



Evolutionary theory, web-search technology combine for DNA analysis

October 4, 2012

Evolutionary theory, web-search technology combine for DNA analysis

Bioinformatics breakthrough has clinical & environmental applications

LOS ALAMOS, NEW MEXICO, October 4, 2012—New software from Los Alamos National Laboratory called Sequedex uses evolutionary theory to swiftly identify short “reads” of DNA, calling out the specific organisms from which the DNA came and their likely activity.

“Sequedex makes it possible for a researcher to analyze data hot off a DNA sequencer using a laptop,” said Joel Berendzen, a scientist on the project. “The tool characterizes whole communities of microorganisms such as those in the mouth in a matter of minutes.”

Sequedex works like a web search engine, making exact matches between DNA sequences and a list of “keywords” called phylogenetic signatures, then placing any hits on the appropriate branch of the Tree of Life. Advantages over current methods include a factor of 250,000 in speed and the ability to work with pieces of DNA as short as 30 bases long.

The software, developed by Los Alamos scientists Joel Berendzen, Nicolas Hengartner, Judith Cohn, Mira Dimitrijevic and Benjamin McMahon, recognizes proteins from short DNA sequences, analyzing them both individually for phylogeny and function and collectively for biodiversity and environmental similarities.

“Sequedex is bioinformatics redesigned from the ground up,” said Berendzen, “making use of the wealth of genomic data that has become available in the 20 years since the most commonly used algorithms were written.”

Data analysis is widely perceived as a bottleneck preventing broader use of DNA sequencing for problems such as rapid clinical diagnoses of viral and bacterial diseases, genetic matchmaking between individual tumors and chemotherapy agents, and improved production methods for algal biofuels. A number of ways around this bottleneck have been proposed, including special computer hardware and farming out analysis to large numbers of computers on computing clouds.

The Sequedex team was originally tasked with investigating DNA analysis on the Laboratory’s Roadrunner supercomputer, but quickly realized that improvements in

the algorithm made having so much hardware unnecessary. “They asked us to build a rocket ship,” Berendzen said, “but instead we built a 10,000 mph motorcycle.” Sequedex software running on a single CPU core can analyze sequences at a rate of 6 billion DNA bases per hour. This rate is more than twice the speed of data generated by today’s fastest sequencing instruments, and it is also more than twice the rate of typical upload speeds to a cloud-computing site.

A journal article on the project, “Rapid Phylogenetic and Functional Classification of Short Genomic Fragments with Signature Peptides,” was published in the open-access, peer-reviewed journal BMC Research Notes.

Sequedex was recently announced as one of this years' winners of [R&D Magazine's "R&D 100" awards](#), one of four from Los Alamos National Laboratory and its partners.

The project was funded with Laboratory Directed Research and Development dollars.

A free demo version is available online at <http://sequedex.lanl.gov/>. The laboratory’s technology transfer office is actively seeking strategic partnership opportunities.

DNA sequencing came to prominence as a result of the Human Genome Project, which was completed in 2003 and found some 25,000 genes in the 3 billion chemical bases that make up the sequence of human DNA. The Human Genome Project arose out of research at Los Alamos and elsewhere in the U.S. Department of Energy into the effects of energy use on human health.

DNA sequencing technology is evolving at a dramatic rate. Costs have dropped by a factor of roughly 300,000 in the past 10 years and the resulting increased flows of sequence data have placed more stress on an already overburdened analysis process.

The current most-widely-used piece of DNA analysis software, a package called the Basic Local Alignment Search Tool (BLAST), was a refinement of software written by Los Alamos scientists Temple Smith and Michael Waterman in 1981.

Los Alamos National Laboratory

www.lanl.gov

(505) 667-7000

Los Alamos, NM

Managed by Triad National Security, LLC for the U.S Department of Energy's NNSA

