PHYS.ORG

Massive data for miniscule communities

August 1 2012

It's relatively easy to collect massive amounts of data on microbes. But the files are so large that it takes days to simply transmit them to other researchers and months to analyze once they are received.

Researchers at Michigan State University have developed a new computational technique, featured in the current issue of the Proceedings of the National Academy of Sciences, that relieves the logjam that these "big data" issues create.

Microbial communities living in soil or the ocean are quite complicated. Their genomic data is easy enough to collect, but their data sets are so big that they actually overwhelm today's computers. C. Titus Brown, MSU assistant professor in bioinformatics, demonstrates a general technique that can be applied on most microbial communities.

The interesting twist is that the team created a solution using small computers, a novel approach considering most bioinformatics research focuses on supercomputers, Brown said.

"To thoroughly examine a gram of soil, we need to generate about 50 terabases of genomic sequence – about 1,000 times more data than generated for the initial human genome project," said Brown, who co-authored on the paper with Jim Tiedje, University Distinguished professor of microbiology and molecular genetics. "That would take about 50 laptops to store that much data. Our paper shows the way to make it work on a much smaller scale."

Analyzing DNA data using traditional computing methods is like trying to eat a large pizza in a single bite. The huge influx of data bogs down computers' memory and causes them to choke. The new method employs a filter that folds the pizza up compactly using a special data structure. This allows computers to nibble at slices of the data and eventually digest the entire sequence. This technique creates a 40-fold decrease in memory requirements, allowing scientists to plow through reams of data without using a supercomputer.

Brown and Tiedje will continue to pursue this line of research, and they are encouraging others to improve upon it as well. The researchers made the complete source code and the ancillary software available to the public to encourage extension.

"We want this program to continue to evolve and improve," Brown said. "In fact, it already has. Other researchers have taken our approach in a new direction and made a better genome assembler."

Provided by Michigan State University

Citation: Massive data for miniscule communities (2012, August 1) retrieved 9 April 2024 from https://phys.org/news/2012-08-massive-miniscule.html