Interpretierbarkeit der Genomsequenzen-basierten Deep Learning Modelle

Yichen Han
Helmholtz Zentrum für Infektionsforschung &
Institut für Statistik, LMU München

1. Hintergrund

Das Ziel ist es, sequenzbasierte Deep-Learning-Modelle zu nutzen, um potenzielle Biomarker, die einen Phänotyp beschreiben, zu identifizieren und diese der biowissenschaftlichen Forschungscommunity als Ressourcen bereitzustellen.

1. Hintergrund

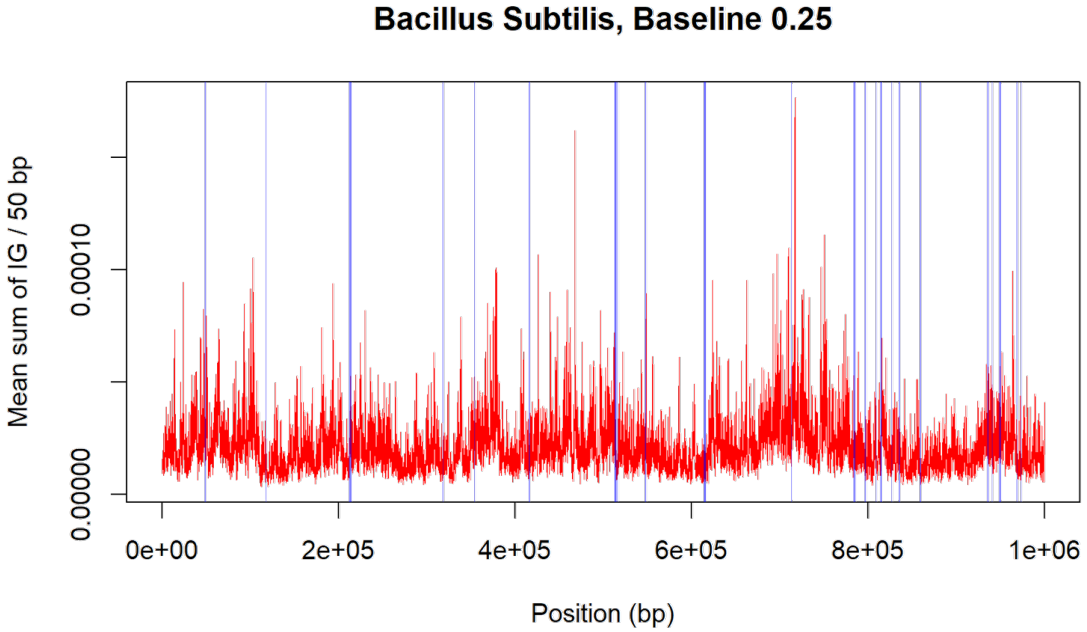
Modell: CNN
Daten: Bakterielle Genomsequenzen
Aufgabe: Phänotypvorhersage, Motiverkennung, Taxonomie, usw.

Integrated Gradients (Sundararajan et al. 2017)

- Importance scores sind schwer zu interpretieren und weisen nur einen geringen Zusammenhang mit phänotypspezifischen Genen auf.
- Nukleotide werden isoliert betrachtet, mögliche Interaktionseffekte bleiben unberücksichtigt.
- Eine geeignete Vergleichssequenz (*Baseline*) muss noch subjektiv gewählt werden.

Beispiel: Sporenbildung

Blaumarkierte Segmente: sporenbezogene Gene, annotiert mit *Prokka*.



Auch das genbasierte Random-Forest-Modell wählt keine sporenbezogenen Gene als wichtige Merkmale aus.

2. Hypothesen

IG kann sequenzbasierte Modelle meistens nicht gut interpretieren, denn…

i. Die Methode ist für Genomsequenzen ungeeignet, da z. B. Interaktionen ignoriert werden.

ii. Das Modell „cheatet“, indem es artspezifische statt aufgabenspezifischer Eigenschaften erkennt.

iii. Durch die Schichten und das Pooling im CNN-Modell gehen kleinere Signale verloren.

iv. Gene oder Eigenschaften, die biologisch relevant sind, können dennoch suboptimale Prädiktoren für die spezifische Aufgabe sein.

3. Zwischenstand und Ziele

i. Eine Erasure-Methode basierend auf einem evolutionären Algorithmus.
   - Best Individual: Blende 10% (ca. 100 Segmente à 1000 bp) einer 1 Mbp Sequenz aus, und die Konfidenz der korrekten Prädiktion sinkt von 99.9% auf 6%.
   - Optimierung, Varianzreduktion, *ad hoc, post hoc*?

ii. Synthetische Daten generieren und kontrollierbare Aufgaben definieren, die der Komplexität bakterieller Genomszenarien ähneln, um die Methoden zu evaluieren.

iii. Modelle mit unterschiedlichen Konfigurationen trainieren, um die passendste Spezifikation im genomischen Kontext zu identifizieren.

Vielen Dank für Ihre Aufmerksamkeit!
Ich freue mich auf Rückfragen und den Austausch, falls Interesse besteht.

Yichen.Han@campus.lmu.de