Technische Hochschule Ingolstadt

Technische Hochschule Ingolstadt

Seminar Künstliche Intelligenz

Studiengang Künstliche Intelligenz, M.Sc.
Fakultät Informatik

Vorhersage von Off-Target-Effekten durch KI in der CRISPR-Cas9 Gen-Modifizierung

Vor- und Zuname: **Robert Kessler**

Matrikelnummer: **00153751**

Zusammenfassung

Die CRISPR-Cas9-Technologie hat die biomedizinische Forschung revolutioniert, indem sie gezielte genetische Modifikationen ermöglicht. Trotz ihrer Potenziale stehen der Anwendung von CRISPR/Cas9 in der klinischen Praxis erhebliche Herausforderungen entgegen, insbesondere das Risiko von Off-Target-Effekten, die zu unerwünschten Mutationen führen können. Diese Arbeit untersucht und vergleicht verschiedene Deep-Learning-Modelle zur Vorhersage dieser Off-Target-Effekte, darunter Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) und das hybride Modell CRISPR-Net. Die Ergebnisse zeigen, dass CRISPR-Net durch die Kombination der Stärken von CNNs und RNNs eine überlegene Leistungsfähigkeit bei der Vorhersage von Off-Target-Mutationen bietet. Diese Erkenntnisse tragen dazu bei, die Sicherheit und Effizienz der CRISPR-Cas9-Genmodifikation zu erhöhen und ihre Anwendung in der klinischen Praxis voranzutreiben.

Keywords— CRISPR/Cas9, Off-Target-Effekte, Deep Learning, Künstliche Intelligenz (KI), CRISPR-Net

Inhaltsverzeichnis

Abkürzungsverzeichnis

AUROC Area Under the Receiver Operating Characteristic Curve

CNN Convolutional Neural Network

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

GUIDE-seq Genome-Wide Unbiased Identification of DSBs Enabled by Sequencing

KI Künstliche Intelligenz

RNA Ribonukleinsäure (engl. Ribonucleic acid)

RNN Recurrent Neural Network

Glossar

Endonuklease Gruppe von Enzymen, die bestimmte DNA-Sequenzen spezifisch erkennen und schneiden können

Genom Vollständige genetische Information eines Organismus, codiert in DNA oder Ribonukleinsäure (engl. Ribonucleic acid) (RNA)

Indel-Frequenz Häufigkeit von Insertionen und Deletionen in einer DNA-Sequenz

Nukleotid Bausteine der DNA und RNA, bestehend aus einem Zucker, einer Phosphatgruppe und einer organischen Base. Nukleotide verbinden sich zu langen Ketten und speichern die genetische Information.

Phage Virus, welcher ausschließlich Bakterien infiziert

1 Einleitung

Die Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas9-Technologie, eine Methode zur Genmodifizierung, hat die biomedizinische Forschung revolutioniert. Diese Technologie ermöglicht es, gezielt genetische Veränderungen vorzunehmen, indem sie auf spezifische DNA-Sequenzen abzielt und präzise Schnitte setzt (Redman et al. 2016).

1.1 Problemstellung

Trotz der beeindruckenden Fortschritte stehen der Anwendung von CRISPR-Cas9 in der klinischen Praxis noch erhebliche Herausforderungen entgegen. Eine der größten Hürden ist das Risiko von Off-Target-Effekten, bei denen die Cas9-Endonuklease unbeabsichtigte Stellen im Genom schneidet. Diese Off-Target-Mutationen können zu genetischer Instabilität und potenziell schädlichen Folgen führen (Lin & Wong 2018). Daher ist die genaue Vorhersage dieser Off-Target-Effekte von entscheidender Bedeutung für die Sicherheit und Wirksamkeit von CRISPR-Cas9 in klinischen Anwendungen.

1.2 Zielsetzung

Die Zielsetzung dieser Arbeit besteht darin, verschiedene Deep-Learning-Modelle zur Vorhersage von Off-Target-Effekten zu vergleichen. Insbesondere werden Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) und das hybride Modell CRISPR-Net untersucht. Diese Modelle sollen daraufhin bewertet werden, wie gut sie in der Lage sind, Off-Target-Mutationen vorherzusagen und dadurch die Sicherheit der Genmodifizierung zu erhöhen.

