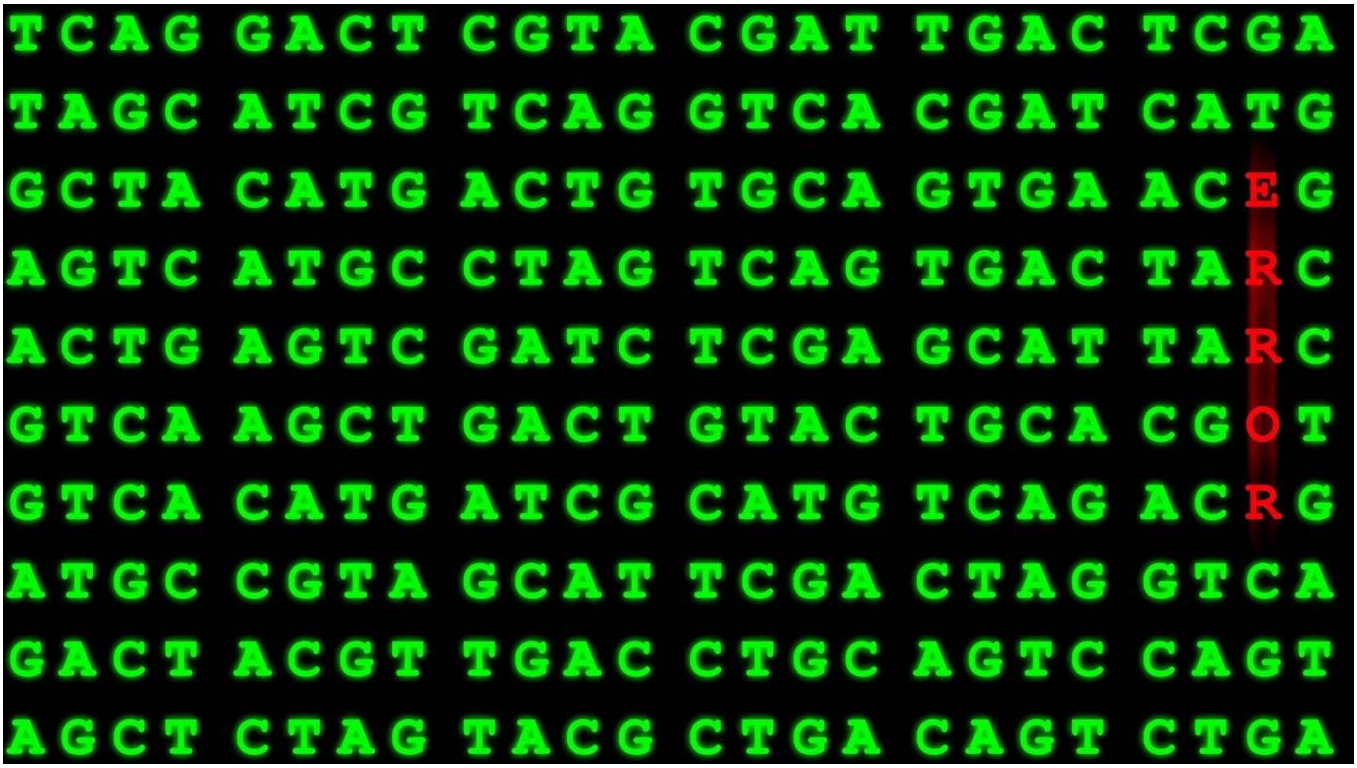
WISSEN BIG DATA

Genomforschung überholt YouTube in der Datenmenge

VON ANNETT STEIN UND SIMONE HUMML

vor 1 Tag



Die Analyse von Erbgutdaten ist so komplex wie die von astronomischen Daten zum Beispiel von Weltraumteleskopen

Foto: picture alliance / blickwinkel/M

DNA-Analysen werden immer billiger. Dabei fallen riesige Mengen an Daten an. US-Forscher sehen einen Engpass in der Infrastruktur. Europäische Kollegen wundern sich eher über den Fortschrittsglauben.

Immer mehr Menschen werden in den kommenden Jahren ihr Erbgut entziffern lassen – gewaltige neue Datenmengen werden die Folge sein. Um das Jahr 2025 herum werde sich die Genomik mit aktuellen Big-Data-Rekordhaltern wie Twitter, YouTube und astronomischer Forschung messen oder sie sogar übertrumpfen, sind Wissenschaftler der University of Illinois in Urbana-Champaign überzeugt.

Das eigene Erbgut zumindest teilweise entziffern zu lassen, ist unter anderem für die aufstrebende personalisierte Medizin bedeutsam. Zudem werden die Genomdaten von immer mehr Tieren, Pflanzen und anderen Organismen erfasst, gespeichert, weiterverbreitet und analysiert.

