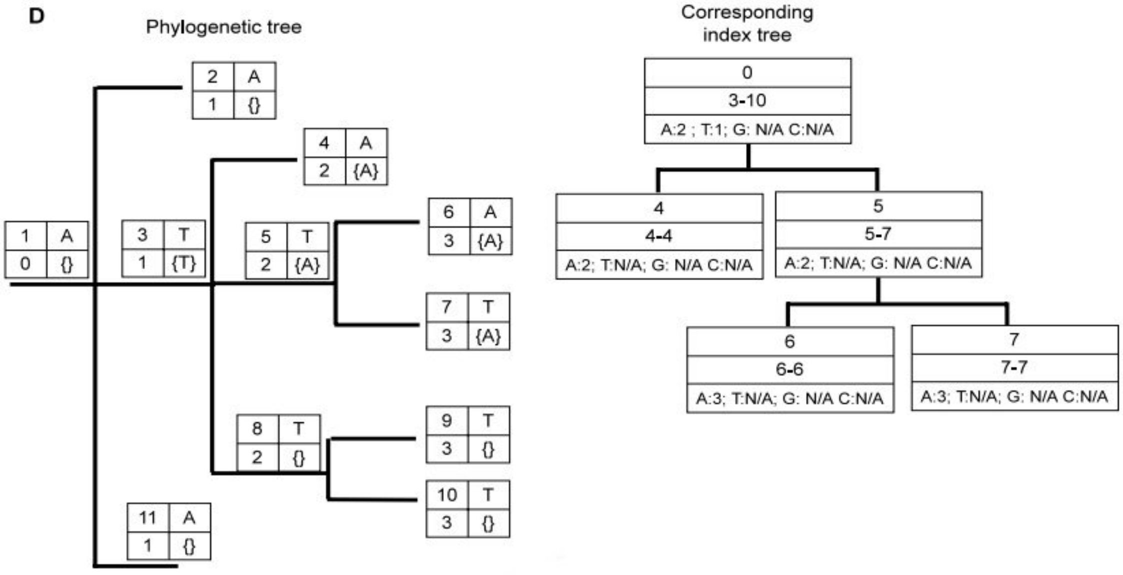
Medical Xpress

New phylogenetic tool can handle the SARS-CoV-2 data load

June 23 2022, by Kiran Kumar and Katherine Connor



An example phylogenetic tree (left) and its corresponding index tree. Credit: University of California - San Diego

Researchers at UC San Diego, in collaboration with UC Santa Cruz, have developed a new software tool for tracing and mapping the evolution of the SARS-CoV-2 virus, that is capable of handling the unprecedented amount of genetic data being generated by the quickly evolving pathogen. The software is used to efficiently and accurately track new variants of this virus on what's known as a phylogenetic tree: a visual history or map of an organism's genetic changes and variations over time and geography. Using this new optimization tool, called matOptimize, researchers are now able to more accurately track the viral genome of SARS-CoV-2, mapping new variants onto the phylogenetic tree as they develop, and tracking the evolutionary and transmission dynamics of the virus.

The tool was described in the journal *Bioinformatics*, with UC San Diego undergraduate computer engineering student Cheng Ye as first author. Hear more about Ye's journey to research as an undergraduate, and his experience working on such a timely project, in this Q&A.

"With over 10 million SARS-CoV-2 genome sequences now available, maintaining an accurate, comprehensive phylogenetic tree of all available SARS-CoV-2 sequences is becoming computationally infeasible with existing software, but is essential for getting a detailed picture of the virus' evolution and transmission," the researchers, under the direction of UC San Diego Electrical and Computer Engineering Professor Yatish Turakhia, write in the paper.

Currently, the program used for SARS-CoV-2 phylogeny is called UShER: Ultrafast Sample placement on Existing tRee. UShER was developed by Turakhia as a postdoctoral researcher at UC Santa Cruz, and is used by UC Santa Cruz to maintain the SARS-CoV-2 phylogeny. It is publicly viewable at—https://taxonium.org/?backend=https://api.cov2tree.org.

A few months into the pandemic, UShER faced a challenge with adding new genetic sequences onto the tree; the team would add sequences step-wise, one at a time, but when the genetic sequence input was incorrect or ambiguous, the system would lose accuracy.

"UShER would make a guess: an educated guess, but still a guess," said

Turakhia.

Thus, these sequences would occasionally be sub-optimally placed on the tree, producing false mutations. In order to refine these placements, a tree optimizing method was needed. However, existing tree optimizers were unable to keep up with the amount of SARS-CoV-2 genetic data being generated, with currently 10 million sequences mapped and up to 100,000 sequences added daily.

That's when Turakhia worked with Ye and other students in his lab on the challenge of creating a better tree optimizer. Ye had joined Turakhia's lab through the Electrical and Computer Engineering Summer Research Internship Program (SRIP) in January 2021. When it became clear to Turakhia that Ye's fundamentals in data structures, parallel algorithms, programming, and bioinformatics were quite strong, he entrusted him with taking a leading role on this task.

"I was initially assigned to work on accelerating sequence alignment on graphic processing units, but I thought the SARS-COV-2 phylogeny project might be more exciting, and it indeed was," said Ye.

"In those days [Cheng] became an expert in tree-optimization," said Turakhia.

Many of the existing tree optimizers were closed source, so Ye was forced to work with what was available in the literature to devise a solution to the data challenge. After a few months of research, Ye developed matOptimize, currently the only tool capable of keeping up with the amount of rapidly evolving SARS-CoV-2 genetic data.

In order to achieve this, Ye created a true parallel software, with processing distributed over several CPUs, and a significantly lower memory requirement. This allows it to be scaled to the level of data required in the SARS-CoV-2 phylogeny.

Today, UShER as the phylogenetic tree software and matOptimize as the tree optimization method, are being used together to characterize the SARS-CoV-2 phylogeny. There is now an entire catalog of genetic sequences which, from phylogenetic inferences, are highlighted as more dangerous or transmissible sequences which UC San Diego and UC Santa Cruz scientists continue to track.

Moving forward, Turakhia's team is using this information to study the recombination of SARS-CoV-2, a phenomenon that may lead to newer, dangerous variants.

"In collaboration with Professor Russell Corbett-Detig's group at UC Santa Cruz, Cheng and I developed a software called RIPPLES, that can sensitively detect recombinants in 1000x larger datasets," said Turakhia. "This software will help monitor the emergence of new SARS-CoV-2 recombinants and is likely to be applied to other pathogens as well in the future."

More information: Cheng Ye et al, matOptimize: A parallel tree optimization method enables online phylogenetics for SARS-CoV-2, *Bioinformatics* (2022). DOI: 10.1093/bioinformatics/btac401

Provided by University of California - San Diego

Citation: New phylogenetic tool can handle the SARS-CoV-2 data load (2022, June 23) retrieved 7 April 2023 from https://medicalxpress.com/news/2022-06-phylogenetic-tool-sars-cov-.html