Virginia Tech to tackle the 'Big Data' challenges of next-generation sequencing with HokieSpeed

October 3 2012

Wu Feng, associate professor of computer science in the College of Engineering at Virginia Tech, will engage in Big Data research with promising advances for genomics. Credit: Virginia Tech

The National Science Foundation (NSF) and the National Institutes of Health (NIH) today announced nearly $15 million in new big data fundamental research projects. These awards aim to develop new tools and methods to extract and use knowledge from collections of large data sets to accelerate progress in science and engineering research.

Among the awards is a $2 million grant to Iowa State, Virginia Tech, and Stanford University to develop high-performance computing techniques on massively parallel heterogeneous computing resources for large-scale

data analytics.

Such heterogeneous computing resources include the NSF Major Research Instrumentation (MRI) funded HokieSpeed supercomputing instrument with in-situ visualization. HokieSpeed was the highest-ranked commodity supercomputer in the U.S. on the Green500 when it debuted in November 2011.

Specifically, the three-university team intends to develop techniques that would enable researchers to innovatively leverage high-performance computing to analyze the data deluge of high-throughput DNA sequencing, also known as next generation sequencing (NGS).

The research will be conducted in the context of grand challenge problems in human genetics and metagenomics or the study of metagenomes, the genetic material received directly from environmental samples.

On this grant, working together are Srinivas Aluru, a chaired professor of computer engineering at Iowa State University and principal investigator; Patrick S. Schnable, a chaired professor of agronomy, also at Iowa State; Oyekunle A. Olukotun, a professor of electrical engineering and computer science at Stanford University; and Wu Feng, www.cs.vt.edu/user/feng who holds the Turner Fellowship and who is an associate professor of computer science at Virginia Tech. Olukotun and Feng are co-principal investigators.

In previous research Aluru has advanced the assembly of plant genomes, comparative genomics, deep-sequencing data analysis, and parallel bioinformatics methods and tools. Aluru and Schnable previously worked together on generating a reference genome for the complex stalk of corn genome that will help speed efforts to develop better crop varieties.

Feng's relevant prior work lies at the synergistic intersection of life sciences and high-performance computing, particularly in the context of big data. For example, in 2007, Feng and his colleagues created an ad-hoc environment called ParaMEDIC, short for Parallel Metadata Environment for Distributed I/O and Computing, to conduct a massive sequence search over a distributed ephemeral supercomputer that enabled bioinformaticists to "identify missing genes in genomes."

Feng said, "With apologies to the movie Willy Wonka and the Chocolate Factory, one can view ParaMEDIC as WonkaVision for Scientific Data – a way to intelligently teleport data using semantic-based cues. "

Feng also recently studied how heterogeneous computing resources at the small scale and large scale, e.g., HokieSpeed, could be used for short-read mapping and alignment of genetic sequences in support of the philanthropic award that he received from NVIDIA Foundation as part of its "Compute the Cure" program.

With this award, Feng, the principal investigator, and his colleagues created a framework for faster genome analysis to make it easier for genomics researchers to identify mutations that are relevant to cancer.

In all, NSF and NIH announced a total of eight new projects in response to a call for proposals on "Core Techniques and Technologies for Advancing Big Data Science & Engineering," or "Big Data," in March of 2012. They run the gamut of scientific approaches with possible future applications in scientific disciplines, such as physics, psychology, economics, and medicine.

"I am delighted to provide such a positive progress report just six months after fellow federal agency heads joined the White House in launching the Big Data Initiative," said NSF Director Subra Suresh. "By funding new types of collaborations—multi-disciplinary teams and communities

enabled by new data access policies—and with the start of an exciting competition, we are realizing plans to advance the complex, science and engineering grand challenges of today and to fortify U.S. competitiveness for decades to come."

"To get the most value from the massive biological data sets we are now able to collect, we need better ways of managing and analyzing the information they contain," said NIH Director Francis S. Collins. "The new awards that NIH is funding will help address these technological challenges—and ultimately help accelerate research to improve health—by developing methods for extracting important, biomedically relevant information from large amounts of complex data."

The TopCoder Open Innovation platform and process allows U.S. government agencies to conduct high risk/high reward challenges in an open and transparent environment with predictable cost, measurable outcomes-based results and the potential to move quickly into unanticipated directions and new areas of software technology.

Big data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage and process the data within a tolerable elapsed time. Big data sizes are a constantly moving targets currently ranging from a few dozen terabytes to many petabytes of data in a single data set.

Provided by Virginia Tech