"Je besser und billiger die Techniken der Genomsequenzierung werden, umso mehr erwarten wir eine Explosion der Erbgutentzifferung, die eine gewaltige Datenflut zur Folge haben wird", erklärt Gene Robinson, einer der Autoren der Studie. In zehn Jahren werde es wohl Millionen sequenzierter

überhaupt handhaben zu können.

Doppelt so viele Genomdaten alle sieben Monate

Schon jetzt hätten die aus der Entzifferung des Erbguts etlicher Organismen gewonnenen Daten die Größenordnung von Petabyte – also Millionen Gigabyte – erreicht, schreiben die Forscher im Fachjournal "PLOS Biology". Weltweit gebe es mehr als 2500 Hochgeschwindigkeits-Geräte verschiedener Hersteller in fast 1000 Sequenzierungszentren. Im zurückliegenden Jahrzehnt habe sich die Menge gesammelter Genomdaten alle sieben Monate verdoppelt.

Künftig werde die Datenmenge noch weit schneller anwachsen, weil es immer verbreiteter wird, sich das eigene Genom entziffern zu lassen oder seine Daten großen Forschungsprojekten zur Verfügung zu stellen. In Großbritannien und Saudi-Arabien zum Beispiel gäbe es den Plan, das Erbgut von je 100.000 Bürgern erfassen zu lassen.

Eine wesentliche Voraussetzung für solche Projekte: Die stetig sinkenden Sequenzierungskosten. Nach einer Auswertung des staatlichen Forschungsinstituts NHGRI in den USA seien die Kosten für das Entziffern eines Genoms vor allem zwischen 2007 und 2012 erheblich gefallen.

Bis 2015 mehr Daten aus Genomen als in YouTube

Bis 2025 könnten weltweit – theoretisch – bis zu zwei Milliarden menschliche Genome sequenziert sein, heißt es in der Studie. Zudem gebe es etliche bereits begonnene oder vor dem Start stehende Großprojekte zur genetischen Entzifferung wichtiger Energie- und Agrarpflanzen in all ihren Varianten. Die US-Forscher schätzen, dass bis 2025 Exabyte an Genomik-Daten vorliegen – kaum vorstellbare Milliarden Gigabyte.

Neue Methode entschlüsselt Erbgut von Ungeborenen

Damit werde der derzeitige Rekordhalter für das größte gespeicherte Datenvolumen, YouTube, übertrumpft, sind die Forscher überzeugt. "Und die Sequenzen allein sind nur ein Element der Genomik." Das entzifferte Erbgut müsse auch noch analysiert werden, um biologische oder medizinische Schlüsse zu erlauben. Die Algorithmen dafür benötigten oft sehr hohe Rechenleistungen.

Zum Vergleich: Am ASKAP-Teleskop in Australien würden derzeit etwa 7,5 Terabyte Bilddaten pro Sekunde erfasst – bis 2025 könnten es allein an dieser Anlage 100-fach mehr sein. Bei YouTube würden derzeit minütlich Videos mit einer Gesamtlänge von etwa 300 Stunden hochgeladen. Bis 2025 könnten es 1000 bis 1700 Stunden sein, rechnen die Forscher aus der bisherigen Entwicklung hoch. Nicht ganz so immens sei der zu erwartende Zuwachs bei Twitter: 500 Millionen Tweets seien es derzeit am Tag, mit etwa 15 Milliarden könnten es 2025 etwa 30 Mal mehr sein.

Viele Datenformate erschweren Genom-Auswertung

Mit den derzeitigen Big-Data-Playern gebe es viele Gemeinsamkeiten, aber auch immense Unterschiede, erklären die Wissenschaftler. Ähnlich wie bei YouTube und Twitter würden auch Erbgutdaten weit verteilt generiert und stammten aus vielen verschiedenen Quellen. Die Online-Plattformen nutzten jedoch zumindest einheitliche Standardformate für das einlaufende Material. Für Genomdaten hingegen würden derzeit verschiedene Formate genutzt, was das Teilen und die Lagerung komplexer mache.

Die Analyse von Erbgutdaten sei ähnlich komplex wie die astronomischer Daten etwa von Weltraumteleskopen, schreiben die Forscher weiter. Auch hier gebe es aber einen wichtigen Unterschied: Schon beim Sammeln der Astro-Daten griffen spezielle Bearbeitungsprozesse, so dass die folgende Auswertung weniger Zeit und weniger Rechenleistung in Anspruch nehme. Ähnliche Techniken seien zwar auch für Genomdaten denkbar – dabei drohe aber eine Falle: Möglicherweise sind im gesamten Erbgut derzeit noch unbekannte Zusammenhänge verborgen, auf die sich mit einer abgespeckten Datenversion nicht mehr schließen lässt.

