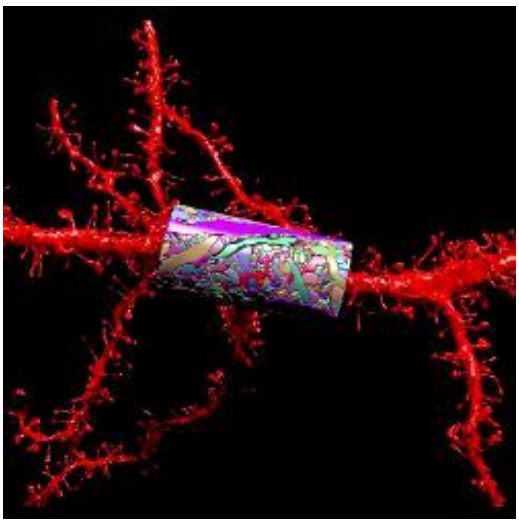
Combating Cancer With Data

By Esther Shein
Communications of the ACM, Vol. 60 No. 5, Pages 10-12
10.1145/3057735
Comments (2)



Researchers used scanning electron microscope images of nanometers-thick mouse brain slices to reconstruct cells into a neocortex structure (center), whose various cell types appear in different colors.

Credit: Argonne National Laboratory

For decades, scientists have worked toward the 'holy grail' of finding a cure for cancer. While significant progress has been made, their efforts have often been worked on as individual entities. Now, as organizations of all kinds seek to put the massive amounts of data they take in to good use, so, too, are the health care industry and the U.S. federal government.

The National Cancer Institute (NCI) and the U.S. Department of Energy (DOE) are collaborating on three pilot projects that involve using more intense high-performance computing at the exascale level, which is the push toward making a billion billion calculations per second (or 50 times faster than today's supercomputers), also known as exaFLOPS (a quintillion, 10^{18}, floating point operations per second). The goal is to take years of data and crunch it to come up with better, more effective cancer treatments.

The DOE had been working on building computing infrastructure capable of handling big data and entered into discussions with the NCI, which houses massive amounts of patient data. The two organizations realized there were synergies between their efforts and that they should collaborate.

The time is right for this particular collaboration because of the application of advanced technologies like next-generation sequencing, says Warren Kibbe, director of the NCI Center for Biomedical Informatics and Information Technology. In addition, data is becoming more readily available from vast repositories, and analytics and machine learning tools are making it possible to analyze the data and make better sense of it.

Says Kibbe, "There is ever-better instrumentation and data acquisition from that instrumentation, such as using cryoEM (cryo-electron microscopy) to generate structural data in biology, that lets us now look at molecules that up until now have been very difficult

to look at." Recently, he adds, "there's been a tremendous infusion of technology in biology," enabling, for example, the ability to interrogate a tissue and determine the types of cells in the tissue and their spatial organization.

Many big challenges still exist, such as learning how individual cells work together in the tumor micro-environment and how they contribute to the overall aggressiveness of cancer and its ability to resist therapies, Kibbe adds.

The opportunity to work with the DOE meant exposure to a tremendous amount of computational expertise and thinking about problems in deep learning and natural language processing (NLP), as well as being able to do very detailed simulations, he says. Taking the available cancer data and using it to build mechanistically informed models and predictive models will enable researchers to better understand, as they perturb a particular cell, how that perturbation is going to impact the tissue and the biological system. It will also tell researchers whether they can "do a better job providing patients with optimal therapies based on the modeling."

For the NCI/DOE collaboration, the goal is not understanding individual cells and tissues, but whether researchers can glean from a huge population how patients respond when they are given a particular therapy. "That's a data aggregation problem and a natural language processing problem," Kibbe says. "The DOE has a lot of expertise in looking not only at energy grids, but thinking about integrating data from a number of different sources and technologies, and building up simulations and models."

