'Condor' brings genome assembly down to Earth

July 20 2010, by Chris Barncard

(PhysOrg.com) -- Borrowing computing power from idle sources will help geneticists sidestep the multimillion-dollar cost of reconstituting the flood of data produced by next-generation genome-sequencing machines.

A team of computer scientists from the University of Wisconsin-Madison and the University of Maryland recently assembled a full human genome from millions of pieces of data — stepping up from commonly assembled genomes several orders of magnitude less complex — and they did it without a big-ticket supercomputer.

Genome-sequencing machines now read at a dizzying rate the molecules that make up DNA, shrinking the time required to spell out a bacteria genome from weeks to hours. The cost comes in analyzing the data.

"You have lots of data being produced by sequencing machines, and the only solution for assembly is to maintain a very expensive computing cluster dedicated to this process," says Mihai Pop, computer science professor at Maryland's Center for Bioinformatics and Computational Biology.

To quickly read through DNA, sequencers chop a strand into sections just tens or hundreds of base pairs long.

"It produces very cheap but very poorly sequenced data," says David Schwartz, a UW-Madison genetics professor and director of the

Laboratory for Molecular and Computational Genomics. "You've got a pile of short sequences, but you don't know the order they should fit back together."

Pop has been trying to fit the pieces back together for a decade.

"After sequencing, it's like having this huge jigsaw puzzle, but you don't actually have that picture on the box that shows you what you're trying to reconstruct," Pop says. "And there are a lot of sky regions. And a lot of the pieces will fit in more than one place."

Until Pop and fellow Maryland professor Michael Schatz brought their "Contrail" assembly software to UW-Madison's Center for High Throughput Computing, the popular working genome assemblers — such as one called Velvet at the European Bioinformatics Institute — were reconstituting bacteria genomes that measure in the thousands of base pairs. The human genome contains more than 3 billion pairs.

Schwartz' work in optical mapping of genomes helped identify the puzzle shapes that should fit together.

"He can give us a map of where certain landmarks are in the genome, and that can help us with ordering these short sequences," Pop says.

But the computing power — measured in complexity and cost — needed to handle the mass of data far outstrips that of the sequencing machines cranking out base pairs by the billion.

"There really is no standard approach to doing this for a human genome," Pop says. "It's virtually impossible to do that on a single machine. We needed access to a large cluster."

UW-Madison had that cluster and had its own tool — called Condor, a

program run by professor Miron Livny at the Center for High Throughput Computing — to manage the work of Maryland's software across a network of computers. Condor breaks up long lists of heavy computing tasks and distributes them across networked computer workstations whose intended use leaves their processors with a little or a lot of idle time.

"In the assembly, you have a very complex job workflow. You must take the data and do this analysis and that analysis, and when that analysis is done you take the results from the first two and do a third," says Todd Tannenbaum, project manager for Condor. "You have this big chain of events that need to happen, and that's what Condor does very well."

To manage both the complex workflow chain and the large data management problems, the Condor group added features of Hadoop, another distributed-computing tool adept at spreading data storage and retrieval across networks, to the mix to help haul around the billions of letters gleaned from human DNA by a sequencing machine.

"By running them together, we're able to efficiently run this biological application — efficient not just in terms of computer time, but efficient in terms of dollars," says Greg Thain, a UW-Madison systems programmer who worked closely on the effort with Condor Project graduate student Faisal Khan. "Because Condor could efficiently schedule the work, Maryland didn't have to buy a multimillion-dollar disc cluster."

And on the first successful run, they needed just four days and about 150 computer processors.

"It's two plus two equals five, if you will," Tannenbaum says. "Condor integrated with Hadoop is a software system powerful enough to tackle problems as complex as human genome assembly without the need for

expensive supercomputers or dedicated special-purpose hardware, lowering the barrier of entry for labs across the country to make contributions in this important area of research."

While there is more work to do before the process can be made available to the genomic computing public, it should be flexible enough to assemble genomes on all sorts of networks, including rented computer time available from sources such as amazon.com.

"The combination of Pop and colleagues' algorithms and Miron and colleagues' computing I think really make this all work," Schwartz says. "And what I mean by making it work is making effective use out of the data to create the full picture of a genome and allow us to discern genomic differences between individuals."

More information: www.cs.wisc.edu/condor/

Provided by University of Wisconsin-Madison