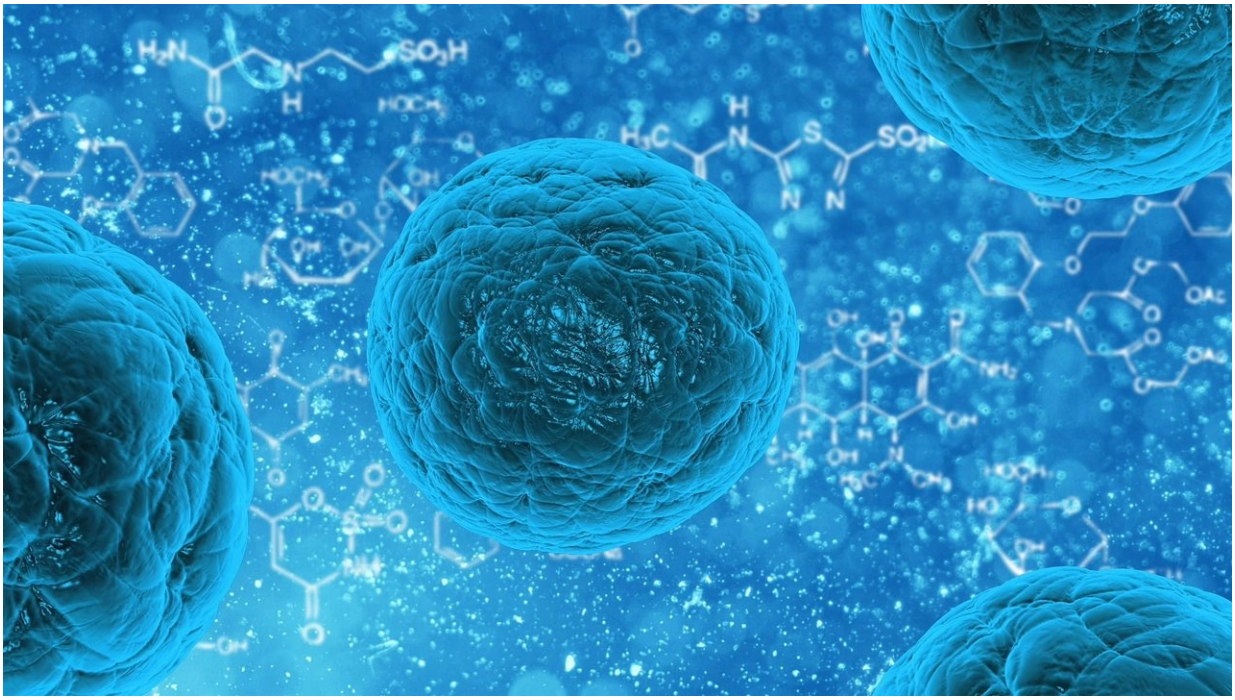
New algorithm uses online learning for massive cell data sets

April 19 2021, by Kelly Malcom



Credit: CC0 Public Domain

The fact that the human body is made up of cells is a basic, well-understood concept. Yet amazingly, scientists are still trying to determine the various types of cells that make up our organs and contribute to our health.

A relatively recent technique called single-cell sequencing is enabling

researchers to recognize and categorize cell types by characteristics such as which genes they express. But this type of research generates enormous amounts of data, with datasets of hundreds of thousands to millions of cells.

A new algorithm developed by Joshua Welch, Ph.D., of the Department of Computational Medicine and Bioinformatics, Ph.D. candidate Chao Gao and their team uses online learning, greatly speeding up this process and providing a way for researchers world-wide to analyze large data sets using the amount of memory found on a standard laptop computer. The findings are described in the journal *Nature Biotechnology*.

"Our technique allows anyone with a computer to perform analyses at the scale of an entire organism," says Welch. "That's really what the field is moving towards."

The team demonstrated their proof of principle using data sets from the National Institute of Health's Brain Initiative, a project aimed at understanding the human brain by mapping every cell, with investigative teams throughout the country, including Welch's lab.

Typically, explains Welch, for projects like this one, each single-cell data set that is submitted must be re-analyzed with the previous data sets in the order they arrive. Their new approach allows new datasets to the be added to existing ones, without reprocessing the older datasets. It also enables researchers to break up datasets into so-called mini-batches to reduce the amount of memory needed to process them.

"This is crucial for the sets increasingly generated with millions of cells," Welch says. "This year, there have been five to six papers with two million cells or more and the amount of memory you need just to store the raw data is significantly more than anyone has on their computer."

Welch likens the online technique to the continuous data processing done by social media platforms like Facebook and Twitter, which must process continuously-generated data from users and serve up relevant posts to people's feeds. "Here, instead of people writing tweets, we have labs around the world performing experiments and releasing their data."

The finding has the potential to greatly improve efficiency for other ambitious projects like the Human Body Map and Human Cell Atlas. Says Welch, "Understanding the normal compliment of cells in the body is the first step towards understanding how they go wrong in disease."

More information: Chao Gao et al, Iterative single-cell multi-omic integration using online learning, *Nature Biotechnology* (2021). DOI: 10.1038/s41587-021-00867-x

Provided by University of Michigan