MedWiss.Online

Die Software Scanpy verarbeitet riesige Mengen an Einzelzelldaten

Wissenschaftler des Helmholtz Zentrums München haben ein neues Programm entwickelt, das große Datensätze beherrschbar machen soll. Die Software mit dem Namen Scanpy ist beispielsweise ein Kandidat für die Auswertung des Human Cell Atlas Projekts und wurde nun in ‚Genome Biology' publiziert.

„Es geht um die Analyse von Genexpressionsdaten* zahlreicher einzelner Zellen", erklärt Erstautor Alex Wolf vom Institute of Computational Biology (ICB) des Helmholtz Zentrums München. Er hat Scanpy entwickelt, gemeinsam mit seinem Kollegen Philipp Angerer in der Machine Learning Gruppe von Institutsdirektor Prof. Dr. Dr. Fabian Theis, der neben seiner Position am Helmholtz Zentrum auch Professor für Mathematische Modelle biologischer Systeme an der TU München ist. „Die neue technische Möglichkeiten generieren um Größenordnungen mehr Daten mit dementsprechend höherer Information", schildert Theis. „Allerdings war die historisch gewachsene Software-Infrastruktur zur Genexpressionsanalyse nicht auf die neuen Herausforderungen ausgelegt." Entsprechend groß sei hier der Bedarf nach neuen Analysemethoden.

Im Rennen für den Human Cell Atlas

Auch ein großes internationales Forschungsvorhaben könnte Theis zufolge von der Software profitieren. Unter dem Namen ‚Human Cell Atlas' tragen zahlreiche internationale Wissenschaftler eine Referenzdatenbank zusammen, in der die Genaktivität aller menschlichen Zelltypen erfasst ist. „Für dieses Projekt oder auch bei der immer häufiger werdenden Zusammenlegung von bestehenden Datensätzen ist es wichtig, eine skalierbare Software zu haben", so Theis. Entsprechend sei Scanpy aktuell in der Auswahl für die Analysesoftware des Human Cell Atlas.

„Mit Scanpy publizieren wir die erste Software, die eine umfängliche Analyse großer Genexpressionsdatensätze mit einem breiten Spektrum aus Methoden des maschinellen Lernens und Statistik erlaubt", beschreibt Alex Wolf den Fortschritt. „Bereits jetzt wird die Software in diversen Gruppen weltweit eingesetzt, insbesondere auch am Broad Institute von Harvard und dem Massachusetts Institute of Technology."

Technologisch beschreitet die Anwendung neue Wege: Während entsprechende Biostatistik-Software traditionell in der Programmiersprache R geschrieben wurde, basiert Scanpy auf der Sprache Python, die die Machine Learning Community dominiert. Neu ist zudem, dass Graph-basierte Algorithmen das Herz von Scanpy bilden. Anstatt Zellen wie bisher üblich als Punkte im Koordinatensystem des Genexpressionsraums zu betrachten, verwenden die Algorithmen ein graphartiges Koordinatensystem. Das heißt, anstatt eine Zelle mit dem Expressionswert einiger Tausend Gene zu charakterisieren, wird sie einfach durch die Angabe ihrer nächsten Nachbarn charakterisiert – vergleichbar mit Verbindungen in sozialen Netzwerken. Wenn es zum Beispiel um die Identifikation von Zelltypen geht, verwendet Scanpy also die gleichen Algorithmen wie Facebook zur Identifikation von Communities.

Weitere Informationen

* Die Expression beschreibt, wie häufig ein Gen abgelesen wird, gibt also Aufschluss über die Aktivität des Gens.

Hintergrund:
Das Team und Alex Wolf konnte erst kürzlich einen der vorderen Plätze beim Data Science Bowl belegen, einem der weltweit höchstdotierten Wettbewerbe zum Thema Big Data. In ihrem Beitrag hatte das Team einen Algorithmus programmiert, der binnen weniger Millisekunden Lungenkrebs auf Basis von 300 Schichten eines dreidimensionalen Computertomographie (CT)-Scans erkennt – ein Vorgang für den ein Radiologe im schlechtesten Fall mehrere Stunden benötigen würde. Des Weiteren publizierte das Team kürzlich einen Artikel in Nature Communications zur Rekonstruktion von Zellentwicklung aus Einzelzellbildern: Malen nach Zahlen: Algorithmus rekonstruiert Prozesse aus Einzelbildern**.**

Original-Publikation:
Wolf, A. et al. (2018): Scanpy: large-scale single-cell gene expression data analysis. Genome Biology, DOI: 10.1186/s13059-017-1382-0

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.

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.

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 – Fax: +49 89 3187 3324 – E-Mail: presse@helmholtz-muenchen.de

Fachlicher Ansprechpartner:
Dr. Dr. Alexander Wolf, Helmholtz Zentrum München – Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), Institute of Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg – Tel. +49 89 3187 4217 – E-Mail: alex.wolf@helmholtz-muenchen.de