2 Grundlagen

2.1 CRISPR-Cas9 Gen-Modifizierung

CRISPR-Cas9 ist eine Technologie zur Genmodifizierung, welches es ermöglicht, Fehler im Genom zu korrigieren (Redman et al. 2016). Die Studien von Yin et al. (2014) zeigen, dass mithilfe von CRISPR-Cas9 genetisch-bedingte Krankheiten von Mäusen geheilt werden konnten und diese auch auf Menschen anwendbar sei.
Dieser Mechanismus stammt ursprünglich von Bakterien zur Abwehr von Phagen. Hier beziehen sich die CRISPR-Sequenzen auf spezifische Abschnitte der DNA, die sich palindromisch wiederholen. Das Cas9 ist ein Schlüsselprotein, welches als Endonuklease fungiert und in der Lage ist, beide Stränge der DNA zu schneiden. Um das Ziel zu finden wird Cas9 von einer einzelsträngigen (engl. single guide) RNA geleitet. Nach dem DNA-Schnitt wird entweder mithilfe nicht-homologem Endjoining zufällige Einfügungen oder Deletionen durchgeführt, oder anhand von homologiegesteuerter Reparatur homologe DNA-Stücke zur Reparatur verwendet und eingefügt.
In der theoretischen Annahme, dass eine homologe DNA-Sequenz gezielt zusammen mit Cas9 und der sgRNA bereitgestellt wird, ermöglicht CRISPR-Cas9 eine präzise Genmodifikation, die Veränderungen auf der Ebene eines einzelnen Basenpaares erlaubt (Redman et al. 2016).

Allerdings kann es vorkommen, dass obwohl spezifische DNA Sequenzen angesteuert werden, die sgRNA auch andere Regionen beeinflussen kann. Diese ungewollten Mutationen in den Genen werden als „Off-Target Mutationen“ bezeichnet und sind noch eine große Herausforderung in der Anwendung von CRISPR-Cas9 (Lin & Wong 2018).

2.2 Künstliche Intelligenz (KI)

Künstliche Intelligenz (KI), insbesondere Deep Learning, bietet eine Möglichkeit, um die Genauigkeit dieser Vorhersagen zu verbessern. In der aktuellen Forschung werden speziell entwickelte neuronale Netzwerke eingesetzt, um das Muster und die Häufigkeit potenzieller Off-Target-Stellen zu erkennen. Die Arbeit von Lin & Wong (2018) zeigt erstmalig auf, wie eine Kombination aus CNNs und RNNs verwendet werden kann, um die Wahrscheinlichkeit und das Ausmaß der Off-Target Mutationen vorherzusagen.
Diese KI-Modelle nutzen eine Kombination aus genetischer Sequenzinformationen und eine Vielzahl an lernbaren Parametern, um Beziehungen zwischen den verschiedenen Sequenzmerkmalen und den Off-Target Aktivitäten zu modellieren (Lin & Wong 2018).

3 Methodik

In diesem Abschnitt wird das methodische Vorgehen der aktuellen Forschung beschrieben. Die methodischen Schritte umfassen die Erhebung und Aufbereitung der Daten, das Design und Implementierung der Deep Learning Architekturen, sowie die Optimierung der Modelle. Abschließend wird die Validierung und Bewertung der entwickelten Modelle beschrieben.

3.1 Datenerhebung und -vorbereitung

Aktuell wird in der Forschung häufig der CRISPOR-Datensatz verwendet, da er umfassende Informationen zur Vorhersage von Off-Target Effekten bereitstellt. Der CRISPOR-Datensatz umfasst Daten zu mehr als 150 Genomen und ermöglicht die Auswahl von geeigneter Guide-RNA für eine Vielzahl von Organismen. CRISPOR kann Off-Target-Stellen im Genom vorhersagen, gRNA-Sequenzen nach ihrer Effizienz und Sicherheit ranken und problematische Guides hervorheben (Concordet & Haeussler 2018). CRISPOR ermöglicht diese umfassenden Analysen und Bewertungen der gRNA, indem es spezifische Scores zur Vorhersage von On-Target- und Off-Target-Effekten berechnet. Diese Scores basieren auf umfangreiche experimentelle Daten und bieten eine gute Grundlage für die Auswahl der besten gRNAs für spezifische Anwendungen.
Ein wesentlicher Vorteil von CRISPOR ist, dass der komplette Quellcode, sowie eine eigenständige Kommandozeilenversion als Open Source auf GitHub unter Haeussler (2024) verfügbar ist. Dies ermöglicht es Forschern, den Code an ihre spezifischen Projekte anzupassen, zu erweitern und zu integrieren.

