Novel algorithms and computational techniques speed up genome assembly from months to minutes

July 1 2015

Human Chromosomes. Credit: Jane Ades, NHGRI

Genomes are like the biological owner's manual for all living things. Cells read DNA instantaneously, getting instructions necessary for an organism to grow, function and reproduce. But for humans, deciphering this "book of life" is significantly more difficult.

Nowadays, researchers typically rely on next-generation sequencers to translate the unique sequences of DNA bases (there are only four) into letters: A, G, C and T. While DNA strands can be billions of bases long, these machines produce very short reads, about 50 to 300 characters at a time. To extract meaning from these letters, scientists need to reconstruct portions of the genome—a process akin to rebuilding the sentences and paragraphs of a book from snippets of text.

But this process can quickly become complicated and time-consuming, especially because some genomes are enormous. For example, while the human genome contains about 3 billion bases, the wheat genome contains nearly 17 billion bases and the pine genome contains about 23 billion bases. Sometimes the sequencers will also introduce errors into the dataset, which need to be filtered out. And most of the time, the genomes need to be assembled de novo, or from scratch. Think of it like putting together a ten billion-piece jigsaw puzzle without a complete picture to reference.

By applying some novel algorithms, computational techniques and the innovative programming language Unified Parallel C (UPC) to the cutting-edge de novo genome assembly tool Meraculous, a team of scientists from the Lawrence Berkeley National Laboratory (Berkeley Lab)'s Computational Research Division (CRD), Joint Genome Institute (JGI) and UC Berkeley, simplified and sped up genome assembly, reducing a months-long process to mere minutes. This was primarily achieved by "parallelizing" the code to harness the processing power of supercomputers, such as the National Energy Research Scientific Computing Center's (NERSC's) Edison system. Put simply, parallelizing

code means splitting up tasks once executed one-by-one and modifying or rewriting the code to run on the many nodes (processor clusters) of a supercomputer all at once.

"Using the parallelized version of Meraculous, we can now assemble the entire human genome in about eight minutes using 15,360 computer processor cores. With this tool, we estimate that the output from the world's biomedical sequencing capacity could be assembled using just a portion of NERSC's Edison supercomputer," says Evangelos Georganas, a UC Berkeley graduate student who led the effort to parallelize Meraculous. He is also the lead author of a paper published and presented at the SC Conference in November 2014.

"This work has dramatically improved the speed of genome assembly," says Leonid Oliker computer scientist in CRD. "The new parallel algorithms enable assembly calculations to be performed rapidly, with near linear scaling over thousands of cores. Now genomics researchers can assemble large genomes like wheat and pine in minutes instead of months using several hundred nodes on NERSC's Edison."

Supercomputers: A game changer for assembly

High throughput and relatively low cost next-generation DNA sequencers have allowed researchers to look for biological solutions to everything from generating clean energy and environmental cleanup to identifying connections between genetic mutations and cancer. For the most part, these machines are very accurate at recording the sequence of DNA bases. But sometimes errors such as substitutions, repetitions, transpositions and omissions do occur—akin to "typos" in a book. These errors complicate analysis by making it harder to assemble genomes and identify genetic mutations. They can also lead researchers to misinterpret the function of a gene.

One technique that researchers often use to identify errors is called shotgun sequencing. This involves taking numerous copies of a DNA strand, breaking it up randomly into numerous smaller pieces and then sequencing each piece separately. This produces a number of overlapping short reads that allow scientists to eventually reassemble the whole DNA strand. Sequencing numerous copies of the same DNA strand also helps identify errors. But for a particularly complex genome, this process also generates a tremendous amount of data, sometimes several terabytes.

To identify errors in this data quickly and effectively, the Berkeley Lab and UC Berkeley team relied on "Bloom filters" and massively parallel supercomputers. Conceived by Burton H. Bloom in 1970, Bloom filters are very efficient at recognizing whether or not an element is a member of the set. Thus, researchers can rely on this tool to tell them if a base is out of place and is likely a mistake. Because bit arrays comprise a Bloom filter's underlying structure, they also require relatively little memory, making them ideal for querying massive datasets.

"Applying Bloom filters to this part of the genome assembly problem is not new, it has been done before. What we have done differently is to get Bloom filters to work with distributed memory systems," says Aydin Buluç, a research scientist in CRD. "This task was not trivial, it required some computing expertise to accomplish."