One pilot by Argonne National Laboratory focuses on deep learning and building predictive models for drug treatment response using different cell lines and patient-derived xenografts (tissue grafts from a donor of a different species than the recipient). "We're trying to build models where we can predict where tumors we haven't screened will respond to a drug," explains Rick Stevens, associate laboratory director for computing, environment, and life sciences research at Argonne, who is spearheading the deep learning pilot. This is the underlying concept of precision medicine.

Tumor cells have thousands of different types of molecules and tens of thousands of genes that change all the time, so there are fundamental points that researchers don't understand, Stevens explains. Building a model based on principles of what is happening in cancer cells is incomplete; if a researcher tried to make predictions of how a cancer cell will respond without taking into consideration the properties of the treatments, it wouldn't be as effective. That's where the team hopes deep learning applied to drug combination therapies will be useful.

A second pilot, at Lawrence Livermore National Laboratory, is aimed at understanding the predictive paths in the Ras cancer gene, mutations of which are responsible for about 30% of all cancers, Stevens says. Work there is also focused on the oncogene which, when mutated, becomes the driver for causing cancer. "It's one of the core targets we're trying to understand [as well as] how to drug it," says Stevens. "It's stuck in the 'on' position; it's like a switch and it tells your cells when to divide."

A third pilot, under way at Oak Ridge National Laboratory, is mining data from millions of patient records in search of large-scale patterns to optimize drug treatments. The pilot is

working with the Surveillance, Epidemiology and End Results (SEER) Registries, which NCI has used since 1974 to assess the incidence and outcomes for cancer patients across the country and covers roughly 30% of the U.S. population, says Stevens. However, the challenge is that because it was built over 40 years ago, it "has seen a lot of technologies, and the hope is we can transform the SEER Registries into something that has very different characteristics" using NLP and deep learning features.

This is where the partnership with DOE will be especially valuable, says Kibbe, because the department has a lot of expertise working with sensor networks and data aggregation interrogation and analysis.

The common thread among all three pilots is that each has a deep learning component to them, Stevens says. To fund the initiatives, he and his co-investigators received $5 million in fall 2016 from the Exascale Computing Project (ECP) to build a deep neural network code called the CANcer Distributed Learning Environment (CANDLE). This year, Argonne, Lawrence Livermore, and Oak Ridge all will deploy their highest-performing supercomputers available and the teams will use these systems to start evaluating existing open source software from various vendors and test machine learning capabilities. That way, Stevens notes, they won't have to reinvent the wheel.

"We'll add what we need on top of the frameworks and make it possible to use the large-scale hardware we have and feed it back into the open source community," Stevens says. "A wonderful feature of the artificial intelligence community is that it's very open. You have collaborations that span companies that are competing with each other," including Microsoft, Google, and Facebook.

The teams working on the three pilots will "run big benchmark problems on the DoE hardware," and will have the first code release that can serve all three pilots and eventually other application areas in the summer, he says.

One of the problems, in Stevens' case, is a classification problem, in which tumor expression data, known as SNP (single nucleotide polymorphisms) data, is used to try to determine what type of cancer is being studied from the SNPs alone. "That hasn't been done before; it's related, but not the same to classification of gene expression," he says. And there are several other problems as well, including trying to predict the response to an individual drug based on its formula and profile, and the auto encoder problem, in which a network is trained to learn the compressed representation of a drug structure, for example, and then has to be trained to accurately reproduce the input so the team can build an improved algorithm.

The benchmarks will change over time, but they are a way to develop a common language among the vendors and the teams working on the pilots, Stevens says.

Once the first iteration of the model has been built and validated, it should be able to analyze tumor information from a newly diagnosed cancer patient and predict which drug will be the most effective at attacking the tumor.

Meanwhile, to help foster existing collaborations and pursue new ones, the first of a series of meetings was held in July 2016. The Frontiers of Predictive Oncology and Computing meeting focused on predictive oncology and computing in a few areas of interest in NCI/DOE collaboration: basic biology, pre-clinical, clinical applications and computing, says Eric Stahlberg, a contractor working on the high-performance computing strategy within the Data Science and Information Technology Program at the Frederick National Laboratory for Cancer Research in Rockville, MD.

