Using deep autoencoder feature embeddings to explore single-cell phenotypes in pediatric cancer

Timothy Keyes

4/22/2020

Project Category: Health Care

1 - Problem Description

In the clinical evaluation of leukemia (blood cancer), most diagnostic and prognostic tests rely on the identification and enumeration of leukemic "blasts" in the blood and bone marrow of patients. In short, blasts are immature blood cells that - due to genetic and epigenetic abnormalities - develop abberrances in cellular maturation that lead them to become cancerous. Blast phenotypes differ widely between patients both because of individual differences in the biology of each patient's cancer and because of instrumentation differences between the clinics where testing is conducted. This means that the current gold standard of diagnostic and prognostic testing for leukemia relies on pathologists manually inspecting the protein-level phenotypes of cancer patients by eye using microscopy, flow cytometry, and related methods.

Due to the labor-intensiveness of current clinical testing protocols for leukemia, the development of high-throughput, automated methods of enumerating leukemic blasts would have great clinical impact in the diagnosis and prognosis of the disease. Thus, here we are interested in using deep learning to take a step towards that goal by building an autoencoder framework capable of denoising and batch-correcting protein-level data collected by single-cell cytometry such that individual differences in protein marker expression are preserved but instrument-to-instrument differences (due to collecting samples between multiple clinical labs) are removed (or reduced). This project will lean heavily on previous implementations of a similar method on single-cell data in non-cancer cells (called SAUCIE) that we hope to adapt for this project.

2 - Challenges

The main challenge for this project will be identifying what components of SAUCIE will need to be altered in order to adapt it to work with cancer data, which often has much larger person-to-person variability than the non-cancer cell datasets on which SAUCIE was originally developed. This may require adjusting the loss function used by SAUCIE and/or adding/removing layers to the autoencoder based on what gives the best results.

3 - Dataset

In 2018, a hallmark paper in the field of leukemia biology claimed that one useful way of understanding the heterogeneity of cancer cells was to "align" them with their most similar healthy cell subtype - with the overall idea being that comparing cancer cell subtypes to healthy cell subtypes might help us to infer how the cancer cells will behave. In their paper, the authors developed a "single-cell classifier" that assigns cancer cells to bins that are similar to known healthy cell subtypes based on protein measurements obtained using mass cytometry. In this project, we will use this paper's dataset to see if our autoencoder denoising/batch correction/dimensionality reduction approach is able to cluster similar developmental cell types (as identified by the original authors) in shared latent space within the autoencoder.

The mass cytometry data for this paper are formatted as .fcs files and provided here. In addition, additional clinical data for the samples are located in a public repository from the Children's Oncology Group (in case this ends up being interesting at later stages of the project).

4 - Preprocessing

Preprocessing of the data according to existing gold-standards for mass cytometry has already been performed. Thus, the data available in the original authors' GitHub repository are already formatted as normalized and debarcoded .fcs files, which means that artifacts from data acquisition have already been removed and individual samples are saved into separate files. This means that the data are already in a relatively clean, tabular form and that analysis should be relatively straightforward. The one challenge will be finding a way to read .fcs files into Python, as they are a relatively uncommon data type (with a lot of inefficiencies due to historical data formatting conventions).

5 - Learning Method

The proposed learning method will use an autoencoder based on SAUCIE, a multitasking network previously developed to batch normalize, denoise, and cluster single-cell data collected from human patients in the context of viral infection. While the TensorFlow code available on the original authors' GitHub will serve as a starting point for this project, we will have to adapt their model by adjusting their hyperparameters and potentially changing the specific loss functions they use for the dimensionality reduction step.

6 - Evaluation

Using our adapted version of SAUCIE, we will assess our results both qualitatively and quantitatively.

- **Qualitative:** We will compare the quality of our feature embedding by comparing plots of our approach to other standard dimensionality reduction algorithms including tSNE, UMAP, and PCA. Our goal is that our method will be able to represent both the differences and similarities between healthy and cancer cell subpopulations at least as well as these existing methods.
- **Quantitative:** We will compare clusters identified using our method to manually-identified clusters (i.e. gold-standard identification of clusters) by clinical experts as indicated in the original dataset's annotations. We will calculate the F-ratio between our clustering and the gold standard as a course metric of success.

7 - References

Citations are hyperlinked throughout this proposal.