The team also developed solutions for parallelizing data input and output (I/O). "When you have several terabytes of data, just getting the computer to read your data and output results can be a huge bottleneck," says Steven Hofmeyr, a research scientist in CRD who developed these solutions. "By allowing the computer to download the data in multiple threads, we were able to speed up the I/O process from hours to minutes."

The assembly

Once errors have been sifted out, researchers can begin the genome assembly. This process relies on computer programs to join k-mers—short DNA sequences consisting of a fixed number (K) of bases—at overlapping regions, so they form a continuous sequence, or contig. If the genome has previously been sequenced, scientists can use the recorded gene annotations as a reference to align the reads. If not, they need to create a whole new catalog of contigs by performing de novo assembly.

De novo assembly is very memory-intensive and until recently nobody had successfully figured out how to parallelize this process in distributed memory. So many researchers use specialized large memory nodes, several terabytes in size, to do this work. But even the largest commercially available memory nodes are not big enough to assemble massive genomes like wheat or pine. Although researchers previously tried to overcome this memory limitation with supercomputers, inefficient codes meant that it still took several hours, days or even months to assemble a single genome.

To make efficient use of massively parallel systems, Georganas created a novel algorithm for de novo assembly that takes advantage of the one-sided communication and Partitioned Global Address Space (PGAS) capabilities of the UPC (Unified Parallel C) programming language. PGAS essentially allows researchers to treat the physically separate memories of each supercomputer node to be addressed as one address space, which cuts down on the time and energy the supercomputer expends swapping information between nodes.

"The new parallelized version of Meraculous shows unprecedented performance and efficient scaling up to 15,360 processor cores for the human and wheat genomes on NERSC's Edison supercomputer," says

Georganas. "This performance improvement sped up the assembly workflow from days to seconds."

And like the process of piecing together a jigsaw puzzle, "missing pieces" can further complicate genome assembly. This is like having enough of a puzzle pieced together so that you have an idea of what the picture is and where all the pieces should fit, but there are still gaps in the picture. In genome assembly, Meraculous scans the whole picture to identify where these gaps are and then uses an established technique to fill them in. Computationally, this process is done in two-phases. With Hofmeyr's help, both these phases have been converted to UPC and parallelized.

"The result of converting this part of the pipeline to UPC is a speedup of 20 to 30 times faster than the original Meraculous code, which was written in Perl," says Hofmeyr. "This conversion also allows the user to load balance dynamically, so it's a lot cleaner."

Tackling the metagenome

"The value of robust genome assembly is clear. It is the starting point for characterizing the genes of an organism, for doing a comparative analysis across species and for assessing genetic variations. It also gives us a reference to judge the accuracy of new sequence-base methods," says Jarrod Chapman, who developed Meraculous at JGI.

"Before this version of Meraculous, it often took longer to compute the analysis than it did to sequence the data. Because computation was so time-consuming, I would choose a set of parameters based on some educated guesses, set my job to run and that would be my result," says Chapman.

Now that computation is no longer a bottleneck, Chapman can try out a

number of different parameters and run as many analyses as necessary to produce very accurate results. He also believes this achievement means that Meraculous could also be used to analyze metagenomes—microbial communities recovered directly from environmental samples. This work is important because many microbes exist only in nature and cannot be grown in a laboratory. These organisms may be the key to finding new medicines or viable energy sources.

"Analyzing metagenomes is a tremendous effort," says Chapman. "If assembling a single genome—for instance wheat—is like piecing together one novel, then assembling metagenomic data is like rebuilding the Library of Congress. Using Meraculous to effectively do this analysis would be a game changer."

"I think the key to our project's success is the tight synchronization with the domain scientists," says Hofmeyr. "When you look at a project like this, you can see the different perspectives. The domain scientist has a problem that they are trying to solve and a process to get there. As computer scientists we're equipped with a 'bag of tricks.' By understanding the scientific perspective we can see which of these tools are most useful for speeding up their process."

More information: For more information about Meraculous: jgi.doe.gov/big-plant-genomes- … nger-insurmountable/

Provided by University of California - Berkeley

Citation: Novel algorithms and computational techniques speed up genome assembly from months to minutes (2015, July 1) retrieved 26 April 2024 from https://phys.org/news/2015-07-algorithms-techniques-genome-months-minutes.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.