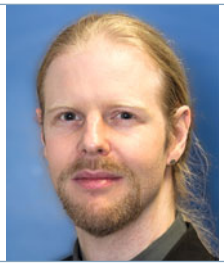
„VARIANT CALLING IST EIN WICHTIGER SCHRITT IN DER MEDIZINISCHEN FORSCHUNG, DIAGNOSTIK, VERSORGUNG UND THERAPIE. JE ZUVERLÄSSIGER UND PRÄZISER DIESE ANALYSE IST, DESTO GENAUER KÖNNEN VARIATIONEN MIT KRANKHEITS- UND THERAPIEVERLAUF ASSOZIIERT WERDEN."


Axel Fürstberger


Matthias Platzer


Hans A. Kestler

Googles DeepVariant: eine Methode für die Medizin- und Bioinformatik?

DOI: 10.1007/s12268-018-0911-z

„Big Data" und „Personalisierte Medizin" sind die biomedizinischen Schlagwörter der letzten Jahre. Große Konzerne entwickeln und verwenden Algorithmen, die uns gezielt z. B. Werbung anbieten und/oder den idealen Partner versprechen. Parallel dazu gelangt die Erkennung und Vorhersage von Profilen und Posts in den sozialen Netzwerken immer wieder in die Medien. Im Moment ist Facebook wegen weitergeleiteter Nutzerdaten einer App für Persönlichkeitsprognose namens *„thisisyourdigitallife"* an die Analysefirma Cambridge Analytica in den Schlagzeilen [1].

Mit einem Variant Calling-Werkzeug im Bereich der Bioinformatik macht derzeit Google von sich reden: DeepVariant, ein Programm zur Vorhersage von genetischen Variationen (d. h. durch Mutationen entstandene heterozygote und von einer Referenzsequenz abweichende homozygote Allele) in resequenzierten Genomdaten. DeepVariant basiert dabei auf Googles Cloud-Software Development Kit (SDK) und TensorFlow.

Genetische Varianten können sowohl neutrale Polymorphismen darstellen als auch Prädispositionen oder Resistenzen für verschiedene Krankheiten verursachen, darunter monogene Erbkrankheiten oder komplexe Volkserkrankungen wie Diabetes und Krebs. Das Auffinden dieser Varianten, also das Variant Calling, ist ein wichtiger Schritt in der medizinischen Forschung, Diagnostik, Versorgung und Therapie. Je zuverlässiger und präziser diese Analyse ist, desto genauer können Variationen mit Krankheits- und Therapieverlauf assoziiert werden.

Die Daten sind Sequenzabschnitte aus dem Genom (*reads*, das heißt die unmittelbaren Ergebnisse des Sequenzierprozesses), die hierbei verarbeitet werden müssen. Sie liegen im Bereich von mehreren Gigabyte oder Terabyte, abhängig vom Organismus und von der Sequenziertiefe.

DeepVariant baut beim Variant Calling auf dem Open-Source-*machine learning framework* TensorFlow auf. TensorFlow stellt unter anderem Lernmethoden für neuronale Netze bereit. Die Merkmale vorverarbeiteter Sequenzierdaten werden dazu zuerst als Multikanaltensordarstellung encodiert, ähnlich einem multidimensionalen Vektor oder einer Zahlenmatrix, und dann für die Prädiktion verwendet [2]. Dabei wird zwischen homozygot, heterozygot und alternativ homozygot unterschieden. Das Finden und Zusammentragen dieser Variationen ist ein etablierter Schritt bei der Genomanalyse.

Der Vergleich eigener Daten aus Nanopore- und Illumina-Sequenzierungen mit anderen Variant Calling-Programmen zeigt, dass DeepVariant nicht immer das beste Ergebnis liefert. Die Fehlerraten sanken in Tests, waren jedoch im Bereich der Insertion/Deletion(Indel)-Erkennung sogar je nach Datensatz auch höher [3]. Encodieren der Daten und Prädiktion der Mutationen sind rechenintensive Schritte. Im Hinblick auf Computerressourcen stellt das Tool höhere Anforderungen an die Hardware – etwa doppelt bis zehnmal so viele Rechen-Core-Stunden im Vergleich zu anderen Programmen. Diese Verarbeitung in die Cloud auszulagern, kann je nach Daten (z. B. menschliches Genom) und geltendem Datenschutzrecht schwierig bis unmöglich sein.

Für das Variant Calling gibt es bereits einige Software-Werkzeuge. Die dort verwendeten Methoden werden stetig verbessert, auf neue Hardware optimiert und parallelisiert. Dieser Analyseschritt ist ein etablierter Teil der Verarbeitung genomischer Sequenzdaten. Das auf Googles TensorFlow basierende DeepVariant bringt also keine grundsätzlichen Neuerungen ins Spiel, schafft es aber, einen Denkanstoß zu geben, ob diese Art der Methodik, quasi *unbiased* aus vielen Daten zu lernen, sich auch in anderen Bereichen der Bioinformatik und Medizininformatik sinnvoll anwenden lässt.

DeepVariant übernimmt lediglich *einen* Teilschritt des Sequenzanalyseprozesses. Dies ist auch nicht der primäre Schritt bei der Genomentschlüsselung, wie es die *Frankfurter Allgemeine Zeitung* in ihrem Artikel suggeriert [4]. Ebenso kann es auch nicht als „das angesehenste Programm auf dem Markt" in diesem Bereich bezeichnet werden. Es erzielt beim Variant Calling gute Ergebnisse, benötigt jedoch deutlich mehr Rechenressourcen.

Wie gut das Programm auch im Bereich der Forschung und Diagnostik angenommen wird, muss die Zukunft zeigen.

Axel Fürstberger
Institut für Medizinische Systembiologie, Universität Ulm

Matthias Platzer
Leibniz-Institut für Alternsforschung, Jena

Hans Armin Kestler
Institut für Medizinische Systembiologie, Universität Ulm

Korrespondenzadressen:
PD Dr. Matthias Platzer
Leibniz-Institut für Alternsforschung
Beutenbergstraße 11
D-07745 Jena
matthias.platzer@leibniz-fli.de

Dr. Axel Fürstberger
Prof. Dr. Hans A. Kestler
Institut für Medizinische Systembiologie
Universität Ulm
D-89081 Ulm
hans.kestler@uni-ulm.de
axel.fuerstberger@uni-ulm.de

Literatur

[1] Cambridge Analytica: Facebook schließt umstrittenes Unternehmen aus, www.heise.de/newsticker/meldung/Cambridge-Analytica-Facebook-schliesst-umstrittenes-Unternehmen-aus-3997615.html
[2] Poplin R, Newburger D, Dijamco J et al. (2017) Creating a universal SNP and small indel variant caller with deep neural networks. BioRxiv, doi: 10.1101/092890
[3] Carroll A, Thangaraj N (2017) Evaluating DeepVariant: a new deep learning variant caller from the Google Brain team. Inside DNAnexus, http://blog.dnanexus.com/2017-12-05-evaluating-deepvariant-googles-machine-learning-variant-caller
[4] Frankfurter Allgemeine, Google veröffentlicht Künstliche Intelligenz zur Genom-Entschlüsselung, www.faz.net/aktuell/wirtschaft/kuenstliche-intelligenz/google-macht-ki-software-deep-variant-zu-open-source-15332900.html