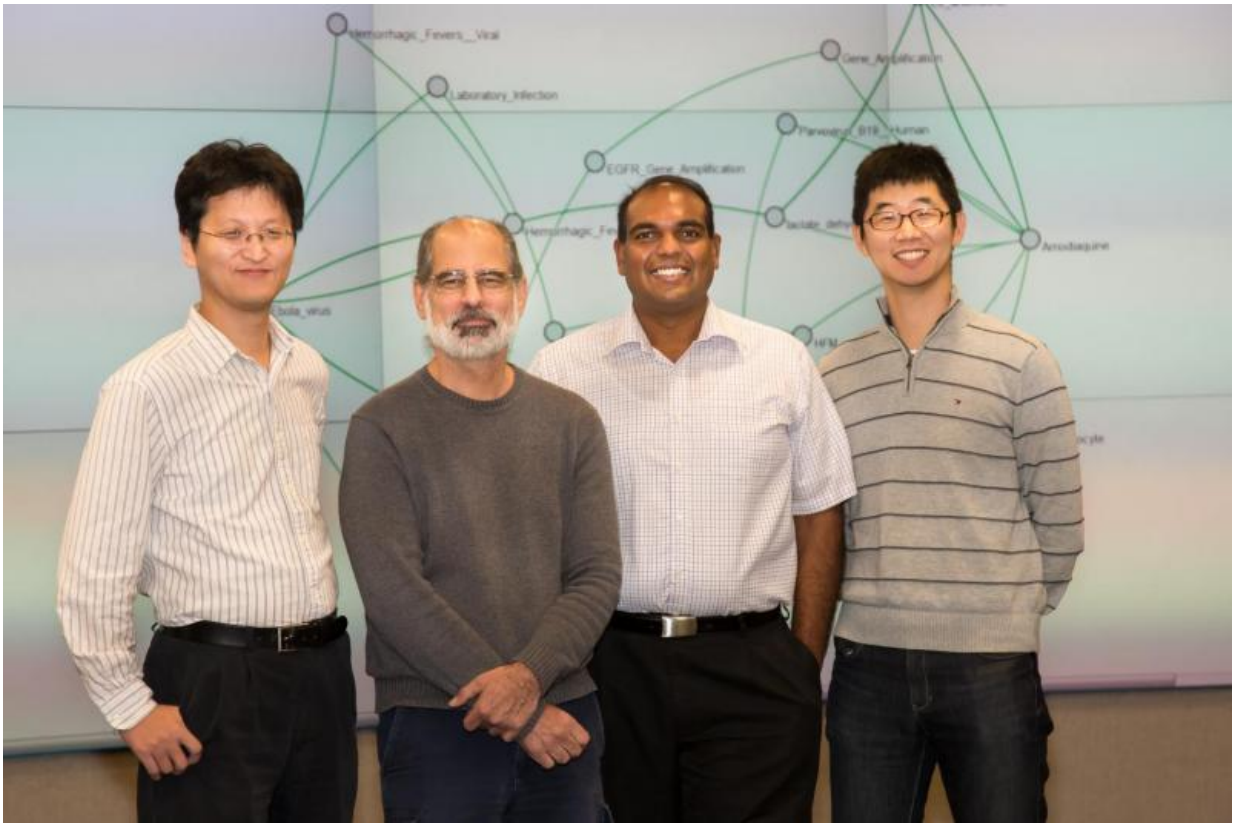
A cure for medical researchers' big data headache

December 7 2015



ORNL researchers (from left) Seung-Hwan Lim, Larry Roberts, Sreenivas Rangan Sukumar and Matt Lee developed a new smart data tool for medical research called ORiGAMI that has the potential to accelerate medical research and discovery. ORiGAMI is the result of collaboration between ORNL and the US National Library of Medicine that made use of CADES resources. Credit: ORNL

As medical research has become more specialized, the scientific community's understanding of the human body has increased, resulting in enhanced treatments, new drugs, and better health outcomes.

A side effect of this information explosion, however, is the fragmentation of knowledge. With thousands of new articles being published by medical journals every day, developments that could inform and add context to medicine's global body of knowledge often go unnoticed.

Uncovering these overlooked gaps is the primary objective of literature-based discovery, a practice that seeks to connect existing knowledge. The advent of online databases and advanced search techniques has aided this pursuit, but existing methods still lean heavily on researchers' intuition and chance discovery. Better tools could help uncover previously unrecognized relationships, such as the link between a gene and a disease, a drug and a side effect, or an individual's environment and risk of developing cancer.

For the past five years, Sreenivas Rangan Sukumar, a data scientist at the Department of Energy's Oak Ridge National Laboratory, has been working with health data and the high-performance computing resources of ORNL's Compute and Data Environment for Science (CADES) to improve health care in the United States. His most recent success, called Oak Ridge Graph Analytics for Medical Innovation (ORiGAMI), supplies researchers with an advanced data tool for literature-based discovery that has the potential to accelerate medical research and discovery.

