PHYS.ORG

Speeding up sequence alignment across the tree of life

April 12 2021

Credit: CC0 Public Domain

A team of researchers from the Max Planck Institutes of Developmental Biology in Tübingen and the Max Planck Computing and Data Facility in Garching develops new search capabilities that will allow to compare the biochemical makeup of different species from across the tree of life. Its combination of accuracy and speed is hitherto unrivaled.

Humans share many sequences of nucleotides that make up our genes with other species—with pigs in particular, but also with mice and even bananas. Accordingly, some proteins in our bodies—strings of amino acids assembled according to the blueprint of the genes—can also be the same as (or similar to) some proteins in other species. These similarities might sometimes indicate that two species have a common ancestry, or they may simply come about if the evolutionary need for a certain feature or molecular function happens to arise in the two species.

Beating the gold standard of comparative genomics research

But of course, finding out what you share with a pig or a banana can be a monumental task; the search of a database with all the information about you, the pig, and the banana is computationally quite involved. Researchers are expecting that the genomes of more than 1.5 million eukaryotic species—that includes all animals, plants, and mushrooms—will be sequenced within the next decade. "Even now, with only hundreds of thousand genomes available (mostly representing small genomes of bacteria and viruses), we are already looking at databases with up to 370 million sequences. Most current search tools would simply be impracticable and take too long to analyze data of the magnitude that we are expecting in the near future," explains Hajk-Georg Drost, Computational Biology group leader in the Department of Molecular Biology of the Max Planck Institute of Developmental Biology in Tübingen.

"For a long time, the gold standard for this kind of analysis used to be a tool called BLAST," recalls Drost. "If you tried to trace how a protein was maintained by natural selection or how it developed in different phylogenetic lineages, BLAST gave you the best matches at this scale. But it is foreseeable that at some point the databases will grow too large for comprehensive BLAST searches."

Finding the needle in the haystack—but quickly!

At the core of the problem is a tradeoff between speed versus sensitivity: just like you will miss some small or well-hidden Easter eggs if you scan a room only briefly, speeding up the search for similarities of protein sequences in a database typically comes with downside of missing some of the less obvious matches.

"This is why some time ago, we started to devise the DIAMOND algorithm, in the hope that it would allow us to deal with large datasets in a reasonable amount of time," remembers Benjamin Buchfink, collaborator and Ph.D. student in Drost's research group who has been developing DIAMOND since 2013. "It did, but it also came with a downside: it couldn't pick up some of the more distant evolutionary relationships." That means that while the original DIAMOND may have been sensitive enough to detect a given human amino acid sequence in a chimpanzee, it may have been blind to the occurrence of a similar sequence in an evolutionary more remote species.

A powerful tool for future research

While being useful for studying material that was directly extracted from environmental samples, other research goals require more sensitive tools than the original DIAMOND search algorithm. The team of researchers from Tübingen and Garching was now able to modify and extend DIAMOND to make it as sensitive as BLAST while maintaining its superior speed: with the improved DIAMOND, researchers will be able to do comparative genomics research with the accuracy of BLAST at an 80- to 360-fold computational speedup. "In addition, DIAMOND enables researchers to perform alignments with BLAST-like sensitivity on a supercomputer, a high-performance computing cluster, or the Cloud in a truly massively parallel fashion, making extremely large-scale sequence alignments possible in tractable time," adds Klaus Reuter, collaborator from the Max Planck Computing and Data Facility."

Some queries that would have taken other tools two months on a supercomputer can be accomplished in several hours with the new DIAMOND infrastructure. "Considering the exponential growth of the number of available genomes, the speed and accuracy of DIAMOND are exactly what modern genomics will need to learn from the entire collection of all genomes rather than having to focus only on a smaller number of particular species due to a lack of sensitive search capacity," Drost predicts. The team is thus convinced that the full advantages of DIAMOND will become apparent in the years to come.

More information: Benjamin Buchfink et al, Sensitive protein alignments at tree-of-life scale using DIAMOND, *Nature Methods* (2021). DOI: 10.1038/s41592-021-01101-x

Provided by Max Planck Society

Citation: Speeding up sequence alignment across the tree of life (2021, April 12) retrieved 26 June 2024 from https://phys.org/news/2021-04-sequence-alignment-tree-life.html