Milestone 2 - Using deep autoencoder feature embeddings to explore single-cell phenotypes in pediatric cancer

Timothy Keyes

2020-05-27

Background and project motivation

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 abberrancies in cellular maturation that cause 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 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,[1] flow cytometry,[2] 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, batch-correcting, and clustering 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 reduced. This project leans heavily on existing implementations of a network method called SAUCIE that was developed using non-cancer cells and that we have adapted for this project.[3]

SAUCIE ("sparse autoencoder for unsupervised clustering, imputation, and embedding") is a multitasking autoencoder that "extends" the architecture of a simple autoencoder. It does so by adding several layers that are regularized in such a way that their outputs are biologically meaningful. These layers include a clustering layer that is regularized to penalize within-cluster distances between cells *and* that performs information dimension (ID) regularization (to encourage cluster sparsity); in addition, the "embedding" layer of the autoencoder (i.e. the middle layer between the encoder and the decoder) is regularized such that the pairwise distances between cells from different batches is minimized, allowing for denoising and batch correction. SAUCIE has been previously validated on single-cell data collected from the immune system, in which phenotypic differences between cell types are relatively large. However, it is yet to be shown that SAUCIE can provide meaningful results on cancer cell data, which are more generally characterized by subtle phenotypic differences within expanded cellular sublineages within a tumor.

Current Progress

So far, we have been able to apply the implementation of SAUCIE to our dataset with some success but with a great deal of room for further development. Challenges to date have included navigating SAUCIE's installation and syntax (which turned out to be nontrivial, as the implementation is relatively new and thus poorly documented to date) in order to run it locally.

Description of the dataset

In this project, I am working with data originally published in a previous paper from my lab about leukemia biology that was published in one of the 2018 issues of the biomedical journal **Nature Medicine**. The paper devised a method of "aligning" cancer cells with the healthy cell type to which they are most similar. Overall, their idea was that comparing cancer cell subtypes to healthy cell subtypes might help us to infer how cancer cells behave or where they come from. The link to the original paper can be found here.[4]

Specifically, the dataset contains mass cytometry data from 60 patients with B-cell precursor acute lymphoblastic leukemia (BCP-ALL) and 5 healthy control patients. The data were collected on a cytof2 mass cytometer and appropriately normalized, batch-corrected, and clinically annotated prior to our analysis. The data are stored as .fcs files, a file format developed by the International Society for the Advanced of Cytometry.[5]

These single-cell data can be represented as an [m x n] matrix in which `m` represents the number of cells that you've measured and `n` represents the number of proteins that you've measured within each cell. Each cell was collected from one of the 60 patients enrolled in the study (thus, many thousands of cells come from each patient). A summary table of how many cells came from each sample is provided in the Appendix (note that there are sometimes several samples that correspond to a single patient as well!).

For this report, we worked with a cleaned version of these data to eliminate some of the frustration of working with samples that vary significantly in size and quality (often, when there are relatively few cells in a sample, it means that the sample was not very "viable" when it was collected, which means that many cells were dead or dying. Low-viability samples are often best to throw away entirely, as even the cells that lived were probably on their way to cell death). Specifically, we limited the dataset to solely diagnostic specimens taken from the blood or bone marrow, and we sampled 10,000 cells from each unique patient. All patients that did not have at least 10,000 cells collected in their sample were removed from the analysis.

