Computational Biology

Künstliche Intelligenz gegen den Fehlerteufel

Moderne technische Möglichkeiten erlauben es, einzelne Zellen zu sequenzieren und jeweils individuell herauszufinden, welche Gene gerade abgelesen werden. Diese Methoden sind sehr fein, dadurch aber auch sehr fehleranfällig: Geräte, Umwelt aber auch die Biologie selbst können für Ausfälle und Unterschiede zwischen den Messungen verantwortlich sein. Forscher des Helmholtz Zentrums München haben gemeinsam mit Kollegen der Technischen Universität München (TUM) und des englischen Wellcome Sanger Institute nun Algorithmen entwickelt, die diese Fehlerquellen berechen- und korrigierbar machen. Die Ergebnisse sind in ‚Nature Methods' und ‚Nature Communications' erschienen.

Es ist ein visionäres Vorhaben von enormen Ausmaßen – das Human Cell Atlas-Projekt kartiert alle Gewebe des menschlichen Körpers zu verschiedenen Zeitpunkten mit dem Ziel eine Referenzdatenbank zur Entwicklung personalisierter Medizin zu schaffen, also ‚gesunde' und ‚kranke' Zellen vergleichen zu können. Möglich wird das durch sogenannte Einzelzell-RNA-Sequenzierung – also vereinfacht gesagt: die Möglichkeit, nachzuvollziehen, welche Gene diese winzigsten Bausteine des Lebens gerade an- oder ausschalten. „Das ist methodisch gesehen ein enormer Sprung, denn früher waren solche Daten immer nur aus großen Gruppen von Zellen zu gewinnen, weil die Messungen so viel RNA benötigten", erklärt Maren Büttner. „Die Ergebnisse waren also immer nur der Mittelwert aller eingesetzter Zellen, heute bekommen wir für jede einzelne Zelle exakte Daten", so die Doktorandin am Institute of Computational Biology (ICB) des Helmholtz Zentrums München.

Durch die feineren Messungen steigt allerdings auch die Anfälligkeit für den sogenannten Batch-Effekt. „Dabei handelt es sich um Abweichungen zwischen mehreren Messungen, die beispielsweise bereits entstehen können, wenn die Temperatur des Gerätes leicht abweicht oder sich die Verarbeitungszeit der Zellen verändert", erklärt Maren Büttner. Zwar gäbe es hier verschiedene Modelle, um den Fehler herauszurechnen, allerdings sind diese Methoden stark davon abhängig, wie groß der Effekt eigentlich ist. „Um das herauszufinden, haben wir eine nutzerfreundliches, robustes und sensitives Maß namens kBET entwickelt, dass Unterschiede zwischen Experimenten quantifiziert und damit verschiedene Korrektur-Ergebnisse vergleichbar macht", sagt Büttner.

„Unsere neue Methode kBET ist ein leistungsfähiges Werkzeug für den Vergleich von Batch-Effekt-Korrekturschemata, mit dem Forscher verschiedene Datensätze für Einzelzell-RNA-Sequenzierungen untersuchen können. Dies hat in Zukunft auch wichtige Auswirkungen für die Datenintegration, die für wichtige Initiativen wie den Human Cell Atlas von zentraler Bedeutung ist ", sagt Dr. Sarah Teichmann, Corresponding Author vom Wellcome Sanger Institute in Großbritannien.

Neben dem Batch-Effekt sind sogenannte Null-Messungen (englisch: dropout events) bei der Einzelzellsequenzierung eine große Herausforderung. „Wir sequenzieren also eine Zelle und stellen fest, dass ein bestimmtes Gen in dieser Zelle überhaupt kein Signal von sich gibt", veranschaulicht ICB-Direktor Prof. Dr. Dr. Fabian Theis. „Dahinter kann sich nun ein biologischer oder ein technischer Grund verbergen: Entweder wird das Gen nicht abgelesen, weil es in diesem Moment schlicht keine Rolle spielt, oder aber die Sequenz ist aus technischen Gründen nicht erfasst worden", so der Professor für Mathematische Modellierung biologischer Systeme an der TUM.

Um diese Fälle zu erkennen, nutzten die Bioinformatiker Gökcen Eraslan und Lukas Simon aus Theis' Gruppe die große Anzahl der Datenpunkte und entwickelten einen sogenannten Deep Learning Algorithmus. Dabei handelt es sich um künstliche Intelligenz, die Lernprozesse simuliert, wie sie auch beim Menschen vorkommen (neuronale Netze).*

Über ein neues Wahrscheinlichkeitsmodell und Vergleich der ursprünglichen und rekonstruierten Daten ermittelt der Algorithmus, ob in diesem Fall ein biologischer oder ein technischer Ausfall zugrunde liegt. „Durch dieses Modell lassen sich sogar Zelltyp-spezifische Korrekturen ermitteln, ohne dass sich zwei unterschiedliche Zelltypen künstlich ähnlicher werden", so Fabian Theis. „Als einer der ersten Deep Learning Methoden im Bereich Einzelzell-Genomik hat der Algorithmus den weiteren Vorteil, gut auf Datensätze mit Millionen von Zellen zu skalieren."

