
Big Data und Körpergröße: Algorithmische Innovation bei genetischer Analyse komplexer Merkmale

18.02.2026, 17:00 | Wissenschaft, Forschung, Bildung

Pressemitteilung von: *idw - Informationsdienst Wissenschaft*



Das Extrahieren und Analysieren relevanter medizinischer Informationen aus großen Datenbanken wie Biobanken stellt eine erhebliche Herausforderung dar. Frühere Versuche solche „Big Data“ zu nutzen, konzentrierten sich auf große Stichprobenalgorithmen, die einzelne Datenpunkte modellieren. Da diese Algorithmen jedoch den gesamten Datensatz millionenfach abtasten müssen, ist ihre theoretisch sehr hohe Genauigkeit mit unerschwinglichen Rechenkosten verbunden und bleibt daher praktisch unerreichbar. Um diese Hürde zu überwinden, haben Wissenschaftler:innen bisher Ansätze entwickelt, die Genauigkeit zugunsten von Geschwindigkeit opfern. Aber gibt es die Möglichkeit, Genauigkeit und Verarbeitungs-Geschwindigkeit in einem ausgewogeneren Ansatz zu vereinen?

In dem Bestreben, Präzision und Leistung zu optimieren, entwickelten Forscher aus den Gruppen von Matthew Robinson und Marco Mondelli am Institute of Science and Technology Austria (ISTA) einen Algorithmus, der Informationen aus der weltweit umfangreichsten Biobank mit beispielloser Genauigkeit und Geschwindigkeit extrahieren und analysieren kann. Letztendlich könnte ihre Methode, die hier anhand des genetisch-komplexen Merkmals der Körpergröße vorgestellt wird, die personalisierte Medizin im Bereich der Diagnostik voranbringen – und sogar die Forensik.

Körpergröße und algorithmische Innovation

Der Ansatz des Teams stützt sich auf das kürzlich etablierte mathematische Framework „Approximate Message Passing“ (AMP), zu dem Mondelli maßgeblich beigetragen hat. Ihre neue Methode mit dem Namen „Genomic Vector Approximate Message Passing“ oder gVAMP verbessert die Fähigkeit des Frameworks, komplexe Informationen aus dem vorliegenden Datensatz zu extrahieren. „Während andere Methoden dazu neigen, jeweils nur einen Ausschnitt zu analysieren, bevor sie die Ergebnisse kombinieren, funktioniert gVAMP als Methode der „gemeinsamen Schätzung“. Daher bietet es einen Gesamtüberblick über die Auswirkungen auf ein Merkmal im Kontext aller Varianten in massiven genetischen Datensätzen“, sagt ISTA-Doktorand Al Depope, der Erstautor der Studie. „Wir können von einer algorithmischen Innovation sprechen.“

Um ihre Methode zu entwickeln, wählte das Team die Körpergröße des Menschen, ein etabliertes Modell für die genetische Analyse komplexer Merkmale. „Die Untersuchung der Körpergröße ermöglichte es uns, die Grenzen der

rechnerischen Skalierbarkeit mit gVAMP sowohl hinsichtlich der Anzahl der Genomsequenzen als auch der Anzahl der beteiligten Varianten zu erforschen“, sagt Depope. Tatsächlich wird dieses Merkmal von beeindruckenden 17 Millionen Varianten beeinflusst, die das Team gleichzeitig in Hunderttausenden von Gesamt-Genomsequenzen anonymisierter Freiwilliger aus der UK Biobank, dem weltweit umfassendsten Datensatz mit Informationen zu Biologie, Gesundheit und Lebensstil, analysieren konnte.

„Besonders wichtig finde ich die Interpretierbarkeit unseres Algorithmus bei der Anwendung in der Biologie. Damit können wir nicht nur die Körpergröße von Menschen anhand ihrer DNA genauer als bisher vorhersagen, sondern auch die spezifischen DNA-Regionen identifizieren, die daran beteiligt sind“, sagt ISTA-Postdoc und Mitautor Jakub Bajzik.

Leistungsstärker als bestehende Methoden

Wenn gVAMP die Körpergröße des Menschen und den Beitrag einzelner genetischer Varianten vorhersagt, erstellt der Algorithmus diese Daten zum ersten Mal. Daher gibt es keine bereits vorhandenen Daten zur genetischen Grundlage der Körpergröße, anhand derer die Methode bewertet werden könnte. „Im Wesentlichen lautet die Frage hier: ‚Woher wissen wir, dass gVAMP die richtigen Varianten ausgewählt hat?‘“, erklärt Depope.

