

Digitale Werkzeuge für effektive Viren-Forschung

Die digitale Infrastruktur „Serratus“ ermöglicht Forschenden, öffentliche Sequenzdatenbanken effektiv nach biologischen Viren zu durchsuchen. Bislang konnten über 130.000 neue RNA-Viren identifiziert werden - von Corona-Viren über Verwandte des Hepatitis-D-Virus bis zu Bakteriophagen. Das internationale Team hinter dem Projekt, an dem auch Forschende des Heidelberger Instituts für Theoretische Studien und des Max-Planck-Instituts für Biologie beteiligt sind, berichtet über die Ergebnisse im Fachjournal „Nature.“

Die Vielfalt der Viren auf unserem Planeten ist sprichwörtlich unfassbar, denn die Wissenschaft kennt bislang nur einen Bruchteil der existierenden Viren. Welch verheerende Folgen neu auftretende Viruserkrankungen für die Menschheit haben, hat die derzeitige SARS-CoV2-Pandemie gezeigt. Daher ist es wichtig, die Diversität der global vorkommenden Viren mit Mitteln der Informatik zu katalogisieren und für die Wissenschaft nutzbar zu machen.

Zufallsfunde im Regenwald

Öffentliche Sequenzdatenbanken sind zu einem riesigen Speicher für genetische Daten geworden, den Forschende aus aller Welt befüllen. Diese Daten stammen von biologischen Forschungsgruppen, die Sequenzdaten erzeugen, sei es zur Untersuchung des Bodenmikrobioms des Amazonas-Regenwaldes oder zur Erforschung der Ausbreitung von Krankheiten wie dem SARS-CoV-2-Virus. In der Regel werden bei solchen Studien genetische Sequenzdaten nicht nur von dem Organismus gewonnen, der untersucht werden sollte, sondern auch von anderen Organismen, deren DNA zufällig in der Probe enthalten ist. Solche zufälligen Daten können für andere Forschende besonders interessant sein, da diese Daten nicht im Mittelpunkt der ursprünglichen Studie stehen und daher in der Regel ignoriert werden. Sie sind aber dennoch in den öffentlichen Datenbanken hinterlegt.

Eine Infrastruktur für effiziente Suche

Diesen verborgenen Schatz zu heben bedeutet, dass die Forschenden in ungeheuer großen und verteilten Datenmengen suchen müssten. Denn in den frei zugänglichen öffentlichen Datenbanken liegen Sequenzdaten in der Größenordnung von Petabytes (d.h. Millionen von Gigabytes). Die Forschenden im internationalen Serratus-Projekt haben hierfür eine Cloud-basierte Infrastruktur entwickelt. Serratus ist eine open source Cloud-Computing-Infrastruktur, die den Sequenzabgleich im Petabyte-Maßstab ermöglicht.

„Unsere Infrastruktur ermöglicht eine effiziente Suche im Sequence Read Archive, einem der beliebtesten öffentlichen Sequenzspeicher“, erläutert Pierre Barbera, der als Mitglied der Computational Molecular Evolution Gruppe am Heidelberger Institut für Theoretische Studien (HITS) Ko-Autor der Studie war. Er erstellte Software zur Berechnung und Analyse der phylogenetischen Stammbäume aller untersuchten Spezies. Am Projekt beteiligt sind auch Forschende am Max-Planck-Institut für Biologie in Tübingen. Sie brachten ihre Biocomputing-Software „DIAMOND“ in das Projekt ein, die wie eine Internet-Suchmaschine in wenigen Stunden Übereinstimmungen von Proteinbausteinen sequenzierter Lebewesen auflistet. Bis vor kurzem war für solche Berechnungen selbst mit Hochleistungsrechnern und dem bisherigen Goldstandard BLAST noch ein Zeitraum von Monaten notwendig. Die erweiterte Version „DIAMOND v2“ wird in

Zusammenarbeit mit der Max Planck Computing and Data Facility in Garching entwickelt.

Zahl der neu entdeckten Viren verzehnfacht

Mit den entwickelten Werkzeugen konnten die Forscher über 130.000 neue RNA-Viren identifizieren, was eine Verzehnfachung der bekannten Virenspezies bedeutet. Darunter befanden sich bisher unbekannte Mitglieder der Coronavirus-Familie, die eng mit dem SARS-CoV-2-Virus verwandt sind, sowie neuartige Viren, die mit dem Hepatitis-D-Virus verwandt sind, und neuartige Bakteriophagen, d. h. Viren, die speziell gegen Bakterien gerichtet sind.

Neben den beiden deutschen Teams waren Forschende vom Institut Pasteur (Paris, Frankreich), der Universität St. Petersburg (Russland), der Universität Valencia, der University of British Columbia (Kanada) und der UC Berkeley (USA) an der Studie beteiligt. Erstautor ist der Bioinformatiker Artem Babaian (University of Cambridge, Großbritannien).

Die Ergebnisse wurden jetzt im Fachjournal „Nature“ veröffentlicht. Die Daten aus dem Projekt sind öffentlich zugänglich und finden sich auch auf der Website <https://serratus.io>, so dass Forschende jederzeit darauf zugreifen und sie weiter untersuchen können.

Titel der Publikation:

Edgar, R.C., Taylor, J., Lin, V. et al. Petabase-scale sequence alignment catalyses viral discovery. Nature, 26 January 2022.

DOI: 10.1038/s41586-021-04332-2 / <https://www.nature.com/articles/s41586-021-04332-2>

Weitere Informationen:

<https://www.h-its.org/de/2022/01/27/serratus-nature/> HITS-Pressemitteilung

<https://serratus.io/> Serratus-Website

<https://github.com/bbuchfink/diamond> DIAMOND-Software