Zur Validierung der entwickelten Modelle wird Genome-Wide Unbiased Identification of DSBs Enabled by Sequencing (GUIDE-seq) verwendet. GUIDE-seq bietet eine unvoreingenommene und hochempfindliche Erkennung von Off-Target-Effekten, die für die Bewertung der Spezifität und Sicherheit von CRISPR-Cas9-Experimenten entscheidend ist. Diese Methode wird zur Validierung von Vorhersagemodellen für Off-Target-Effekte eingesetzt, um sicherzustellen, dass die Modelle genaue und zuverlässige Ergebnisse liefern. Dabei können Off-Target-Stellen mit einer Indel-Frequenz von über 0,1% detektiert werden, was GUIDE-seq zu einem leistungsfähigen Werkzeug für die genaue Charakterisierung von Nuklease-Aktivitäten macht (Zhu et al. 2017).
Wie bei CRISPOR ist auch GUIDE-seq als Open Source auf GitHub unter *Aryeelab/Guideseq* (2024) verfügbar und kann von Forschern angepasst und erweitert werden.

3.2 Entwicklung des KI-Modells

In den letzten Jahren wurden bedeutende Fortschritte bei der Anwendung von Deep Learning zur Vorhersage der Off-Target-Effekte von CRISPR-Cas9 erzielt. Dieses Kapitel beschreibt die Entwicklung und Optimierung von Deep Learning-Modellen zur Vorhersage von Off-Target-Effekten, insbesondere unter Verwendung von Convolutional Neural Networks (CNNs) und CRISPR-Net.

3.2.1 Sequence Encoding

Einer der ersten Schritte bei der Entwicklung von Deep Learning-Modellen zur Vorhersage von Off-Target-Effekten ist die richtige Kodierung der sgRNA-DNA-Sequenzpaare. Die Forschung von Charlier et al. (2021) hat das bestehende Sequence-Encoding verbessert, indem sie die diese Paare als eine Matrix für die Eingabe in ein CNN kodiert haben. Sie nutzen eine $8xL$-Matrix, wobei 8 die Anzahl der Nukleotide und L die Länge der Sequenz ist, um die sgRNA-DNA-Sequenzpaare zu kodieren (Charlier et al. 2021).

3.2.2 Convolutional Neural Networks (CNNs)

Das CNN ist eine der am häufigsten verwendeten Deep Learning-Architekturen zur Vorhersage von Off-Target-Effekten. Diese stellt die sgRNA-DNA-Sequenzpaare als Bild dar und verwendet Convolutional- und Pooling-Schichten, um die Merkmale zu extrahieren und zu verarbeiten. Die Architektur wird wie folgt umfasst:

- **Eingabeschicht**: Die kodierte Matrix der Größe $8xL$ wird als Eingabe in das CNN-Modell gegeben.
- **Erste Faltungsschicht**: Diese Schicht verwendet ein Kernel von $3x3$ und eine ReLU-Aktivierungsfunktion.
- **Zweite Faltungsschicht**: Diese Schicht verwendet ein Kernel von $1x1$ und eine ReLU-Aktivierungsfunktion. Die ersten beiden Faltungsschichten extrahieren die Informationen der sgRNA-DNA.
- **Pooling-Schicht**: Diese Schicht verwendet eine Max-Pooling-Operation der Größe $(2, 2)$, um die Dimensionalität der Daten zu reduzieren.
- **Dense-Schicht**: Diese Schicht besteht aus 128 Neuronen, gefolgt von einer Dropout-Schicht zur Vermeidung von Overfitting und einer Softmax-Aktivierungsfunktion zur Vorhersage von Off-Target-Effekten.

3.2.3 Recurrent Neural Networks (RNNs)

Die Forschung von Charlier et al. (2021) stellt ebenfalls eine Form des RNN vor, welches wie das CNN fungiert und wie folgt aussieht:

- **Eingabeschicht**: Die kodierte Matrix der Größe $8xL$ wird als Eingabe in das RNNs-Modell gegeben.
- **LSTM-Schicht**: Diese Schicht besteht aus 92 Neuronen. Sie dient dazu, sequentielle Abhängigkeiten zu erfassen.
- **Dense-Schichten**: Die Ausgabe der LSTM Schicht wird an zwei aufeinanderfolgenden fully connected Schichten mit jeweils 92 Neuronen übergeben. Diese sind per Batch-Normalisierung getrennt, um die Eingaben während des Trainingsprozesses zu normalisieren.
- **Dropout-Schicht**: Eine Dropout-Schicht wird verwendet, um Overfitting zu vermeiden.
- **Ausgabeschicht**: Die letzte Schicht ist eine Ausgabeschicht mit einer Softmax Aktivierungsfunktion. Diese Funktion wandelt die Ausgabe in Wahrscheinlichkeiten um, die die Vorhersage der Off-Target-Effekte darstellen.

