Datenbanktechnologie beflügelt personalisierte Medizin

Dr. Matthieu-P. Schapranow, Leiter des Themengebiets „Lebenswissenschaften" am Fachbereich Enterprise Platform and Integration Concepts, Hasso-Plattner-Institut;
Prof. Dr. Christoph Meinel, Direktor und Geschäftsführer des Hasso-Plattner-Instituts und Leiter des Fachbereichs „Internet-Technologien und -Systeme", Hasso-Plattner-Institut;
Prof. Dr. h.c. mult. Hasso Plattner, Leiter des Fachbereichs „Enterprise Platform and Integration Concepts", Hasso-Plattner-Institut

Genomdaten fließen mehr und mehr in die Diagnostik der personalisierten Medizin mit ein. Dabei ist der Weg bis zu den endgültigen Analyse-Ergebnissen heute aufgrund der Vielzahl unterschiedlicher Prozessschritte, der schieren Datenmengen und der unterschiedlichen Formate mitunter steinig. Hier kann eine in Potsdam erforschte Datenbanktechnologie Abhilfe schaffen und das Analysieren riesiger Mengen medizinischer Daten in Echtzeit ermöglichen. Die In-Memory-Technologie ist dabei das Herzstück künftiger Entscheidungshilfen für die Anwendung individueller Therapien, zum Beispiel bei der Therapie von Krebserkrankungen.

Die ursprünglich für Unternehmenssoftware entwickelte und mit dem Deutschen Innovationspreis 2012 ausgezeichnete In-Memory-Technologie kann helfen, Genomdaten in Echtzeit zu analysieren und auszuwerten. Komplizierte und teure Krebstherapien können so im Rahmen der personalisierten Medizin künftig schneller und passender auf jeden Patienten individuell zugeschnitten werden. Die durch das Potsdamer Hasso-Plattner-Institut (HPI) auf dem diesjährigen World Health Summit vorgestellten Forschungsergebnisse des „High-Performance In-Memory Genome"-Projekts (HIG) sind vielversprechend. Bei der personalisierten Medizin steht der Patient mehr denn je im Vordergrund. Neben der Symptomatik fließen auch die persönlichen Umwelteinflüsse und die individuellen Dispositionen beim Finden einer zielgerichteten, individuellen Behandlungsentscheidung, etwa bei der Behandlung spezieller Tumore, mit ein. Jedoch stellen die dabei pro Patient entstehenden diagnostischen Datenmengen, zum Beispiel im Rahmen einer Genomsequenzierung, Computersysteme vor Herausforderungen. Beispielsweise fallen bei der Genomsequenzierung mehrere Gigabyte Daten je Patient an. Die am HPI erforschte In-Memory-Datenbanktechnologie unterstützt die Echtzeit-Analyse dieser riesigen Datenmengen.