Um die Leistungsfähigkeit ihrer Methode zu bewerten, führten die ISTA-Forscher eine Datensimulation durch – ein gängiger Ansatz in diesem Bereich. Sie entwickelten ein künstliches Merkmal mit ungefähr derselben Anzahl genetischer Varianten wie die Körpergröße des Menschen und führten eine umfangreiche Simulationsstudie mit mehreren Datensätzen durch, wobei sie die Leistung des Algorithmus mit anderen Methoden verglichen. Auf diese Weise konnten sie zeigen, dass gVAMP bestehende Methoden sowohl in Bezug auf die Genauigkeit als auch auf die Verarbeitungszeit deutlich übertrifft.

„Unsere Methode erreicht eine Spitzenpräzision und ist gleichzeitig so effizient, dass sie eine echte gemeinsame Analyse riesiger genetischer Datensätze in nur wenigen Tagen durchführen kann. Dadurch können wir die zugrunde liegende Biologie aufdecken, die zuvor aufgrund des begrenzten Umfangs der Datenanalyse verborgen blieb“, sagt Depope. „Die algorithmische Innovation ist genau das, was diese Größenordnung der Analyse und die daraus resultierenden biologischen Erkenntnisse möglich macht.“

Von der personalisierten Medizin zur Forensik?

Diese interdisziplinäre Studie vereint Fachwissen aus den Bereichen Informationstheorie, Mathematik, Genomik und Softwareentwicklung. Bajziks Hintergrund in der Informatik ergänzte Depopes Schwerpunkt auf Theorie und Mathematik. Robinson, der sich auf modernste statistische Modelle für Genomdaten spezialisiert hat, betreute das Projekt gemeinsam mit Mondelli, der robuste Inferenzmethoden in der Informationstheorie entwickeln möchte, um datengesteuerte Herausforderungen in den Ingenieur- und Naturwissenschaften anzugehen.

Derzeit baut das Team auf dieser Arbeit auf, um sie auf personalisierte Medizin und diagnostische Anwendungen auszuweiten. Dazu könnten Prognosen zum Zeitpunkt des Ausbruchs einer Krankheit, zu ihrem Schweregrad und zum wahrscheinlichen Auftreten bestimmter Symptome gehören. Darüber hinaus versuchen sie, die Methode auf Protein- und epigenetische Daten auszuweiten, also Informationen, die nicht allein aus den Genomsequenzen hervorgehen.

Letztendlich könnte das Potenzial von gVAMP in personalisierten medizinischen Anwendungen auch Ärzt:innen dabei helfen, gezielte Patient:innen-Profile für klinische Studien auszuwählen. Laut Depope könnte die Methode jedoch sogar noch weitere Anwendungsmöglichkeiten finden. „Ich denke, unser Algorithmus könnte auch in der Forensik nützlich sein, um anhand der am Tatort gefundenen DNA die Größe eines oder einer Verdächtigen vorherzusagen“, sagt er.

--

Projektförderung:

Dieses Projekt wurde durch Mittel aus dem Lopez-Loreta-Preis, einem SNSF Eccellenza Grant (PCEGP3-181181), einem ERC Starting Grant (INF2, Projektnummer 101161364) und durch die Grundfinanzierung des ISTA unterstützt.

Die Hochleistungsrechner wurden von den Scientific Service Units (SSU) des ISTA mit Ressourcen des Scientific Computing (SciComp) unterstützt.

wissenschaftliche Ansprechpartner:

<https://ista.ac.at/de/forschung/robinson-gruppe/> Forschungsgruppe "Medizinische Genomik" am ISTA

<https://ista.ac.at/de/forschung/mondelli-gruppe/> Forschungsgruppe "Data Science, Machine Learning und Informationstheorie" am ISTA

Originalpublikation:

Al Depope, Jakub Bajzik, Marco Mondelli, and Matthew R. Robinson. 2026. Joint modelling of whole genome sequence data for human height via approximate message passing. Cell Genomics. DOI: 10.1016/j.xgen.2026.101162 / <https://doi.org/10.1016/j.xgen.2026.101162>

Institute of Science and Technology Austria

VeronikaOleksyn (Mitarbeiter in der Presse- und Öffentlichkeitsarbeit)

+43 664 88326665

veronika.oleksyn@ista.ac.at

News-ID: 1304277 • Views: 350 (Stand: 11.07.2026)

Link zur Pressemitteilung:

<https://www.openpr.de/news/1304277/Big-Data-und-Koerpergroesse-Algorithmische-Innovation-bei-genetischer-Analyse-komplexer-Merkmale-idw.html>