New technology reduces storage needs and costs for genomic data

July 6 2010

A new computer data compression technique called Genomic SQueeZ (G-SQZ), developed by the Translational Genomics Research Institute (TGen), will allow genetic researchers and others to store, analyze and share massive volumes of data in less space and at lower cost.

Created specifically for genomic sequencing data, the encoding method underlying G-SQZ and its software use are described in a paper published today in the journal *Bioinformatics*.

Tests show that G-SQZ can compress data by as much as 80 percent while maintaining the relative order of the data and allowing for selective content access. This could save researchers and others millions of dollars worldwide.

Plans are to make the G-SQZ program freely available for research and academic use, and to explore commercial opportunities in genomic data storage and processing. TGen has filed a patent application for the G-SQZ technology.

"Data storage and processing costs are becoming a large factor in research planning as high-throughput genomic sequencing studies continue to generate increasing amounts of data. G-SQZ has the potential to save individual institutes hundreds of thousands of dollars per year in storage costs," said Dr. Waibhav Tembe, the paper's lead author and TGen's Senior Computational Scientist, who led the development of the G-SQZ algorithm and its software.

Enormous computing power is required to conduct today's cutting-edge analysis of large volumes of genomic sequencing data. This data is critical in studying the genes that are a part of the 3-billion-letter DNA sequence, the entire genome of one person. Such analysis is enabling researchers to identify those genomic components that either prevent or contribute to diseases, such as cancer, diabetes and Alzheimer's, and to discover treatments tailored to individual patients that can prolong and increase their quality of life.

Today's genomic sequence analysis requires analyzing terabytes of data. Large sequencing centers are planning or have installed petabyte-scale storage. One terabyte is more than 1 trillion bytes of data. One petabyte is 1,000 terabytes.

Benefits shared with other institutes

Dr. Edward Suh, TGen's Chief Information Officer, described G-SQZ as a significant breakthrough in storing and analyzing ever-increasing genomic sequencing data.

"As a non-profit research institute dedicated to advancing science for the public good, we at TGen are proud to be able to share aspects of this technology with other non-profit research institutes, especially in these times of tightened budgets," said Dr. Suh, who also is a Senior Investigator at TGen and co-author of the paper.

James Lowey, TGen's Director of High-Performance Biocomputing and the third co-author of the paper, said reducing storage costs for genomic technology has the potential to eventually lead to a chain reaction of lower health costs for medical institutions and, ultimately, for patients.

"When you reduce the need for storage, you also are reducing your overhead costs, such as electricity and space, and that can save money,"

Lowey said.

The software is available for download from http://public.tgen.org/sqz.

Technology springs from Next-Gen research

Dr. Tembe's motivation for G-SQZ came from the challenges involved in storing, processing, parsing and transferring enormous Next-Generation Sequencing data, which primarily is stored in plain text formats.

"Generating this data is one thing. It is quite another to store, query and manage it in an efficient manner, minimizing data-analysis bottlenecks and expediting the discovery process," Dr. Tembe said.

The G-SQZ approach is a novel application of Huffman coding of information, an idea first developed in the 1950s, which uses shorter codes for most frequently-occurring pieces of information.

Dr. Tembe's solution is specific to genomic sequencing data. In addition to analyzing the frequency of the ACGT letters that make up DNA, G-SQZ also can encode the annotation information, including the data's quality, as well as erroneous entries, such as unidentified bases.

The indexing system used in G-SQZ allows access at regular intervals, such as every millionth data point, so all the information need not be decoded from the start.

"It's not enough to compress the information. The compressed representation should allow quick retrieval and querying," Dr. Tembe said. "To that end, G-SQZ has been designed as an efficient practical approach, rather than a theoretically optimal compression algorithm."

Even faster advancements on the horizon

Dr. Tembe is moving ahead with improving his current design to accommodate what he calls "parallel computing."

Because G-SQZ compression keeps the data ordered and indexed, the squeezed data can be split into smaller "chunks," allowing multiple computer processors to decode and analyze different parts of the same file simultaneously, he said. For example, if a file is indexed at 1,000 places, it can be fed into a supercomputer, allowing 1,000 processors to analyze the data at the same time, speeding up the results. Analysis tools using parallel programming approaches can take advantage of the G-SQZ encoding format.

"While indexed and compressed representation is ready, the parallel computing functionality is undergoing a testing phase," Dr. Tembe said. "But this is where it is headed. Sequencing hundreds of billions of bases per run is now a reality. The real impact of G-SQZ lies in the storage, transfer and processing of genomic sequencing data, where substantial room for improvement still exists."

Provided by The Translational Genomics Research Institute