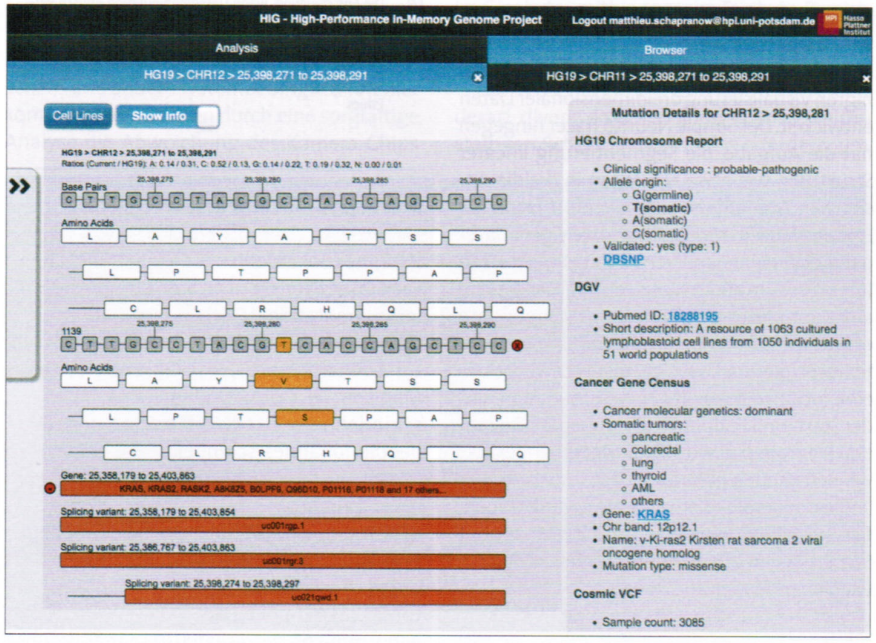

Abb. 1: Der Genom-Browser der Cloud-Anwendung des Hasso-Plattner-Instituts zur Echtzeit-Analyse von Genomdaten im Rahmen der personalisierten Medizin. Auf der linken Seite sind das Referenzgenom und das Patientengenom gegenüber- und eine ausgewählte Mutation auf Ebene der DNA dargestellt. Auf der rechten Seite sind die Ergebnisse der Forschungsdatenbanken aus aller Welt zur spezifischen Mutation dargestellt.

Umfassende Wissensdatenbank

Um genetische Veränderungen in Echtzeit zu analysieren, werden aufbauend auf der In-Memory-Technologie weltweite medizinische Forschungsergebnisse in einer riesigen Wissensdatenbank kombiniert. Durch die Verbindung von Hochleistungsrechnern mit riesigen Arbeitsspeichern können so Erkenntnisse aus der Forschung direkt in kurative Behandlungsentscheidungen einfließen. Durch die Verwendung moderner Next Generation Sequencing (NGS)-Geräte können Gewebeproben, z.B. von Krebstumoren, schon heute binnen Stunden, statt wie noch vor einem Jahrzehnt über Wochen hinweg, sequenziert werden. Als Vorbereitung wird dazu eine Gewebeprobe aus dem Tumor, etwa mittels einer Biopsie, entnommen. Die aus dem Tumorgewebe gewonnen Zellen werden in ausreichender Quantität im Labor herangezogen und für das Sequenzierungsverfahren vorbereitet. Am Ende des herstellerspezifischen Verfahrens liegen die unbearbeiteten Sequenzierungsdaten in digitaler Form vor, zum Beispiel als FASTQ-Dateien in Größen von 50 GB und mehr.

Nun schließen sich IT-gestützte Verarbeitungsschritte der Rohdaten an. Nach der Rekonstruktion des ursprünglichen Genoms – Alignment genannt –, folgen die Identifizierung von spezifischen Variantenausprägungen, das Auffinden möglicher Mutationsloci, sowie der Abgleich mit weltweiten Forschungsdatenbanken. Dieser Schritt ist heute sehr zeitaufwendig und erfordert oft lange Recherchen bei weltweiten Forschungsinstituten.

Die Rekonstruktion des Ursprungsgenoms aus einer Vielzahl kurzer Leseabschnitte – Reads genannt – stellt dabei ein kombinatorisches Problem dar. Die Ausführungszeit heutiger Alignment-Algorithmen, wie BWA, Bowtie, oder Bowtie2, werden dabei durch die zur Verfügung stehende Rechenkapazität bestimmt. Im Gegensatz dazu werden das Identifizieren spezifischer Varianten und ihre Annotation durch die verwendete Hauptspeicherkapazität beschränkt. Beispielsweise erfordert das „Variant Calling" den Abgleich der spezifischen Variantenausprägungen eines jeden Patienten mit den ca. 80 Millionen weltweit bekannten Mutationen.

Die Forscher am HPI rücken diesen riesigen Datenbergen mit einem Hochleistungsrechenverbund bestehend aus 1.000 Kernen zu Leibe, einem von weltweit drei Exemplaren dieser Art. Dieser Rechenverbund besteht aus 25 Knoten und hat dadurch zwei einmalige Eigenschaften.

