PHYS.ORG

Scientists help tame tidal wave of genomic data using SDSC's trestles

September 18 2013

Sequencing the DNA of an organism, whether human, plant, or jellyfish, has become a straightforward task, but assembling the information gathered into something coherent remains a massive data challenge. Researchers using computational resources at the San Diego Supercomputer Center (SDSC) at the University of California, San Diego, have created a faster and more effective way to assemble genomic information, while increasing

In a paper presented the past month at the 39th International Conference on Very Large Databases (VLDB2013) in Riva del Garda, Italy, Xifeng Yan, the Venkatesh Narayanamurti Chair of Computer Science at the University of California, Santa Barbara, explains how he used SDSC's *Trestles* compute cluster to help develop a new algorithm called MSP (minimum substring partitioning) that helps to assemble genomes with extreme efficiency. MSP is a critical part of a pipeline, or a group of software that assembles entire genomes, with each piece of the software doing one part of the job. Yan and his colleagues were able to optimize one of two steps to use a mere 10 gigabytes of memory without runtime slowdown.

"High-quality genome sequencing is foundational to many critical biological and medical problems," said Yan. "With the advent of massively parallel DNA sequencing technologies how to manage and process the big sequence data has become an important issue. Experimental results showed that MSP can not only successfully complete the tasks on very large datasets within a small amount of memory, but also achieve better performance than existing state-of-the-art algorithms."

According to Yan, his experimental results demonstrate that MSP's improvement in efficiency might soon make it possible to assemble large genomes using smaller, less expensive, commodity clusters rather than requiring high-cost, high performance resources.

Knowing the whole genome of various species underlies biological and medical research, such as understanding evolution pathways or identifying the cause of diseases. However, existing sequencing techniques produce huge amounts – billions for a higher organism such as a human – of overlapping short sequence randomly sampled from the genome. A major challenge in genome research is to assemble those short reads, which vary from ten to several hundred bases, back into the whole genome, a task that requires vast amounts of memory. It would be similar to gluing together an encyclopedia from a haystack of words and sentence fragments.

Using *Trestles*, Yan and his colleagues demonstrated that MSP reduces one of the steps required so that it uses significantly less memory than widely-used algorithms, removing one of the bottlenecks in processing whole genomes. Algorithms such as Velvet and SOAPdenovo struggle to computationally to prepare a virtual scaffolding upon which to assemble the sequence into complete genomes. MSP, a disk-based partition method, streamlines the creation of such scaffolding, known as a De Bruijn graph. A mammalian-sized genome processed using other algorithms would consume hundreds of gigabytes of memory, while MSP allows researchers to complete a key step to ten gigabytes of memory without runtime slowdown.

Yan and his colleagues are working on a second step that also consumes a significant amount of memory, and have so far reduced its memory use by two-thirds with the goals of further reductions in the future. Additional researchers include Yang Li, Pegah Kamousi, Fangqiu Han, Shengqi Yang, and Subhash Suri, all with UC Santa Barbara.

The full paper can be viewed at http://www.cs.ucsb.edu/~xyan/papers/vldb13_debruijn.pdf.

Provided by University of California - San Diego

Citation: Scientists help tame tidal wave of genomic data using SDSC's trestles (2013, September 18) retrieved 20 March 2024 from https://phys.org/news/2013-09-scientists-tidal-genomic-sdsc-trestles.html