3.2.4 CRISPR-Net

Die Modelle des CNN und RNN wurden von Lin et al. (2020) in einem neuen Modell namens CRISPR-Net kombiniert. Es kombiniert die Stärken beider Modelle, um sowohl lokale als auch sequentielle Muster in den gRNA-DNA-Interaktionen zu erfassen.

- **Sequence Encoding**: Die gRNA-Sequenzpaarung wird hier ebenfalls in eine binäre Matrix kodiert.
- **Inception-Modul**: Diese Schicht nutzt CNNs, um Merkmale parallel zu extrahieren.
- **Bi-Directional LSTM**: Diese Schicht erfasst sequentielle Muster in beiden Richtungen der Sequenz und kombiniert die Ausgaben der Vorwärts- und Rückwärts-LSTM-Module. Dies ermöglicht eine umfassende Analyse der Sequenzinformationen.
- **Dense-Schichten**: Die Ausgabe der rekurrenten Schicht wird durch zwei dichte Schichten weiterverarbeitet, die zur finalen Vorhersage der Off-Target Aktivitäten führen.
- **CRISPR-Net-Aggregate**: Hier werden die individuellen Vorhersagen von CRISPR-Net zu einem einzigen Konsens-Score aggregiert.

4 Leistung der Modelle

Die Leistung von Modellen zur Vorhersage von Off-Target-Aktivitäten bei CRISPR-Cas9 wurde in mehreren Studien untersucht. In diesem Abschnitt werden die Ergebnisse des CNN und RNN von Charlier et al. (2021) und dem CRISPR-Net Modell von Lin et al. (2020) miteinander verglichen.
Zur Validierung der Leistung wird die Metrik Area Under the Receiver Operating Characteristic Curve (AUROC) verwendet. Dieser stellt die Leistungsfähigkeit eines Klassifikators dar und reicht von 0 bis 1. Ein AUROC von 0,5 deutet auf einen zufälligen Klassifikator hin, während ein AUROC von 1 auf einen perfekten Klassifikator hinweist. Je näher der AUROC-Wert an 1 liegt, desto besser ist die Leistungsfähigkeit des Modells in der Unterscheidung zwischen den Klassen.

4.1 Leistung des CNN

CNNs wurden verwendet, um die gRNA-DNA-Sequenzinformationen zu analysieren, indem sie diese als Bilddarstellung verarbeiten. Diese Modelle sind effektiv bei der Erkennung lokaler Merkmale und haben sich als nützlich bei der Vorhersage von Off-Target-Effekten erwiesen. In der Studie von Charlier et al. (2021) zeigte das CNN-Modell ein AUROC von 0,968 bei einer 5-schichtigen Architektur.

4.2 Leistung des RNN

RNNs, insbesondere Long Short-Term Memory (LSTM) Netzwerke, wurden ebenfalls zur Vorhersage von Off-Target-Effekten eingesetzt. Diese Modelle sind besonders gut darin, sequentielle Abhängigkeiten in den Daten zu erkennen. Ein typisches RNNs-Modell, das in der Studie von Charlier et al. (2021) verwendet wurde, bestand aus einer LSTM-Schicht mit 92 Neuronen, gefolgt von zwei dichten Schichten mit Batch-Normalisierung und Dropout. Dieses Modell erreichte einen AUROC von 0,907.

4.3 Leistung des CRISPR-Net

Das CRISPR-Net kombiniert CNN und RNN in einem hybriden Modell, das sowohl lokale Merkmale als auch sequentielle Informationen erfasst. Es verwendet eine Inception-basierte Konvolutionsschicht und eine bi-direktionale LSTM-Schicht (B-LSTM). In der Studie von Lin et al. (2020) zeigte CRISPR-Net eine durchschnittliche AUROC von 0.995. Darüber hinaus übertraf CRISPR-Net die bestehenden Methoden zur Vorhersage von Off-Target-Aktivitäten mit Missmatches und Indels.

5 Interpretation der Ergebnisse

Die vorliegenden Ergebnisse zeigen deutlich, dass die Wahl der Modellarchitektur einen erheblichen Einfluss auf die Leistungsfähigkeit bei der Vorhersage von Off-Target-Aktivitäten bei CRISPR-Cas9 hat. Die CNN-Modelle von Charlier et al. (2021) zeigten eine hohe Genauigkeit mit einem AUROC von 0,968 bei einer 5-Schicht-Architektur. Dies bestätigt die Effektivität von CNNs bei der Erkennung lokaler Merkmale. Dennoch haben CNNs Einschränkungen bei der Erfassung sequentieller Muster, die für eine umfassendere Analyse der gRNA-DNA-Interaktionen notwendig sind.

