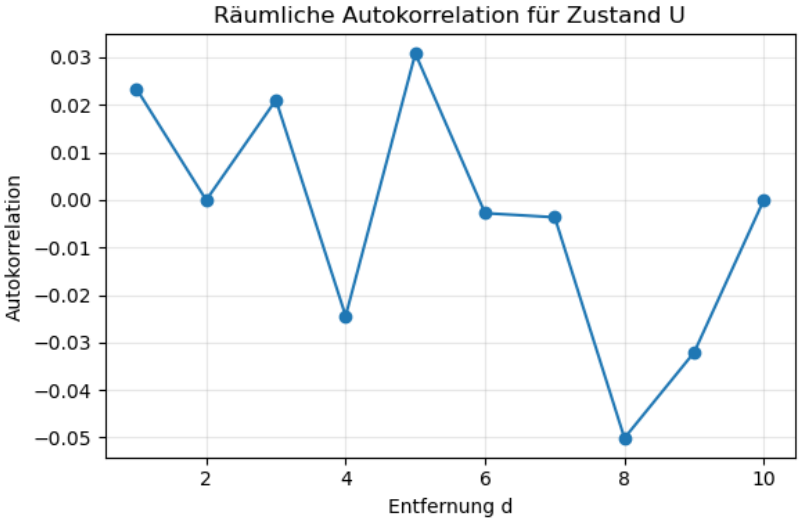

Abbildung: 6: Autokorrelation im gemischten Modell aus Prohaska und Markov

Quellenverzeichnis

[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] Guillaume Filion, Joke G. van Bemmel, Ulrich Braunschweig, Wendy Talhout, Jop Kind, Lucas D. Ward, Wim Brugman, Inês J. de Castro, Ron M. Kerkhoven, Harmen J. Bussemaker, and Bas von Steensel. Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell*, 143, 2010.

[6] 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.