Cancer's big data problem

October 20 2016, by Justin H.s. Breaux

DOE is partnering with the National Cancer Institute in an “all-government” approach to fighting cancer. Called the Joint Design of Advanced Computing Solutions for Cancer, this initial three-year pilot project makes use of DOE supercomputing resources to build sophisticated computational models that facilitate breakthroughs in the fight against cancer on the molecular, patient and population levels. Credit: Argonne National Laboratory

Data is pouring into the hands of cancer researchers, thanks to improvements in imaging, models and understanding of genetics. Today the data from a single patient's tumor in a clinical trial can add up to one terabyte—the equivalent of 130,000 books.

But we don't yet have the tools to efficiently process the mountain of genetic data to make more precise predictions for therapy. And it's

needed: treating cancer remains a complex moving target. We can't yet say precisely how a specific tumor will react to any given drug, and as a patient is treated, cancer cells can continue to evolve, making the initial therapy less effective.

Toward this goal, the U.S. Department of Energy (DOE) is partnering with the National Cancer Institute in an "all-government" approach to fighting cancer. Part of this partnership is a three-year pilot project called the Joint Design of Advanced Computing Solutions for Cancer (JDACSC), which will use DOE supercomputing to build sophisticated computational models to facilitate breakthroughs in the fight against cancer on the molecular, patient and population levels.

The pilot builds on President Obama's Precision Medicine Initiative and Vice President Biden's recent "Cancer Moonshot" to transition cancer therapy away from a "one-size-fits-all" approach. Instead, the goal is to move toward individualized diagnosis and treatment that accommodates a patient's unique body chemistry and genetics.

"Cancer researchers are very good at generating all types of data, from genomic data, proteomic data and imaging data," said Warren Kibbe, director of the Center for Biomedical Informatics and Information Technology at the National Cancer Institute. "What we're not really good at yet is integrating all that information into a consistent model and making predictions on how a tumor will respond to a given treatment."

CANDLE

Key to this collaboration is a computational framework called the CANcer Distributed Learning Environment (CANDLE). Over the years, many projects have amassed a huge volume of cancer data, including tumor genomes, patient data and experiments on potential drugs. CANDLE is designed to use machine-learning algorithms to find

patterns in large datasets. Machine learning is a type of artificial intelligence that focuses on developing programs that can teach themselves to grow and change when presented with new data. These patterns offer insights that may ultimately result in improved treatment or guidance on new experiments.

To date, machine learning studies have produced computational models that estimate drug response for a singular data point—say, a certain mutation. Researchers working with CANDLE, however, envision a higher degree of complexity that integrates many types of information, such as drug interactions and specifics about a patient's genealogy, as well as the tumor's molecular characteristics and how its protein expression varies over time.

"The research community has collected thousands of experiments with hundreds of thousands of data points characterizing tumors and their response to the drugs," said Rick Stevens, an associate laboratory director at the DOE's Argonne National Laboratory and professor of computer science at the University of Chicago. "By working with the national laboratories, the National Cancer Institute can now finally scale and quantify the cancer problem."

Using this computational architecture, participating labs—Argonne, Lawrence Livermore, Los Alamos and Oak Ridge national laboratories—will focus on three problems singled out by the National Cancer Institute as the biggest bottlenecks to advancing cancer research by launching three pilots. These are: understanding key protein interactions, predicting drug response and automating patient information extraction to inform treatment strategies.

Each pilot looks at cancer from a different scale. Lawrence Livermore drives the molecular-level pilot, Argonne leads the patient-level pilot, Oak Ridge undertakes the population-level pilot and Los Alamos looks

at uncertainty quantification across all three pilots.

Molecular level

Thirty percent of all cancers exhibit mutations in the Ras family—a collection of proteins that help trigger cellular machinery to make new cells or kill old ones.

The molecular-level pilot work being led by Lawrence Livermore will use the CANDLE architecture to predict how these proteins behave on the top of cell membranes. Then they will apply this knowledge to what researchers describe as the "Ras pathway problem," where glitches cause genes to remain stuck in the "on" position, leading to cancerous tumors.

Researchers want to produce highly complex simulations that describe how a protein moves and binds to specific locations on the cell membrane. They hope such insights can be applied to the millions of Ras pathways and dramatically enhance our understanding of how they work by predicting the likelihood that a signal will take a certain path.

Patient level

Cancer encompasses hundreds of diseases, each with thousands of possible causes. Thus, bringing precision to therapy selection for a specific patient is the goal of the Argonne-led patient-level pilot.

With the CANDLE platform, researchers at the Argonne Leadership Computing Facility, a DOE Office of Science User Facility, will develop predictive models that guide drug treatment choices for tumors based on a much wider assortment of data than currently used.

To do this, they will merge one type of computational model that uses

data to predict phenomena with another model that uses data to explain them. The hope is that by merging these two methods, they will be able to migrate lessons learned from computer simulations to the research laboratory, where researchers test mice to verify the computer's prediction of how a tumor will respond to a given therapy.

Researchers will also try to find mechanisms for how a particular tumor evades a therapy or develops resistance.

"Conceptually, that's how we're thinking the future of cancer therapies is going to move. Right now we don't understand the biological implications of resistance well enough for any particular therapy to do a good job at predicting combinatorial therapies," said Kibbe. "We think that simulation will allow us to do a much better job of predicting which combination of therapies would be most effective for a specific patient."

Population level

At any one point in time, three to five percent of patients with cancer participate in a cancer clinical trial. And cataloguing all this data is still a manual task.

Oak Ridge will help the National Cancer Institute to scale its ability to monitor cancer patients across the country by automating the process of entering and extracting information. By applying natural language processing and machine learning algorithms to these millions of clinical reports, computers will be able to derive meaning from the notes that doctors and nurses write in their reports.

Once completed, this system would automatically analyze and extract information so that researchers can monitor country-wide outcomes, which can then inform treatment strategies for patients of different lifestyles, environments and cancer types.

Steps are being taken to de-identify data before population-level pilot data is shared with participating labs, Stevens said.

Next steps

Over the next three years, both the National Cancer Institute and DOE have a monumental task; but they have a plan.

The first year will focus on merging statistical models and building machine-learning methods that make the best of their ability to explain and predict phenomena. In the second year, computer scientists will have to computationally estimate how confident they are in those predictions, and in the final year, researchers will put all of these pieces together and integrate experimental design.

"I really think we are at a unique place right now. There are some unbelievably great conversations happening across government right now about how we work together and integrate these brand new tools to enable understanding of these basic processes," said Kibbe. "And if we can really understand the interplay of mutations, normal biological processes and cancer, we have a much better chance of being able to interfere with—and end—cancer."

As the pilots progress from the building and merging of computational models to testing them in the laboratory, Stevens admits that you'd have to be a little fearless to go after a problem like this.

"In my nearly 20 years of working in computational biology, I can say that this is a really hard problem and it's not clear if we know how to do this," said Stevens. "But what the Cancer Moonshot gives us all is the ability to show the world how DOE labs can work in collaboration with the National Cancer Institute in a way that hasn't been done before."

"With that level of collaboration, it starts to look like less of a far-off moonshot," he said, "and more a problem that we have a real shot at addressing."

Provided by Argonne National Laboratory