"Humans' limited bandwidth constrains the ability to reason with the vast amounts of available medical information," Sukumar said. "By design, ORiGAMI can reason with the knowledge of every published medical paper every time a clinical researcher uses the tool. This helps

researchers find unexplored connections in the medical literature. By allowing computers to do what they do best, doctors can do better at answering health-related questions."

The result of collaboration between ORNL and the US National Library of Medicine (NLM), a division of the National Institutes of Health, ORiGAMI unites three emerging technologies that are shaping the future of health care: big data, graph computing, and the Semantic Web, a common framework that allows data to be shared more freely between people and machines.

A Better Way to Search

When medical researchers and clinicians want to know the latest biomedical research, they turn to MEDLINE, NLM's comprehensive database of life sciences and biomedical information. MEDLINE draws from more than 5,600 journals worldwide, adding 2,000 to 4,000 new citations each day to its archive.

A conventional search engine query of MEDLINE can yield results in the thousands—more information than a researcher can review. To improve the usefulness of MEDLINE searches, NLM information research specialist Tom Rindflesch developed software called Semantic MEDLINE that is capable of "reading" key words pulled from the titles and abstracts of articles and summarizing the most relevant information in an interactive graph. The graph, a network of words connected by lines, draws attention to key relationships between the texts and serves as a guide to further exploration. Currently, more than 70 million articles in the MEDLINE database can be searched in this way.

"Semantic MEDLINE is kind of like having a research assistant who looks at a ton of articles and organizes them for you," Rindflesch said.

One of the primary limitations of NLM's Semantic MEDLINE, however, is computing. To produce its results, the application must plow through millions of subject-verb-object groupings pulled from each article and identify the strongest relationships or—better still—the strongest potential relationships, a specialized task that requires a specialized computer. Although conventional data analysis computers excel at reducing large datasets to smaller, more significant datasets, they struggle to compute large graphs capable of linking concepts and weak—yet relevant—associations.

Fortunately, CADES, an integrated compute and data ecosystem within ORNL's Computing and Computational Sciences Directorate, houses a machine with just the right attributes. Apollo, a Cray Urika graph computer, possesses massive multithreaded processors and 2 terabytes of shared memory, attributes that allow it to host the entire MEDLINE database and compute multiple pathways on multiple graphs simultaneously. Combined with Helios, CADES' Cray Urika extreme analytics platform, Sukumar's team had the cutting-edge hardware needed to process large datasets quickly—about 1,000 times faster than a workstation—and at scale.

Once the MEDLINE database was brought into the CADES environment, Sukumar's team applied advanced graph theory models that implement semantic, statistical, and logical reasoning algorithms to create ORiGAMI. The result is a free online application capable of delivering health insights in less than a second based on the combined knowledge of a worldwide medical community.

The Future of Research

In the hands of medical experts and clinicians, ORiGAMI has the potential to increase the efficiency of medical research by directing researchers toward the right questions, an outcome that could reduce

costs and speed up delivery of new treatments. The tool is currently being enhanced beyond literature-based reasoning to data-driven, evidence-supported reasoning using cohort and intervention assessment methods.

Georgia Tourassi, director of ORNL's Health Data Sciences Institute, offered an example of how ORiGAMI is impacting research. Tourassi's team is investigating environmental factors and migration patterns that affect people's cancer risk for a study proposed by the National Cancer Institute's Provocative Questions Initiative. As part of the investigation, the team searched for connections between lung cancer and airborne carcinogens recognized by the US Environmental Protection Agency (EPA).

"When we threw the EPA's top 10 carcinogens at ORiGAMI, we noticed that there were a few elements that appeared over and over as connecting links. Some of these elements made sense from a reasoning point of view, but there was one that we had never seen before," Tourassi said.

The surprising connection was xylene, a common solvent used in the printing, rubber, paint, and leather industries. Past EPA studies focused on xylene as a potential carcinogen have proven inconclusive, but ORiGAMI's results suggested further inquiry. Using publicly available health-related datasets and an advanced web crawler called iCRAWL, Tourassi's team built profiles of xylene exposure for lung cancer patients and non-cancer patients and compared the two.

"The people who had lung cancer had much larger and longer exposures to xylene than the people without cancer," Tourassi said. "This is not confirmation that xylene causes cancer—in order to have confirmation, we need a carefully designed longitudinal cohort study—but this is one more red flag that we should be looking at xylene closely."

In addition to population health, Tourassi's team has used ORiGAMI to explore genomic literature. Tourassi refers to the utility of ORiGAMI as "computer-assisted serendipity," meaning the tool enhances rather than replaces the person making the discovery.

"All of us have those moments of epiphany when certain thoughts click into our head and we move on to explore hypotheses deeper," Tourassi said. "This tool enables that serendipity. It helps guide you in certain ways."

Provided by Oak Ridge National Laboratory