Eines – das ist den Wissenschaftlern wichtig – ist die Methode aber nicht: „Wir bauen hier keine Software, um Ergebnisse beliebig zu ‚glätten'. Unser Ziel ist es vor allem, Fehler ausfindig zu machen und zu korrigieren", so Fabian Theis. „Mit diesen möglichst korrekten Daten können wir dann in den Austausch mit unseren Kollegen weltweit gehen und unsere Ergebnisse mit ihren vergleichen." Beispielsweise, wenn die Helmholtz-Forscher ihren Anteil für den Human Cell Atlas beisteuern, denn gerade hier ist die Verlässlichkeit und die Vergleichbarkeit der Daten von größter Wichtigkeit.

Weitere Informationen

* Die neue Methode, der sogenannte „Deep Count Autoencoder", lernt eine einfachere Darstellung der komplexen Daten, indem diese komprimiert und anschließend wieder rekonstruiert werden.

Hintergrund:
Die Arbeit in ‚Nature Methods' entstand in enger Zusammenarbeit mit Dr. Sarah Amalia Teichmann vom Wellcome Trust Sanger Institute. Sie ist Co-Vorsitzende des Organisationskomitees für den Human Cell Atlas und war 2017 mit einem Helmholtz International Fellow Award ausgezeichnet worden, der die Zusammenarbeit von Helmholtz-Wissenschaftlerinnen und Wissenschaftlern mit hervorragenden Kollegen im Ausland fördern soll, was in diesem Fall offenbar gut gelungen ist.

Original-Publikationen:
Büttner, M. et al. (2019): A test metric for assessing single-cell RNA-seq batch correction. Nature Methods, DOI: 10.1038/s41592-018-0254-1

Eraslan, G. & Simon, L.M. (2019): Single cell RNA-seq denoising using a deep count autoencoder. Nature Communications, DOI: 10.1038/s41467-018-07931-2

Verwandte Artikel:
Ran an die Datenberge – neue Graduiertenschule für Data Science in München
Einheitliche Standards für epigenetische Daten gefordert
Kompletter Zellatlas eines unsterblichen Plattwurms
Die Software Scanpy verarbeitet riesige Mengen an Einzelzelldaten

Das Helmholtz Zentrum München verfolgt als Deutsches Forschungszentrum für Gesundheit und Umwelt das Ziel, personalisierte Medizin für die Diagnose, Therapie und Prävention weit verbreiteter Volkskrankheiten wie Diabetes mellitus und Lungenerkrankungen zu entwickeln. Dafür untersucht es das Zusammenwirken von Genetik, Umweltfaktoren und Lebensstil. Der Hauptsitz des Zentrums liegt in Neuherberg im Norden Münchens. Das Helmholtz Zentrum München beschäftigt rund 2.300 Mitarbeiter und ist Mitglied der Helmholtz-Gemeinschaft, der 18 naturwissenschaftlich-technische und medizinisch-biologische Forschungszentren mit rund 37.000 Beschäftigten angehören. www.helmholtz-muenchen.de

Das Institut für Computational Biology (ICB) führt datenbasierte Analysen biologischer Systeme durch. Durch die Entwicklung und Anwendung bioinformatischer Methoden werden Modelle zur Beschreibung molekularer

Prozesse in biologischen Systemen erarbeitet. Ziel ist es, innovative Konzepte bereitzustellen, um das Verständnis und die Behandlung von Volkskrankheiten zu verbessern. www.helmholtz-muenchen.de/icb

Die Technische Universität München (TUM) ist mit rund 550 Professorinnen und Professoren, 41.000 Studierenden sowie 10.000 Mitarbeiterinnen und Mitarbeitern eine der forschungsstärksten Technischen Universitäten Europas. Ihre Schwerpunkte sind die Ingenieurwissenschaften, Naturwissenschaften, Lebenswissenschaften und Medizin, verknüpft mit den Wirtschafts- und Sozialwissenschaften. Die TUM handelt als unternehmerische Universität, die Talente fördert und Mehrwert für die Gesellschaft schafft. Dabei profitiert sie von starken Partnern in Wissenschaft und Wirtschaft. Weltweit ist sie mit dem Campus TUM Asia in Singapur sowie Verbindungsbüros in Brüssel, Kairo, Mumbai, Peking, San Francisco und São Paulo vertreten. An der TUM haben Nobelpreisträger und Erfinder wie Rudolf Diesel, Carl von Linde und Rudolf Mößbauer geforscht. 2006 und 2012 wurde sie als Exzellenzuniversität ausgezeichnet. In internationalen Rankings gehört sie regelmäßig zu den besten Universitäten Deutschlands. www.tum.de

Ansprechpartner für die Medien
Abteilung Kommunikation, Helmholtz Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Ingolstädter Landstr. 1, 85764 Neuherberg - Tel. +49 89 3187 2238 - E-Mail: presse@helmholtz-muenchen.de

Fachlicher Ansprechpartner
Prof. Dr. Dr. Fabian Theis, Helmholtz Zentrum München, Institut für Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany, Tel. +49 89 3187-4030 - E-Mail: fabian.theis@helmholtz-muenchen.de