Zum Einen besteht er aus 1.000 individuellen Rechenkernen, die alle gleichzeitig je ein kleines Puzzleteil des gesamten Analyseprozesses rasant bearbeiten können. Gerade die Rekonstruktion des Genoms profitiert von dieser parallelen Datenverarbeitung. Statt wie bisher lange auf ein Gesamtergebnis warten zu müssen, können so bereits frühzeitig Teilergebnisse während der Rekonstruktion bearbeitet werden. Der zweite Vorteil ist der riesige Hauptspeicher des Verbunds. Jeder Knoten verfügt über einen Terabyte (TB) Hauptspeicherkapazität, die im Verbund zu riesigen 25 TB kombiniert werden.

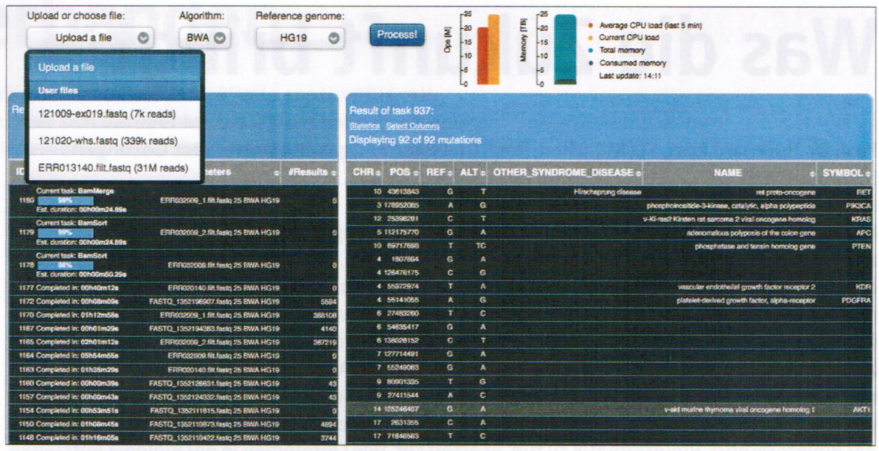

Abb. 2: In der Analyse-Ansicht werden die FASTQ-Datei, der zu verwendende Analyseprozess, sowie das Referenz-Genom ausgewählt und der Prozess gestartet. Die Liste auf der linken Seite zeigt die jüngsten Analysen und deren Fortschritt; in der Liste auf der rechten Seite können frei wählbare Attribute aus weltweiten Forschungsdatenbanken zu spezifischen Mutationsloci interaktiv angezeigt werden.

Ergebnisse in Wimpernschlagdistanz

Die technischen Grundlagen kombiniert mit geeigneten Werkzeugen der In-Memory-Technologie bilden die Basis für die Echtzeit-Analyse von Genomdaten. Zum Beispiel können durch geeignete Anwendung der leichtgewichtigen Kompression und der spaltenorientierten Datenablage tausende Genome, Millionen bekannter Mutationen und unzählige Annotationen gleichzeitig im Hauptspeicher im schnellen Zugriff vorgehalten werden. Neben der Reduktion des Speicherplatzbedarfs, kann so auch ein Vielfaches der unkomprimierten Datenmenge gleichzeitig durch die limitierten Datenbusleitungen zwischen Prozessorkern und Hauptspeicher transportiert und verarbeitet werden. Dies erhöht den Durchsatz und die Anzahl gleichzeitiger bearbeitbarer Aufgaben. Nur so wird es möglich, die etwa 80 Millionen bekannten Mutationen auf dem menschlichen Genom und unzählige Erkenntnisse aus weltweit verteilten Forschungsprojekten während eines Wimpernschlages zu durchsuchen und die relevanten Ergebnisse zu identifizieren.

