Ohio Supercomputer Center

NEWS RELEASE

For Immediate Release: June x, 2010

Contact: Kathryn Kelley, Director of Outreach, Ohio Supercomputer Center, 614-292-6067, kkelley@osc.edu
Jamie Abel, Director of Media and Communications, Ohio Supercomputer Center, 614-292-6495, jabel@osc.edu

Doctoral candidate devises genetic prediction algorithm

Study looks at mid-range inhomogeneous nucleotide sequences

Columbus, Ohio (June x, 2010) – A University of Toledo doctoral candidate in biomedical sciences recently combined the inspiration he received from his grandfather, values learned from his mother, insights gleaned from his mentors and processing power tapped from a supercomputer to unlock a few of the many secrets of the human genome.

Samuel S. Shepard, a native of Bowling Green, Ohio, recently presented his doctoral dissertation, "The Characterization and Utilization of Middle-range Sequence Patterns within the Human Genome." Leveraging high performance resources of the Ohio Supercomputer Center, Shepard was able to compute within days complex, optimized algorithms that he estimated would have taken him more than three and half years to run on a typical desktop computer.

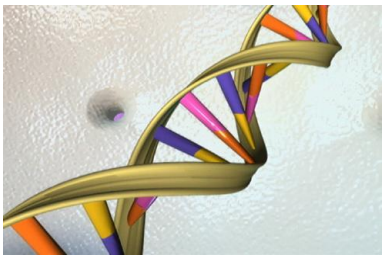
A quick primer on genomics:
Within each cell's nucleus, deoxyribonucleic acid, or DNA, carries the information needed to create and sustain most living organisms. Most DNA is made up of a pair of twisted strands composed of paired subunits, called nucleotides, that comprise the nucleotide bases or "letters" of the genetic alphabet: adenine (A), thymine (T), guanine (G) or cytosine (C). Just as the order of letters determines the meaning of a word, the order of the bases determines the meaning of the information encoded in that part of the DNA. Some sequences or "words", called exons, are translated into proteins that express the genetic instructions, while other sequences, known as introns, often serve as intervening markers between exons or contain "old code" that is no longer used.

Shepard credited his late grandfather, Reynolds Shepard, for the gift of the family's first computer, which "started me on my way." He also dedicated his dissertation to his mother Christy Shepard, "who taught me how to read, how to write, and who instilled in me the love of learning and of excellence."

Shepard's research introduces a novel algorithm for improved prediction of certain genomic sequences, known as exons and introns, within mid-range sequences of 30 to 10,000 nucleotides in length. These genomic "words" are said to display a non-random pattern referred to as "mid-range inhomogeneity," or MRI.

Courtesy: Ohio Supercomputer Center

Shepard leveraged OSC's Glenn Cluster supercomputer (above) to devise an algorithm to predict introns and exons in mid-length sequences of nucleotides found in DNA strands (below).



Courtesy: National Human Genome Research Institute

Courtesy: University of Toledo

Shepard (above, center) pleased his advisory committee with his dissertation on genomic predition techniques.

"We based our approach on Markov chain models, which are the basis for many gene prediction programs," Shepard explained. "During the project, our algorithm read 12 million nucleotides of exons and introns each, and three million are used to train the predictions."

Markov models are limited to analyzing short sequences of nucleotides. However, recent research elsewhere has demonstrated that relationships exist between nucleotides as much as 50 base pairs

– more –

apart in transcribed DNA strands. Shepard and his team hypothesized that MRI was different for exons and introns and would serve as a reliable predictor.

To circumvent the limitations of traditional Markov models, Shepard developed a technique known as binary-abstracted Markov modeling (BAMM). The procedure involves creating rules that reduce mountains of nucleotide information into a much smaller binary code, based upon word length and the nucleotide base found within those words. For instance, if looking for a sequence rich in guanine, Shepard might break the sequence into three-letter words and assign the binary code of "1" to each word containing guanine, "0" if to each word that doesn't.

Shepard was able to test his abstraction rules for words of one or two nucleotides locally at the University of Toledo. As more bases are used to create each binary digit, however, the possible abstraction outcomes increased exponentially, requiring far more computational horsepower. To test rules for longer word lengths, Shepard turned to the Ohio Supercomputer Center (OSC) and its flagship system, the 9500-node IBM Cluster 1350. Shepard and fellow student, Andrew McSweeny, accessed the "Glenn Cluster" to optimize the abstraction process by using "hill-climbing" techniques that determine a single, maximum value for each variable, rather than each of its possible values.

"The trials required approximately 116 individual supercomputer jobs, each using 128 computer cores (32 physical nodes) and taking a little over two hours of wall time per round," Shepard said. "Total optimization for the tetranucleotide abstraction rule took more than ten-and-a-half days for 324 million abstraction rules."

"Researchers at Ohio universities are fortunate to have at their disposal the resources of the Center when their investigations require computational resources beyond those found on their campuses," said Yuan Zhang, client and technology support engineer at OSC. "Beyond the big hardware, OSC also offers researchers the expertise to prepare their jobs to run efficiently on parallel systems."

Shepard and his colleague then combined model variations to improve accuracy. Using support vector machine technology, they determined the best combination of models, achieving a prediction accuracy of greater than 95 percent. Based upon his research, Shepard is preparing a scholarly paper for publication in a professional journal – the sixth he's authored or co-authored while working under advisor Alexei Fedorov, associate professor of Medicine and director of Bioinformatics Lab.

"In his three years, Sam has been involved in dozens of projects in my lab, in different areas of mathematics, genomics and proteomics," said Fedorov. "With his technical expertise, Sam has co-taught the Biomedical Databases summer course for two years, has given a number of lab sessions and lectures for the Perl programming course, and has assisted in the maintenance of the BPG cluster where student data is housed."

Shepard had to miss graduation exercises and asked university officials place his Doctor of Philosophy in Biomedical Sciences diploma and the program's Outstanding Student award in the mail. He had received an Austrian Marshall Plan Foundation scholarship and departed to study in Europe several days before the commencement ceremonies. Funded by the Austrian Marshall Plan Foundation, the prestigious academic exchange program was established with a special focus on universities of applied sciences and technical universities.

XXX

Editor's note:
For high-resolution versions of the images in this release, click on the thumbnails at www.osc.edu/press/releases/2010/shepardMRI.shtml.

*The **Ohio Supercomputer Center** (OSC) is a catalytic partner of Ohio universities and industries, providing a reliable high performance computing infrastructure for a diverse statewide/regional community including education, academic research, industry, and state government. Funded by the Ohio Board of Regents, OSC promotes and stimulates computational research and education in order to act as a key enabler for the state's aspirations in advanced technology, information systems, and advanced industries. For more, visit www.osc.edu.*

*The **Program in Bioinformatics & Proteomics/Genomics at the University of Toledo Health Science Campus** is designed to provide students and researchers with the advanced analytical tools and approaches needed for state-of-the-art biomedical research. The BPG Program is associated with independent, but cooperating programs on the University of Toledo Main Campus and at Bowling Green State University. For more, visit www.utoledo.edu/med/depts/bioinfo/index.html*