So Long, Data Depression

August 31, 2009

Byline: Matthew Dublin
Newsletter: Genome Technology
Genome Technology - September 2009

What if you never had to throw away any data, and if at the same time you could spend just a fraction of what it costs to maintain an ever-expanding collection of hard drives? The way to get there might be data compression.

Traditionally, compression has not been at the forefront of the storage discussion for IT folks dealing with massive amounts of next-gen sequencing data. When everyone started to see hard drive meters pushing past full, what naturally occurred next was a panic about obtaining more storage hardware, quickly followed by a paradigm shift when it came to data management and organization, instead of thinking more about how to make use of what hardware they already had.

So is data compression really a viable option for next-generation sequencing data management? David Lifka, director of the Center for Advanced Computing and an adjunct associate professor of computing and information science at Cornell University, had been on the prowl for a data compression solution. The Center for Advanced Computing is a research service facility for Cornell that provides, in addition to computational and data resources, data storage for the Cornell research community. "The research groups in life sciences are often forced to find the cheapest solution, and I do mean the cheapest solution, which in some cases is scary — like USB drives that you plug into your laptop. … It's not a very manageable solution or reliable one, but that's the price point that we can really swing," Lifka says. "So we looked around and tried to find a device that would hit the right price point but be very manageable, provide an interesting economy of scale so we could attract a broad group of researchers across campus and the med school."

The device his team found came in the form of a commercial data compression solution that is being offered by Ocarina Networks. The company's ECOsystem solution comes packaged with about 100 algorithms that support more than 600 file types. Ocarina claims an average of 50 percent compression of Affymetrix's .dat, .arr, and .cel formats and 45 percent to 85 percent for Illumina's .tiff and .txt files.

Lifka's approach is to use ECOsystem in conjunction with large Data Direct Network storage devices —50 one-terabyte drives, to be exact. Needless to say, the new compression solution is really helping to stretch those storage dollars. "One of the biggest users of the current storage is the bioinformatics group. They were afraid they'd be throwing data away but because of the DDN storage they were able to afford to keep that data around, and now with the Ocarina device they're getting very, very good compression," says Lifka. "They initially bought from us 16

terabytes of useable storage, but with compression they're almost getting two-for-one on their storage, and divisions of life sciences on campus are looking at developing sequencing services that layer on top of us as a solutions provider for the storage."

The only criticism Lifka has for Ocarina is lack of good Windows support, which is no small complaint as the number of life sciences Windows users seems to be growing. "We have a lot of life sciences people that have Windows apps and Windows data and there's not a native decompression there, but anything on the Linux side or Apple side is quite good," he says.

Ocarina founder Carter George says that the real secret sauce to his company's compression rates has to do with a built-in, tailor-made approach that doesn't just use one method of compression for all scenarios. "We are 'content aware.' We don't just have one algorithm and we try to apply that to everything and hope that it shrinks — we look at a file and we have logic that runs that says, 'What is this file?' and based on that we'll figure out what the best strategy is for shrinking it," George says. "In life sciences we've gotten to the point where we're very specific for everything that happens in the pipeline. So the intensity table comes out and we'll recognize what kind of sequencer it came from … and we have compressors that can recognize call bases and quality scores and compress those so it's different from the intensity table."

Right now, the smallest market for Ocarina's compression solution also happens to be the biggest fish in the life sciences pond. George says sites that spend considerable amounts of money on storage to accommodate ever-growing data bases like the Broad Institute, the Wellcome Trust Sanger Institute, European Bioinformatics Institute, and the National Center of Biotechnology Information, could potentially use the solution as an effective cost saving device. "They're spending millions of dollars on storage — really they're some of the biggest storage consumers in the world — so for those guys the proposition is really clear. The cost of buying us is relatively small compared to the amount of storage spending," George says. The fastest-growing market for the young company is big pharma and although universities are interested, he adds, the price tag is still a bit high.

Academic alternatives

To deal with price barriers, the academic community is steadily churning out proof-of-principle open-source compression approaches for genome sequences. "The key idea there is that all of our genomes are almost 99 percent identical, and most of that data is identical and there's no real reason to look at each genome separately and then try to compress that separate data," says Scott Christley, a postdoctoral researcher in the computer science department at the University of California, Irvine. "We don't need to compress the whole genome, we only need to compress differences or variations with what we call a reference genome, like the original human genome that was sequenced."

Christley was part of a team that reduced James Watson's genome to 4 MB — basically the size of an mp3 music file. Christley's colleagues have written a C++ demonstration program called DNAzip,available for download on the team's website, but they have not gotten it up to general utility speed. They are hoping that users can download DNAzip and help develop it into a general utility tool, but Christley knows that there is always a barrier to introducing a new technology.

Marty Brandon, a postdoc at the Institute for Genomics and Bioinformatics also at UC Irvine, recently published a paper presenting his team's compression solution for genomes. Using the Cambridge Reference Sequence as their reference genome, the team was able to compress 3,615 genome sequences taking up roughly 56 MB in GenBank down to 167 KB. "I can tell you

that, by and large, genome sequencing centers are not utilizing compression; so you go to the major databases and they just download raw text, which is enormously expensive not only in the storage space but transfer as well," Brandon says. "I really feel that the bandwidth that you save in transferring these sequences is equally if not more important than the storage aspects."

But it is not just big data repositories that are being targeted by researchers experimenting with compression. Personal genomics and a world in the not-so-distant future where individuals have entire health records on one device will definitely benefit from compression technology. "You're going to have lots of individuals getting large quantities of data that they might carry around with them as part of their personal record between hospitals," says Brandon. "I think [compression] is going to be really important in that aspect. I'm not sure that a hardware solution would even be appropriate for these guys — even though technically the hardware might be capable of storing the data, you still have a large chunk of data that you need to store from point A to point B. These individuals are going to have to manipulate it some way."