Implementing DNA sequencing technology

The first human genome took 15 years and about $3 billion to complete. Soon, doctors at The Research Institute at Nationwide Children's Hospital will be able to sequence two human genomes in just over one week for a fraction of that cost. The institute's recent acquisition of a next-generation HiSeq 2000 sequencing system supports whole-genome DNA sequencing analysis, transcriptome and epigenome analysis, along with multiple other genomic applications.

"We are among the first laboratories in the nation to establish this new instrumentation," said Dr. Peter White, director of the insitute's Biomedical Genomics Core. "This will allow us to enhance the health of children by continuing to expand our programs in areas of strategic emphasis, namely, perinatal research, infectious diseases, congenital and acquired heart conditions, digestive diseases and childhood cancer. Access to this instrument will be particularly helpful for young investigators, postdoctoral research scientists, physician scientists and clinical fellows."

In support, the Ohio Supercomputer Center (OSC) provides additional mass storage for the sequencer's immense data needs. Each eight-day run requires the computational power to align 2,000,000,000 x 100 base-pair reads, using six terabytes of disk space and 600 gigabytes of storage space, plus two additional terabytes if the raw input data is retained.

"OSC also will provide computational backup for local pipeline analysis and resources for more advanced analysis of the datasets that require multi-node support," added Don Stredney, senior research scientist for biomedical applications and director of OSC's Interface Lab. ■

above: The computational power and storage capacity of the Ohio Supercomputer Center support the implementation of a cutting-edge HiSeq 2000 sequencing system by Peter White at The Research Institute at Nationwide Children's Hospital.

Project lead: Peter White, Nationwide Children's Hospital
Research title: Establishing next-generation sequencing technology at Nationwide Children's Hospital
Funding source: National Institutes of Health, American Recovery and Reinvestment Act

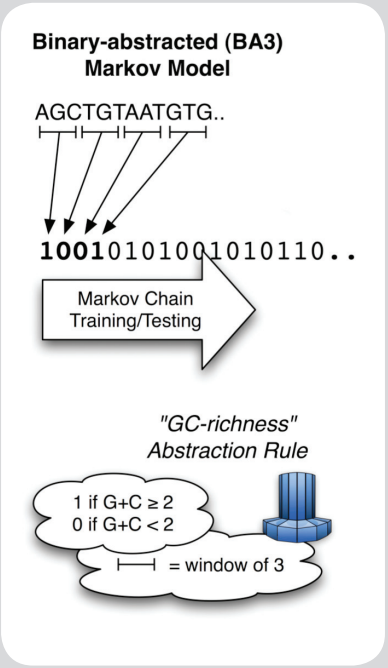
Creating new genomic sequence predictions

Samuel Shepard, Ph.D., a researcher in the University of Toledo bioinformatics lab of Associate Professor Alexei Fedorov, recently developed an algorithm for the prediction of certain genomic sequences, known as exons and introns, using mid-range sequences of 20-50 nucleotides in length. These genomic patterns are said to display a non-random clustering of bases referred to as "mid-range inhomogeneity," or MRI. Shepard hypothesized that the MRI patterns were different for exons and introns and would serve as a reliable predictor.

"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 each were used to test the predictions."

To circumvent the limitations of traditional Markov models, Shepard developed a technique known as binary-abstracted Markov modeling (BAMM). The procedure reduces mountains of nucleotide information into a much smaller binary code. Shepard tested abstraction rules for sequences of one or two nucleotides locally, but as larger sequences were studied, the possible abstraction outcomes increased exponentially.

Requiring far more computational horsepower, Shepard and his colleagues accessed the Ohio Supercomputer Center's Glenn Cluster to optimize the abstraction process by using "hill-climbing" techniques that determine a single, maximal value for each abstraction space, rather than each of its possible values. Shepard and his team then combined different abstraction models to achieve an exon-intron prediction accuracy of greater than 95 percent. ■



Project lead: Samuel S. Shepard, University of Toledo
Research title: The characterization and utilization of middle-range sequence patterns within the human genome
Funding source: National Science Foundation