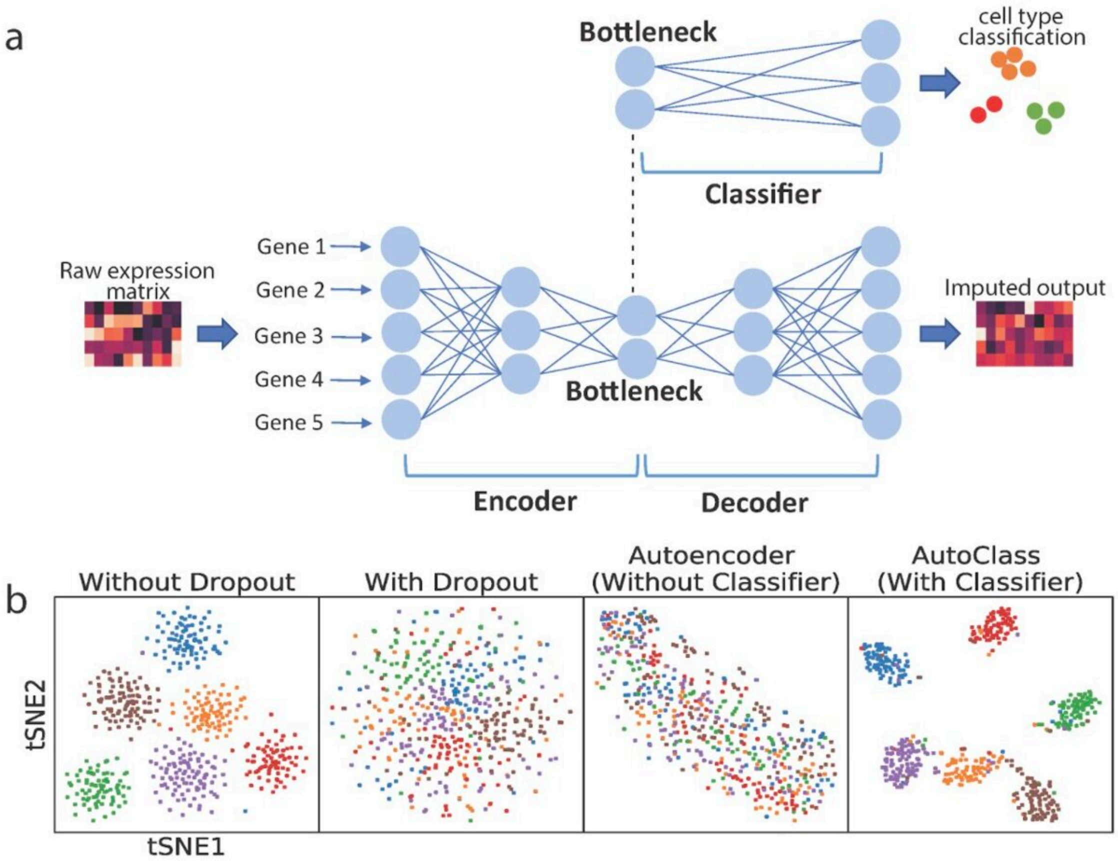
PHYS ORG

Team develops a universal AI algorithm for in-depth cleaning of single cell genomic data

April 7 2022



AutoClass integrates a classifier to a regular autoencoder, as to fully reconstruct scRNA-Seq data. a. AutoClass consists of a regular autoencoder and a classifier branch from the bottleneck layer. The input raw expression data is compressed in the encoder, and reconstructed in the decoder; the classifier branch helps to retain signal in data compression. The output of the autoencoder is the desired imputed data. b. t-SNE plots of Dataset 1 without dropout, with dropout, with dropout imputed by a regular autoencoder and AutoClass. Credit: UNC Charlotte

Just as asking a single person about their health will provide tailored, personalized information impossible to glean from a large poll, an individual cell's genome or transcriptome can provide much more information about their place in living systems than sequencing a whole batch of cells. But until recent years, the technology didn't exist to get that high resolution genomic data—and until today, there wasn't a reliable way to ensure the high quality and usefulness of that data.

Researchers from the University of North Carolina at Charlotte, led by Dr. Weijun Luo and Dr. Cory Brouwer, have developed an artificial intelligence algorithm to "clean" noisy single-cell RNA sequencing (scRNA-Seq) data. The study, "A Universal Deep Neural Network for In-Depth Cleaning of Single-Cell RNA-Seq Data," was published in *Nature Communications* on April 7, 2022.

From identifying the specific genes associated with sickle cell anemia and breast cancer to creating the mRNA vaccines in the ongoing COVID-19 pandemic, scientists have been searching genomes to unlock the secrets of life since the Human Genome Project of the 1990s. Technology has leaped from those early days of batching thousands of cells together to decrypt the millions of base pairs that make up genetic information, and in 2009 researchers created scRNA-Seq, now used widely in biomedical research, which only sequences the transcriptome or the expressed portion of genome in a single cell of a living organism.

Unfortunately, scRNA-Seq data is very noisy and has plenty of errors and quality issues. Sequencing a single cell rather than many cells results in frequent "dropouts"—missing genes in the data. A single cell, like a single person, may have its own health issues or be at an awkward stage in its life cycle—it may have just divided, or be on its way to cell death, which can create more errors or technical variations in the scRNA-Seq data. Besides the single-cell specific problems, genomic profiling usually comes with "normal" issues of sequencing errors. All these errors need to be "cleaned" from the data before it can be used or interpreted, which is where the new AI algorithm comes in.

The algorithm, called AutoClass, is a step up from existing statistical methods. Most existing methods assume that errors (or noises) would follow certain predefined distribution, or how likely it is the errors will occur and how big the errors might be. Existing methods are often unable to fully clean the data as to reveal biological signals, and may even add new errors because of their improper assumptions on data distribution. In the opposite, AutoClass does not make any distributional assumption; hence, it can effectively correct a wide range of noises or technical variations.

"AutoClass is an AI algorithm based on a special deep neural network designed to maximize both noise removal and signal retention." Dr. Luo said, "The AI teaches itself to differentiate signal vs. noise in the data by seeing enough data. Usually the more data it sees, the better it performs."

In the study, Dr. Luo and his team demonstrated AutoClass can reconstruct high quality scRNA-Seq data and enhance downstream analysis in multiple aspects. In addition, AutoClass is robust and performs well in various scRNA-Seq data types and conditions.

AutoClass is highly efficient and scalable, and works well with data of a wide range of sample sizes and feature sizes, and runs smoothly even in a regular PC or laptop. AutoClass is open source and available online.

More information: Hui Li et al, A universal deep neural network for in-depth cleaning of single-cell RNA-Seq data, *Nature Communications* (2022). DOI: 10.1038/s41467-022-29576-y

Provided by University of North Carolina at Charlotte

Citation: Team develops a universal AI algorithm for in-depth cleaning of single cell genomic data (2022, April 7) retrieved 28 April 2024 from https://phys.org/news/2022-04-team-universal-ai-algorithm-in-depth.html