NCBI, UCSC etc.: Alles in einer Wolke

Durch die Kombination weltweit über das Internet verfügbarer Datenquellen können erstmals auch behandlungsrelevante Zusatzinformationen, wie assoziierte Krankheiten, pharmakologische Zusammenhänge oder geeignete klinische Studien, binnen Sekunden in die Auswertung einfließen. Dadurch erhalten medizinische Nutzer der Cloud-Anwendung einen ganzheitlichen Überblick über individuelle Dispositionen eines Patienten, wobei stets die weltweit aktuellsten Forschungsergebnisse in die Behandlung mit einfließen können. Dazu integriert die Cloud-Anwendung des HPIs eine Vielzahl öffentlich verfügbarer medizinischer Forschungsergebnisse. Darunter sind unter anderem Daten etablierter Konsortien, wie des National Center for Biotechnology Information (NCBI) oder der University of California, Santa Cruz (UCSC). Aber auch Erkenntnisse spezifischer Forschungsprojekte, z.B. der Database of Genomic Variants, fließen bereits in die Analyse mit ein. Nicht-öffentliche Forschungsergebnisse können bei der Analyse ebenfalls zur Anwendung kommen. Die Vertraulichkeit dieser Erkenntnisse wird durch Zugriffskontrollen und Benutzergruppenberechtigungen gewahrt.

Aber gerade diese Kombination der weltweiten Datenquellen in Echtzeit erfordert viel technische Raffinesse. Zum Beispiel stehen wissenschaftliche Erkenntnisse in einer Vielzahl unterschiedlicher Datenformate zur Verfügung. Statt alle Quellen in ein einheitliches Format zu überführen, bevor sie genutzt werden können, hilft auch hier die In-Memory-Technologie. Durch die Verwendung sogenannter „Views" können zusätzliche virtuelle Attribute der Datenbasis hinzugefügt werden. Dabei bestehen „Views" aus Transformationsregeln, die im Moment des Zugriffs ausgewertet werden. So können die ursprünglichen Datenformate unverändert beibehalten und nur die verhältnismäßig kleine Menge der ergebnisrelevanten Daten vor der eigentlichen Anzeige abgeleitet werden. Da die Datentransformation direkt durch die Datenbank durchgeführt wird, muss der Anwender hierbei keine zusätzlichen Schritte durchführen. Die verwendeten „Views" ermöglichen den direkten Zugriff auf neueste Erkenntnisse ohne Verzögerung durch langwierige Vorverarbeitung der Daten.

Dank der HPI-Technologie dauert die Genomdatenanalyse nur noch wenige Sekunden. Werden dabei krankheitsrelevante Mutationen entdeckt, erspart die Technologie dem Onkologen viele umständliche Einzelabfragen in einzelnen, spezialisierten Datenbanken. Stattdessen werden die Resultate automatisch ihrer Relevanz nach im Vergleich mit international bekannten Forschungsergebnissen sortiert angezeigt. Behandelnde Mediziner bekommen im Genom-Browser des HIG-Projekts entscheidende Zusatzinformationen zu jeder Mutation angezeigt, wie etwa deren Häufigkeit oder weltweit bekannte Fälle mit ähnlicher Historie, die eine spezifische Behandlung ermöglichen.

Aber nicht nur Exzellenzzentren mit hochwertiger Ausstattung sondern auch Krankenhäuser und Forschungsinstitute in abgelegenen Regionen oder mit geringeren Ressourcen profitieren von den Ergebnissen des HIG-Projekts am HPI. Letztere erhalten die Möglichkeit, Genom-Analysen in ihre eigene personalisierte Medizin zu integrieren, ohne NGS-Geräte, IT-Experten oder Bioinformatiker selbst vor Ort haben zu müssen. So können Sie mit Hilfe der Cloud-Anwendung des HPIs die Rohdaten aus in Auftrag gegebenen Sequenzierungsaufträgen selbst in Echtzeit analysieren und auswerten. Dabei unterstützt die Lösung des HPIs nicht nur die erforderlichen Hardware-Ressourcen und das Know-how, sondern auch durch eine einmalige Kombination weltweiter Forschungserkenntnisse, auf die sogar weltweit zugegriffen werden kann.

Derzeit ist die Nutzung des Systems noch auf ausgewählte Nutzergruppen beschränkt. Interessenten können sich aber bereits heute online informieren[1].

Literatur

[1] http://epic.hpi.uni-potsdam.de/Home/HigProject

Korrespondenzadresse

Dr.-Ing. Matthieu-P. Schapranow
Lehrstuhl Prof. Dr. Hasso Plattner, Hasso-Plattner-Institut für Softwaresystemtechnik GmbH
August-Bebel-Str. 88, 14482 Potsdam
matthieu.schapranow@hpi.uni-potsdam.de