COVID MODELL VORHERSAGEN

INTRO

Das Ziel dieses Projekts ist es, die evolutionären Patterns hinter Covid zu erforschen und diese letztendlich simulieren zu können. Dadurch würde es möglich, neue Varianten frühzeitig zu erkennen und Vorbeugungsmaßnahmen zu treffen. Dafür versuchen wir Modelle, basierend auf künstlicher Intelligenz, einzusetzen, um Tausende von Sequenzen zu bearbeiten und möglichst viel Information zu extrahieren. Das Modell soll die Abhängigkeiten zwischen verschiedenen aufgetretenen Mutationen erlernen, insbesondere Kombinationen, die sich in der Vergangenheit als gut funktionierend erwiesen haben, damit es realistische Evolutionspfade generieren kann.

Die Data ist aus GISAID zu entnehmen, wobei sich da gerade mehr als 15 Millionen Samples zu finden sind. Die Bearbeitung der hohen Anzahl an allen möglichen Mutationskonfigurationen würde rechnerische Schwierigkeiten darstellen, weshalb es sich als sinnvoll aufweist, maschinelles Lernen zu verwenden.

METHODEN

Wir implementieren ein Encoder-Decoder Network, das aus zwei rekurrenten neuralen Netzen (RNN) - der Decoder und der Encoder - zusammengesetzt ist. Ausgehend von einer Eingabesequenz, die in unserem Fall eine bestimmte Variante von Covid Sequenz ist, transformiert sie der Algorithmus in einer neuen Sequenz – eine neue Nachfolger Sequenz, in der sich die ursprüngliche Sequenz entwickeln kann. Die beiden Netze arbeiten zusammen, um die Bedeutung der angegebenen Sequenz zu encoden und diese dann zu übersetzen. Dieser Ansatz wird meistens in Übersetzungsaufgaben benutzt, findet aber auch in der Biologie und genau in der Virenforschung Anwendung.

Wir arbeiten auf der Nukleotid Ebene, damit wir die Evolution ganz vom Beginn betrachten können. Die Covid Sequenzen sind fast 30 000 Basenpaaren lang, was die Berechnung besonders erschwierigt. Wir wollten aber den Genom nicht in Genen aufteilen, sondern war die Idee, die Vorhersagen anhand der ganzen Sequenz zu erfolgen, damit alle Mutationsverhältnisse abgedeckt werden können. Aus diesem Grund haben wir eine vereinfachte Darstellung der Sequenzen gewählt - jede Variante wird als ein Set von Mutations dargestellt.

Die Trainingsdata enthält mehr als 500 000 Beispiele, die das Modell analysieren kann. Es beinhaltet Evolutionspaare – eine Ancestor Sequenz zusammen mit einer Descendant Sequenz. Die Paaren wurden mithilfe der pangolin lineages baum gebildet. Ein Trainingspaar würde zum Beispiel eine Sequenz BA.2.75 mit einer BA.2.75.1, oder BA.2.75.2.

Bei Eingabe der Ancestor Sequenz soll dann das Modell mögliche Evolutionspfade empfehlen. Die Sequenz BA.2.75 kann sich entsprechend zu BA.2.75.1 oder zu BA.2.75.2 entwicklen mit gewissee Wahrscheinlichkeit, was von der Häufigkeit dieser Entwicklung in der reelen Data abhängt. Das würde durch Beam Search erfolgen. Auf jedem Schritt werden die k-höchstwahrscheinlichsten Mutationen gelistet, mit denen dann die Suche sich verzweigt.

BEISPIEL

Das Modell ist noch in der Training Phase, wobei viele Parameter noch angepasst werden können. Trotzdem liefert es vernünftige Resultate. Zum Beispiel beginned mit einer beliebigen Sequenz von GISAID (Variante BA.2.57) und lassen das Modell laufen. Es sagt neue mutationen vorher, die addiert werden konnten.

FURHTER RESEARCH

Das Modell operiert nur basierend auf dem bis jetzt betrachteten Verhalten und der Evolution von Covid Varianten. Es kann weiter generalisiert werden, indem wir die ganzen Sequenzen kodieren, nicht nur die Mutationen. Das würde mehr computational Power verlangen, aber vielleicht liefert es bessere Resultate.

Das Modell kann auch für andere Viren benutzt werden, provided es gibt genug sequenzierten Data. Dieses Thema ist nicht so weit recherchiert, es gibt nicht so viele Papers mit solchen Modellen, most notable: Mutagan etc. Die benutzen dieselben Idee, für die Avianinfluenza, anhand Phylogenetische Bäume.