RNNs, insbesondere LSTM-Netzwerke, erzielten ebenfalls gute Ergebnisse mit einem AUROC von 0,907. Diese Modelle sind besonders gut darin, sequentielle Abhängigkeiten und langfristige Muster zu erkennen.

Das CRISPR-Net-Modell von Lin et al. (2020) übertraf beide Einzelmodelle mit einem AUROC von 0,995. CRISPR-Net kombiniert die Stärken von CNNs und RNNs, indem es sowohl lokale Merkmale als auch sequentielle Informationen erfasst. Die Inception-basierte Konvolutionsschicht ermöglicht eine detaillierte Merkmalsextraktion, während die bi-direktionale LSTM-Schicht sequentielle Muster in beide Richtungen analysiert.

6 Fazit und Ausblick

CRISPR-Net hat durch seine fortschrittliche Architektur gezeigt, dass es besser in der Lage ist, Missmatches und Indels vorherzusagen, was für die Minimierung von Off-Target-Effekten entscheidend ist. Dies macht es zu einem wertvollen Werkzeug für die CRISPR-Cas9-Genmodifizierung, insbesondere in klinischen Anwendungen, bei denen Genauigkeit und Zuverlässigkeit von großer Bedeutung sind.

Ausblick

Für zukünftige Forschung und Anwendungen könnte die Weiterentwicklung von CRISPR-Net durch die Integration weiterer biologischer Daten und fortschrittlicherer Algorithmen zur Mustererkennung noch robustere und genauere Modelle hervorbringen. Darüber hinaus könnte die Anwendung von CRISPR-Net auf eine breitere Palette von genetischen Daten und verschiedenen Organismen dazu beitragen, das volle Potenzial der CRISPR-Cas9-Technologie auszuschöpfen und ihre klinischen und therapeutischen Anwendungen zu erweitern. Langfristig könnte die Verfeinerung dieser Modelle dazu beitragen, personalisierte Medizinansätze zu fördern und die Sicherheit und Effizienz der Genmodifizierung zu maximieren.

Literatur

Aryeelab/Guideseq (2024), aryeelab.
URL: *https://github.com/aryeelab/guideseq*

Charlier, J., Nadon, R. & Makarenkov, V. (2021), 'Accurate deep learning off-target prediction with novel sgRNA-DNA sequence encoding in CRISPR-Cas9 gene editing', *Bioinformatics* **37**(16), 2299–2307.
URL: *https://doi.org/10.1093/bioinformatics/btab112*

Concordet, J.-P. & Haeussler, M. (2018), 'CRISPOR: Intuitive guide selection for CRISPR/-Cas9 genome editing experiments and screens', *Nucleic Acids Research* **46**(Web Server issue), W242–W245.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6030908/*

Haeussler, M. (2024), 'Maximilianh/crisporWebsite'.
URL: *https://github.com/maximilianh/crisporWebsite*

Lin, J. & Wong, K.-C. (2018), 'Off-target predictions in CRISPR-Cas9 gene editing using deep learning', *Bioinformatics* **34**(17), i656–i663.
URL: *https://doi.org/10.1093/bioinformatics/bty554*

Lin, J., Zhang, Z., Zhang, S., Chen, J. & Wong, K.-C. (2020), 'CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-Target Activities with Mismatches and Indels', *Advanced Science* **7**(13), 1903562.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.201903562*

Redman, M., King, A., Watson, C. & King, D. (2016), 'What is CRISPR/Cas9?', *Archives of Disease in Childhood. Education and Practice Edition* **101**(4), 213–215.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975809/*

Yin, H., Xue, W., Chen, S., Bogorad, R. L., Benedetti, E., Grompe, M., Koteliansky, V., Sharp, P. A., Jacks, T. & Anderson, D. G. (2014), 'Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype', *Nature Biotechnology* **32**(6), 551–553.
URL: *https://www.nature.com/articles/nbt.2884*

Zhu, L. J., Lawrence, M., Gupta, A., Pagès, H., Kucukural, A., Garber, M. & Wolfe, S. A. (2017), 'GUIDEseq: A bioconductor package to analyze GUIDE-Seq datasets for CRISPR-Cas nucleases', *BMC Genomics* **18**(1), 379.
URL: *https://doi.org/10.1186/s12864-017-3746-y*