Summary statistics for the dataset are provided here:

```
##    patient              CD45              PLCg2              CD19        
##  Length:115879      Min.   : -1.1313   Min.   :-1.1343   Min.   : -1.122  
##  Class :character   1st Qu.: -0.3058   1st Qu.:-0.7013   1st Qu.:  2.981  
##  Mode  :character   Median :  0.7924   Median :-0.4202   Median :  9.091  
##                     Mean   :  2.3708   Mean   :-0.3214   Mean   : 15.668  
##                     3rd Qu.:  3.0605   3rd Qu.:-0.1413   3rd Qu.: 20.814  
##                     Max.   :1617.0559   Max.   :82.6025   Max.   :399.320  
##       CD22              p4EBP1            Ikaros            CD79b        
##  Min.   :  -1.121   Min.   :-1.1270   Min.   :-1.113   Min.   :  -1.1276  
##  1st Qu.:  -0.385   1st Qu.:-0.3457   1st Qu.: 1.041   1st Qu.:  -0.4054  
##  Median :   0.344   Median : 0.4785   Median : 3.280   Median :   0.2837  
##  Mean   :   1.251   Mean   : 1.2535   Mean   : 4.768   Mean   :   0.8862  
##  3rd Qu.:   1.679   3rd Qu.: 1.9454   3rd Qu.: 6.852   3rd Qu.:   1.4534  
##  Max.   :3947.908   Max.   :36.0411   Max.   :85.211   Max.   :1238.4268  
##       CD20               CD34              CD179a              pSTAT5        
##  Min.   :  -1.1232   Min.   :  -1.108   Min.   :  -1.1302   Min.   :  -1.129  
##  1st Qu.:   0.1103   1st Qu.:  11.049   1st Qu.:  -0.6377   1st Qu.:  -0.580  
##  Median :   2.7838   Median :  32.155   Median :  -0.2929   Median :  -0.177  
##  Mean   :  11.0486   Mean   :  48.279   Mean   :  -0.0403   Mean   :   0.286  
##  3rd Qu.:  11.5035   3rd Qu.:  70.023   3rd Qu.:   0.1788   3rd Qu.:   0.592  
##  Max.   :2362.1838   Max.   :2956.034   Max.   :2791.2075   Max.   :4353.186  
##      CD123              Ki67              IgMi             IgL kappa       
##  Min.   : -1.1136   Min.   :  -1.120   Min.   :  -1.131   Min.   :  -1.1264  
##  1st Qu.: -0.2473   1st Qu.:  -0.319   1st Qu.:  -0.548   1st Qu.:  -0.5782  
##  Median :  0.7159   Median :   0.855   Median :  -0.110   Median :  -0.1744  
##  Mean   :  1.6628   Mean   :   6.304   Mean   :   1.494   Mean   :   0.2510  
##  3rd Qu.:  2.4637   3rd Qu.:   5.311   3rd Qu.:   0.755   3rd Qu.:   0.5888  
```

```
##  Max.   :256.0133   Max.   :7137.480   Max.   :6827.284   Max.   :1119.9121
##     IKAROS_i              CD10              CD179b             pAkt
##  Min.   : -1.11967   Min.   :  -1.076   Min.   :-1.1307   Min.   :-1.12993
##  1st Qu.: -0.00687   1st Qu.: 265.449   1st Qu.:-0.3599   1st Qu.:-0.65985
##  Median :  1.58254   Median : 412.521   Median : 0.3895   Median :-0.33451
##  Mean   :  8.69393   Mean   : 448.916   Mean   : 0.9981   Mean   :-0.10036
##  3rd Qu.:  9.76113   3rd Qu.: 592.952   3rd Qu.: 1.6056   3rd Qu.:-0.00698
##  Max.   :306.35855   Max.   :8207.047   Max.   :79.2710   Max.   :96.83366
##       CD24              CRLF2              CD127               RAG1
##  Min.   :   -1.089   Min.   :  -1.1240   Min.   :  -1.1319   Min.   :  -1.124
##  1st Qu.:  122.943   1st Qu.:  -0.4865   1st Qu.:  -0.6057   1st Qu.:  -0.635
##  Median :  249.645   Median :   0.0086   Median :  -0.2253   Median :  -0.288
##  Mean   :  354.999   Mean   :   0.5176   Mean   :   0.1687   Mean   :   0.257
##  3rd Qu.:  472.004   3rd Qu.:   0.9129   3rd Qu.:   0.4486   3rd Qu.:   0.240
##  Max.   :11929.121   Max.   :1926.6940   Max.   :1422.5455   Max.   :3502.190
##       Tdt               Pax5               pSyk               CD43
##  Min.   :  -1.121   Min.   : -1.120   Min.   : -1.12314   Min.   :  -1.104
##  1st Qu.:   0.368   1st Qu.:  4.077   1st Qu.: -0.61694   1st Qu.:  13.060
##  Median :   1.903   Median : 11.326   Median : -0.24565   Median :  37.675
##  Mean   :   3.327   Mean   : 16.081   Mean   :  0.06261   Mean   :  68.454
##  3rd Qu.:   4.642   3rd Qu.: 22.791   3rd Qu.:  0.38189   3rd Qu.:  86.343
##  Max.   :4339.835   Max.   :266.003   Max.   :141.63513   Max.   :8124.547
##       CD38               CD58              HIT3a               CD16
##  Min.   :  -1.1217   Min.   :  -1.112   Min.   :-1.13152   Min.   :-1.1275
##  1st Qu.:   0.5536   1st Qu.:   1.588   1st Qu.:-0.66915   1st Qu.:-0.4244
##  Median :   2.9406   Median :   4.601   Median :-0.35835   Median : 0.2393
##  Mean   :   9.7400   Mean   :   7.016   Mean   :-0.19297   Mean   : 0.8588
##  3rd Qu.:   9.2547   3rd Qu.:   9.658   3rd Qu.:-0.04804   3rd Qu.: 1.3881
##  Max.   :3106.6267   Max.   :2605.045   Max.   : 8.32404   Max.   :34.1563
##       pS6               pErk              HLADR               IgMs
##  Min.   :  -1.133   Min.   : -1.1196   Min.   :  -1.066   Min.   :  -1.1304
##  1st Qu.:   0.957   1st Qu.: -0.5621   1st Qu.:  68.323   1st Qu.:  -0.4188
##  Median :   3.473   Median : -0.1369   Median : 162.820   Median :   0.2217
##  Mean   :   9.350   Mean   :  0.2829   Mean   : 270.851   Mean   :   0.9521
##  3rd Qu.:   8.272   3rd Qu.:  0.6816   3rd Qu.: 354.250   3rd Qu.:   1.3034
##  Max.   :6388.095   Max.   :104.1688   Max.   :7318.328   Max.   :2265.0278
##      pCreb
##  Min.   :  -1.111
##  1st Qu.:   1.578
##  Median :   5.202
##  Mean   :   9.686
##  3rd Qu.:  12.550
##  Max.   :3152.514
```

Most important to note here is that, as is common with mass cytometry data (particularly in cancer), the distributions are highly skewed such that there are often huge(!) outliers in the positive direction due to instrumentation failure. These values are not biologically informative, so filtering out all measurements that are above the 95th percentile in a given channel was performed.

