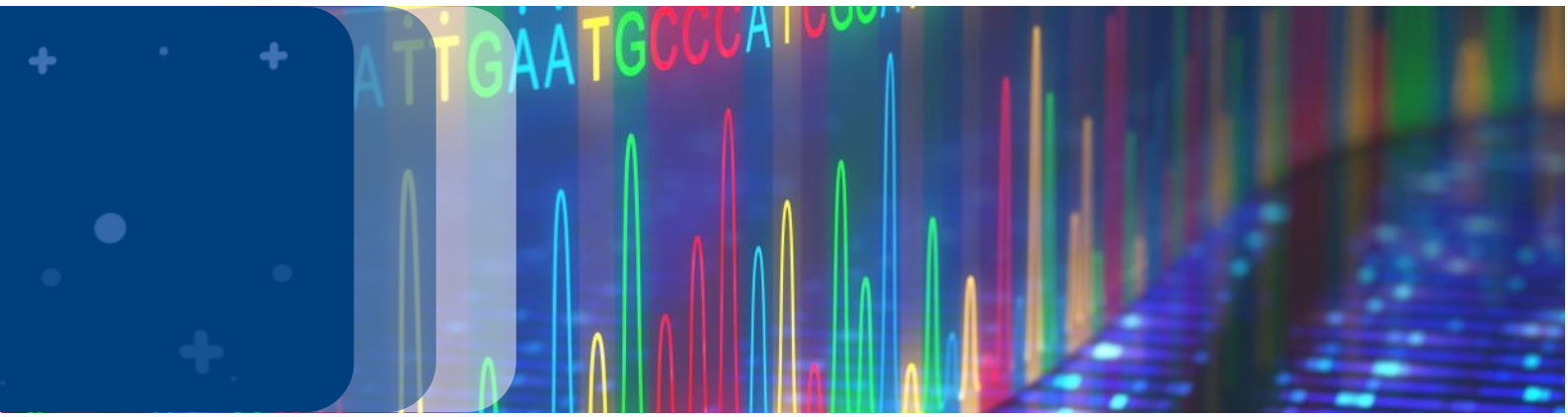


Clustering-Ansatz zur Gruppierung und Identifizierung von DNA-Sequenzen aus Mikrobiom-Sequenzen

Projektziel

Das menschliche Darmmikrobiom ist eine komplexe und dynamische Vereinigung von Mikroorganismen, die den Magen-Darm-Trakt bewohnen. Es trägt zu verschiedenen Aspekten der Wirtsphysiologie bei, darunter Stoffwechsel, Reifung des Immunsystems, Verhalten und sogar die Neurologie. Durch die Beobachtung der Veränderungen der Darm-Mikrobiota kann die Zusammensetzung einen Einblick in die Interaktion zwischen Wirt und Mikrobiom geben und neue Optionen für therapeutische Interventionen vorschlagen.

Ziel des Projekts war es, eine Pipeline zu entwickeln, um die Mikrobiomsequenzen aus dem Dickdarm in der Reihenfolge der Nukleotide zu bündeln und anschließend die gebildeten Cluster mit bekannten Sequenzen durch den Sprengalgorithmus zu kombinieren. Nach der Hypothese können die homogenen Sequenzen im gleichen Cluster zusammengefasst werden, was der gleichen Taxonomie entsprechen könnte. Die repräsentativen Sequenzen aus dem Cluster wurden dann mit der bekannten Bakteriengattung oder -art in der NCBI-Datenbank abgeglichen. Die Ergebnisse des Projekts können für die klinische Behandlung weiterverwendet werden.

Genutzte Daten

Die Proben wurden während der Koloskopie aus dem Darm entnommen. Das Koloskop wurde eingeführt und vom Rektum durch den gesamten Dickdarm bis zum Zäkum nach vorne geschoben. Auf dem Rückweg wurde die Luftinjektion eingesetzt und der Darm entfaltet, so dass die gesamte Schleimhaut gescannt werden kann. Die RNA wurde aus dem Dickdarm isoliert und die 16S rRNA-Sequenzierung wurde zur Identifizierung der Bakterien durchgeführt. Die Rohdaten (.FASTQ) wurden zur Verfügung gestellt.

Herausforderungen

Erstens, die Mischung aus dem menschlichen Genom und dem Bakteriengenom erhöht die Schwierigkeit der Identifizierung der Mikrobiota-Arten. Es wurden keine Vorabinformationen für das Clustering bereitgestellt, daher sollte die unbeaufsichtigte Methode verwendet werden. Zweitens kann der technische Fehler der Maschine während der Sequenzierung auftreten. Es kann nutzlose Informationen (Nukleotide, die als "N" bezeichnet werden) oder sogar falsche Informationen über das Bakteriengenom erzeugen. Drittens, durch die Einschränkung der Technik, können die Proben nicht alle gleichzeitig laufen. Daher sollte der Batch-Effekt normalisiert werden.