Speicherung nur von relevanten Daten

Das Erbgut einer menschlichen Zelle ist aus rund drei Milliarden Bausteinen zusammengesetzt, die bei den verschiedenartigen Analysen bis zu 30-fach erfasst werden, um die Qualität der Ergebnisse abzusichern. Derzeit bedeutet das knapp drei Terabyte Rohdaten pro Sequenz. "In Zukunft stehen wir vielleicht vor der schwierigen Entscheidung, eine bearbeitete Form statt des Originals zu speichern, eine extrem komprimierte Version, um den Speicherbedarf drastisch reduzieren zu können", erklärt Koautor Saurabh Sinha. Es sei dringend nötig, die rechnerbasierten Herausforderungen anzugehen, die die Genomik mit sich bringe.

"Der Grundgedanke der Studie ist richtig – dass das Datenvolumen zunehmen wird und auch die Bedeutung der Daten", meint Christof Kalle vom Deutschen Krebsforschungszentrum in Heidelberg. Auch die Zahl der Anwendungen werde steigen, ergänzte er etwa mit Blick auf die Eiweiß-Daten der Zellen. Die Datenkompression werde sich jedoch zumindest für den klinischen Bereich stärker als von den US-Autoren angenommen erhöhen. Kalle hält daher eine internationale Zusammenarbeit für die Entwicklung intelligenter und aufeinander abgestimmter Speicher- und Kompressionsverfahren für

Speichern mit unterschiedlichem Tempo

"Wir haben beim DKFZ ein Speichervolumen von mehr als zehn Petabyte, und wir werden sicherlich eine weitere Zunahme dieser Aktivität sehen", sagte Kalle. "Es gibt aber eine Tendenz, die Daten je nach Dringlichkeit auf anderen Speichermedien abzulegen. Das bedeutet, sie werden sicher nicht mehr alle in schnell zugänglichen Bereichen gespeichert." Dort blieben die medizinisch bedeutenden Daten.

"Allein am Deutschen Krebsforschungszentrum in Heidelberg sind wir in der Lage, jedes Jahr das Genom von 4500 Patienten zu erfassen", ergänzt Roland Eils, Leiter der Arbeitsgruppe Theoretische Bioinformatik am DKFZ. "Dabei wenden wir eine besonders intensive Sequenzierung an, bei der jede DNA-Base statistisch 120 Mal sequenziert wird." So würden auch Heterogenitäten eines Tumors erfasst. Künftig könne vielen Patienten so eine wesentlich detailliertere genetische Analyse ihrer Erkrankung angeboten und Hinweise auf mögliche zielgerichtete Therapien gefunden werden, ist Eils überzeugt. Schon jetzt würden zehn Terabyte an Genomdaten täglich erzeugt. "Das entspricht exakt dem derzeitigen Datendurchsatz von Twitter."

Fortschrittsgläubig oder ethische Bedenken?

Die Zahlen der US-Forscher hält Eils insgesamt für nachvollziehbar und realistisch. "Was mich an dem Artikel jedoch erstaunt ist, dass hier fortschrittsgläubig auf die erwartete Datenexplosion in Genomics geschaut wird, ohne dass auch nur ein Wort zu den ethischen und rechtlichen Aspekten gesagt wird, wenn in zehn Jahren potenziell jeder Vierte in den entwickelten Ländern sequenziert werden wird." In den USA werde der genomischen Datenexplosion fast ohne Einschränkung mit großer Begeisterung entgegengesehen – in Europa hingegen gebe es deutliche rechtliche und ethische Vorbehalte.

Andreas Keller vom Lehrstuhl für Klinische Bioinformatik der Universität des Saarlandes in Saarbrücken sieht insbesondere die von den US Forschern genannte Obergrenze von zwei Milliarden sequenzierter humaner Genome in den nächsten 10 Jahren skeptisch.

Alte Prognosen bis heute nicht eingetroffen

"Damit sich die Prophezeiungen erfüllen, müssten sich technologische Entwicklung genauer voraussagen lassen, auch was die Kosten und die Durchsatzkapazitäten zukünftiger Geräte angeht", erklärt er. "Inwieweit sich die prognostizierten Entwicklungen in den nächsten zehn Jahren umsetzen, ist sehr schwer abzuschätzen. Bei einem sich so rasant entwickelnden Gebiet ist das selbst für Experten fast unmöglich."

Es müsse sich auch noch weiter zeigen, wie groß der klinische Nutzen der Genomsequenzierung letztlich wirklich sei, ergänzt Keller. Nötig seien entsprechend ausgebildete Ärzte, um in einem solchen Szenario so vielen Patienten die Folgen und die Bedeutung genetischer Tests zu erklären. Es sei auch abzuwarten, ob jeweils ein gesamtes Genom entschlüsselt werden müsse oder es in der Mehrzahl der Fälle reiche, sich gezielt krankheitsrelevante Gene anzuschauen. Und nicht zuletzt zeige ein Blick ein gutes Jahrzehnt zurück, dass damals erstellte Prognosen zur Genomforschung bis heute nicht erreicht sind. "Ob es tatsächlich so kommt wie nun prognostiziert, wissen wir dann in zehn Jahren."

dpa