CMU will tap advanced computer methods to help doctors make sense of their patients' DNA

January 10 2012

Scientists at Carnegie Mellon University say advanced computational tools will be the key to a new research project that, if successful, could enable doctors to routinely use information extracted from a patient's DNA to diagnose and guide treatment of diseases.

Ion Torrent, a unit of Life Technologies Corporation (NASDAQ: LIFE), is sponsoring the project during its first year, and more funding is expected to come through federal grants and other sources. Robert F. Murphy, director of the Lane Center for Computational Biology in Carnegie Mellon's School of Computer Science, will lead a multidisciplinary team of researchers that will collaborate with scientists at the Baylor College of Medicine and Yale University.

The ultimate dream, Murphy said, is to develop what Ion Torrent Founder and CEO Jonathan M. Rothberg dubbed "doctor in a box" software. Doctor-in-a-box would take a patient's DNA sequence and use it to diagnose disease, identify a patient's susceptibility to disease, and predict which therapies might be most effective or cause the fewest side effects. The size and complexity of the human genome, which was first sequenced in its entirety in 2003, has stymied efforts to date to create such software.

"There's just way too much information for doctors to make sense of it all," Murphy said. But new machine learning tools — statistically driven

software that can detect associations within mountains of data — may soon be able to translate the genetic and other hereditary information encoded in the human genome in a way that is clinically relevant to doctors and patients, he added. His team isn't the first to use machine learning to analyze whole genomes, however it will employ some unique software developed at Carnegie Mellon.

The Lane Center includes a number of faculty who are leaders in aspects of the problem, including Eric Xing, Ziv Bar-Joseph, Kathryn Roeder, Russell Schwartz and Seyoung Kim.

The team's software will be trained specifically to analyze the type of whole-genome sequence data produced by Ion Torrent's unique sequencing technology, which is ideal for clinical applications because it is designed to sequence the entire human genome in a day for just $1,000. Up to now, routine clinical use of whole genome sequencing has been impractical because it's taken weeks to complete at a cost of about $10,000. Now that Ion Torrent can reduce the time and expense, the next step is creating a tool to enable doctors to easily integrate whole genome sequencing into medical practice, Rothberg said.

"The promise of 'doctor-in-a-box' is that by using artificial intelligence, like we've seen with IBM's 'Watson' computer, we will be able to associate the variations in the human genome with the vast amount of information we have about human health," said Rothberg (E'85). "The work the Carnegie Mellon team is undertaking opens up the possibility that practicing physicians will be able to diagnose disease, identify disease susceptibility and guide therapy selection as easily as they can now use Apple's Siri on the iPhone."

"It's an enormous undertaking," Murphy agreed, "but we are creating a framework that will allow us to tackle this problem one piece at a time and to do so at a scale that makes sense when all of those pieces are put

together."

The sheer size of the problem necessitates collaboration with other groups trying to understand the genome, so Murphy said the team intends to make its software available as open source.

During the first year, researchers will focus on identifying the genomic features associated with a single disease or patient population, which has yet to be selected. Researchers at Baylor's Human Genome Sequencing Center and Yale's Center for Genome Analysis will perform the whole genome sequencing of patients and provide longitudinal medical records, such as disease treatments and outcomes and results of clinical tests.

This information, scrubbed of patient identity information, will be analyzed by the Carnegie Mellon researchers, who include biologists, statisticians and computational biologists, as well as other computer scientists. Machine learning programs will tease out the relationships between the genomic data and the clinical outcomes for each of the anonymous patients, while incorporating information from biomedical literature regarding gene and protein expression and disease pathways.

This analysis will yield models based on personal genome sequences that can be used to predict disease susceptibility and treatment responsiveness, as well as choose preventive therapies.

To provide impetus to the research program, Rothberg will sponsor an "Analyzing the $1,000 Genome" Conference to be held at Carnegie Mellon sometime in the summer or fall of 2012. The scientific conference will highlight outstanding work on computational analysis of genome sequences and foster discussion of new directions and strategies for extending this research.

In addition to Murphy, the research program leadership includes Jaime

Carbonell, director of CMU's Language Technologies Institute; Tom Mitchell, director of CMU's Machine Learning Department; Richard Gibbs, director of Baylor's sequencing center; and Shrikant Mane, director of Yale's genome center.

Rothberg also established the Rothberg Research Awards in Human Brain Imaging at Carnegie Mellon to support the university's faculty and students in creatively pushing research boundaries in how the brain thinks, learns and ages.

Provided by Carnegie Mellon University