"Efforts at the frontier of pre-clinical predictive oncology ... included developing new models using patient-derived xenografts and predicting drug efficacy through regulatory networks," Stahlberg says. Other areas of focus were how to gain better insights into Ras-related cancers, gathering quality data for use in predictive model development, and improving the SEER database.

"The meeting attendees were very enthusiastic about the prospects for improving cancer patient outcomes with increased use of computing," Stahlberg says. That said, "One of the largest challenges exists in developing interoperability among solutions used in predictive oncology." Others include gathering consistent data and having enough data to understand the complexity of individual cells, he says.

Since the conference, further progress has been made in yet another collaboration: the public-private partnership for Accelerating Therapeutic Opportunities in Medicine (ATOM) involving GlaxoSmithKline, the DOE, and the NCI, he says. Additionally, "most significantly, the 21st Century Cures Act was just signed into law, setting the stage for a very promising future at the intersection of predictive oncology and computing."

Several universities also are actively researching ways to tackle big data, which is a big challenge given the tremendous amount of information collected in the life sciences, notes Sunita Chandrasekaran, an assistant professor in the Center for Bioinformatics and Computational Biology at the University of Delaware, and one of the meeting's organizers.

"Efforts are under way in universities that collaborate with medical research institutes or facilities in order to accelerate such large-scale computations like sequence alignment using accelerators like GPUs (graphics-processing units)," she says. "Efforts are also under way to build suitable and portable software algorithms that can adapt to varying input and generate results dynamically adapting to evolving hardware."

Stevens says what makes it possible now to use data more effectively than several years ago is that researchers have found ways to accelerate deep learning through things like GPUs. "This, coupled with breakthroughs in some methods like convolutional neural networks, has suddenly made deep learning effective on many problems where we have large amounts of training data."

When the single model has been put into effect, researchers will be able to add more information about cancer cells as well as more information about drugs, "and we would have many more instances of 'this drug worked this well on a given tumor,' so many more training pairs between cancers and drugs," says Stevens.

While acknowledging he hates to make predictions, Kibbe feels confident that "in the next 10 years we should see that many of what are very hard-to-treat cancers will be treated," and that regardless of where someone lives and what their socioeconomic status is, they will have access to the same level of care.

"I think that's what will come out of these collaborations and use of computing; as sensors and instrumentation get cheaper and cheaper to implement and become more and more ubiquitous, the hope is there will be a leveling effect on cancer treatment across the country, and perhaps the whole world."

Perhaps working in collaboration, combined with deep learning and highly advanced computing, will prove to be that holy grail. Kibbe calls the DOE/NCI partnership unique in that two very different cultures are working together as a team. While everyone is excited about their individual projects, he says, they are also excited about their joint mission of creating a workforce that has both biomedical knowledge and computational expertise.

"That side of the collaboration is going to continue to pay dividends for as long as we have computation in biomedical research, which I hope is forever."

Author

Esther Shein is a freelance technology and business writer based in the Boston area.

Comments

Braulio Cabral

May 05, 2017 11:50
Great article Ms. Shein,
Just one observation, Dr. Eric Stahlberg is referenced simply as "a contractor", while other references are more specific. Dr. Stahlberg is a scientist in charge of the High-Performance Computing Strategy at the Frederick National Laboratory for Cancer Research, similar to Argonne, a national laboratory owned by the U.S. government and contractor operated.

Best regards,

Braulio J. Cabral, Ph.D.
Acting Deputy Director
DSITP, Frederick National Laboratory for Cancer Research.
Operated by: Leidos Biomedical Research, Inc.

Esther Shein

May 08, 2017 11:04

Thank you for your comments. Eric and I corresponded via email and this was how he signed his notes.