Angewandte Methoden

Self Organizing Map (SOM)

SOM ist eine Art künstliches neuronales Netzwerk (ANN), das durch unbeaufsichtigtes Lernen trainiert wird, um eine niedrigdimensionale diskrete Darstellung der Eingabe zu erzeugen. SOM kann den Batch-Effekt anpassen und den falsch positiven Fehler durch technischen Fehler kontrollieren. Alle Sequenzen müssen vom Primer-Start an bis zu 250 Nukleotide für das Training geschnitten werden. Unter Berücksichtigung der Tatsache, dass SOM nur mit numerischen Daten arbeitet, wurden die Nukleotide zunächst in Zahlen übersetzt. Die Nukleotide A, C, T, G und N wurden durch die Zahlen 1, 2, 3, 4, 5 ersetzt. Während des Trainingszyklus wurden alle Sequenzen 10.000 auf das Netzwerk verteilt und die Batch-Effekt-Anpassungen durchgeführt. Für jeden Cluster wurde eine entsprechende Heatmap gespeichert. Jedes Nukleotid erhielt eine eigene Farbe zur Visualisierung der Cluster-Homogenität. Der Mittelwert der Cluster-Korrelation unter 0,95 war ein entscheidender Wert für den zweiten SOM-Lauf zur Identifizierung der Subcluster von besserer Qualität.

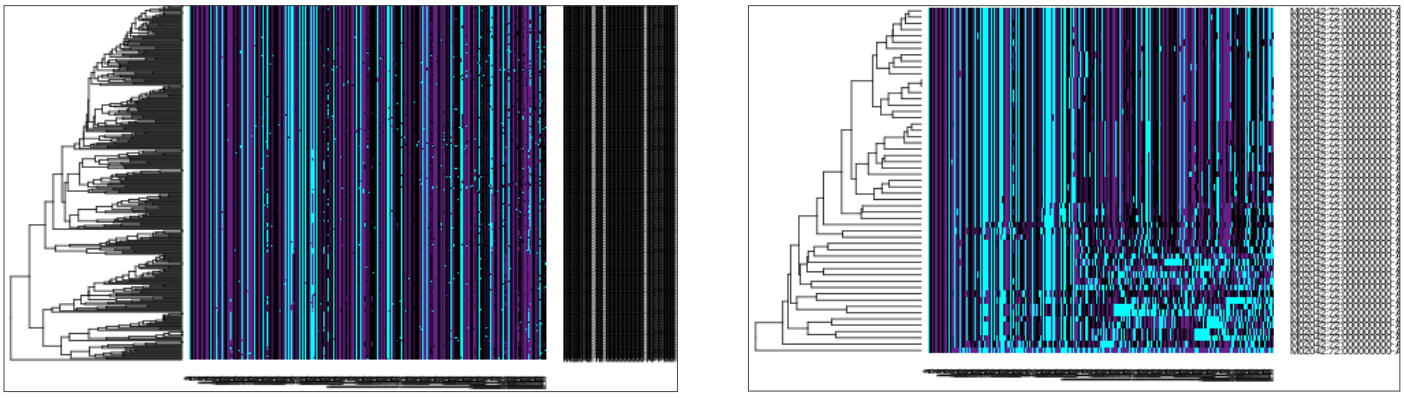
Somit befand sich jede Sequenz nach zwei SOM-Läufen in einem geeigneten Cluster mit einer durchschnittlichen Clusterhomogenität von über 90%.

Basic Local Alignment Search Tool (BLAST)

Der BLAST-Algorithmus wurde verwendet, um zu identifizieren, welche bekannten Sequenzen von Mikroorganismen in den detektierten Proben präsentiert werden. Es kann das menschliche Genom und das Bakteriengenom trennen. Um die Betriebszeit zu minimieren, wurden 50 Vertreter aus jedem der größten Cluster zufällig ausgewählt und als FASTA-Datei gespeichert. Schließlich wurden für jeden Cluster die häufigsten BLAST-Ergebnisse ermittelt.

Projektergebnis

Die SOMs lieferten Clustering mit hoher Homogenität. Neuronale Netze ermöglichten es, Ähnlichkeiten in einer großen Anzahl von DNA-Sequenzen basierend auf der euklidischen Entfernung zu identifizieren. Zusätzlich kann ein Benutzer anhand von Heatmaps erkennen, welches Muster in einem jeweiligen Cluster dominiert.



Heatmaps von Clustern enthielten 389 Sequenzen mit einem mittleren Score von 0,9511 (links) und 57 Sequenzen mit einem mittleren Score von 0,7504053 (rechts).

Nach den BLAST-Ergebnissen ermöglicht uns der identifizierte Mikroorganismus ein tieferes Verständnis der Mikroorganismenzusammensetzung bei den Patienten. Die Ergebnisse können für die Untersuchung der Wirts-Mikrobiota-Interaktion, der Medikamentenwirkung und des Umwelteinflusses weiterverwendet werden.

Cluster	Anzahl Sequenzen	Häufigste BLAST Ergebnisse
4	9505	Ruminococcus flavefaciens
5	5971	Ruminococcus bromii
6	5566	Bacteroides thetaiotaomicron 8713
7	5195	Bacteroides vulgatus strain JCM 5826
8	4397	Clostridium fusiformis
9	3810	Ruminococcus sp. ID1
10	4782	Coprococcus catus VPI-C6-61