Results and Discusssion

Because SAUCIE is an unsupervised learning algorithm, we evaluated its performance by comparing it to the "gold-standard" supervised clustering algorithm that the authors applied to the same data in the original paper for which the data were collected. Specifically, we compared SAUCIE's performance to the original authors' algorithm using a version of the F1-measure of classification accuracy commonly used to compare

single-cell clustering methods to one another.[6] In short, the F1-measure is the harmonic mean of precision and recall for classification compared to a gold-standard method. SAUCIE performed with an F1-measure overall of 0.5, making it about average as far as clustering algorithms applied to mass cytometry datasets are concerned[6].

In addition to this overall metric, subpopulation-specific performance criteria are reported here:

Developmental Population	Precision	Recall	F-measure
Mature_Non_B	0.6305680	0.9788499	0.7670244
HSC	0.1011070	0.9563293	0.1828792
Progenitor_1	0.0692157	0.9669652	0.1291844
Early_Progenitors	0.0481128	0.9534779	0.0916032
Late_Progenitors	0.0349065	0.9765301	0.0674037
Pro_B2	0.0985915	0.0469003	0.0635634
Mature_B	0.0307879	0.9865886	0.0597124
Pre_B2	0.0254251	0.9835885	0.0495689
Progenitor_3	0.0174219	0.9824072	0.0342366
Immature_B2	0.0078320	0.9870535	0.0155408
Pre_Pro_B	0.0154867	0.0142373	0.0148357
Pro_B1	0.0067114	0.0266667	0.0107239
Progenitor_2	0.0050013	0.9890895	0.0099523
Pre_B1	0.0044248	0.0131004	0.0066152
Immature_B1	0.0011186	0.9966102	0.0022347

From these results, we can see that SAUCIE's best performance is on the population of cells called "Mature Non-B" cells, which also happens to be the largest (and most diverse) cell population in our dataset. Thus, it may be possible that SAUCIE is learning to identify cell populations better than others and more even sampling across subpopulations during training might increase its performance compared to the gold-standard we're using here.

Future directions

There are three concrete further directions that I want to carry out from what has been presented here.

- First, I want to run SAUCIE on the entire dataset (rather than just 10,000 cells per patient) now that I have some proof-of-principle that it works. If this does not improve the F-measure significantly, I will implement a biased sampling of the training set with adversarial examples (i.e. smaller cell populations on which SAUCIE is currently performing poorly).
- Second, I want to perform a more rigorous random search over the hyperparameters of the SAUCIE network to find more optimal values for its regularization constants. Here, I informally tested several values until I got a single-digit number of clusters (to make more comparable to the gold-standard) but a more rigorous approach is needed.
- Third, I am interested in using the reconstruction of the denoised features and the clustering information yielded by SAUCIE to see if adding an additional layer or two to the network will allow it to perform a supervised learning task (i.e. predicting which cells come from a patient who will relapse and patients who will not; successfully identifying cells that are cancerous compared to healthy cells, identifying cancer cells that come from an early timepoint of disease compared to a late timepoint of disease, etc.). In order to do this, I am hoping to add a convolutional layer to the end of the network that will automatically detect cellular subsets (even within SAUCIE-identified clusters) associated with the disease outcome. This approach has been used in one paper detailing an algorithm called "Cellular Convolutional Neural Network"[7] and I am curious how it might combine with SAUCIE.

References

1. Abou Dalle I, Jabbour E, Short NJ. Evaluation and management of measurable residual disease in acute lymphoblastic leukemia. Ther Adv Hematol. 2020;11:2040620720910023. Published 2020 Mar 6. doi:10.1177/2040620720910023
2. Wang XM. Advances and issues in flow cytometric detection of immunophenotypic changes and genomic rearrangements in acute pediatric leukemia. Transl Pediatr. 2014;3(2):149-155. doi:10.3978/j.issn.2224-4336.2014.03.06
3. Amodio, M., van Dijk, D., Srinivasan, K. et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods 16, 1139–1145 (2019). https://doi.org/10.1038/s41592-019-0576-7
4. Good Z, Sarno J, Jager A, et al. Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. Nat Med. 2018;24(4):474-483. doi:10.1038/nm.4505
5. https://en.wikipedia.org/wiki/Flow_Cytometry_Standard
6. Aghaeepour, N., Finak, G., Hoos, H. et al. Critical assessment of automated flow cytometry data analysis techniques. Nat Methods 10, 228–238 (2013). https://doi.org/10.1038/nmeth.2365
7. Arvaniti, E., Claassen, M. Sensitive detection of rare disease-associated cell subsets via representation learning. Nat Commun 8, 14825 (2017). https://doi.org/10.1038/ncomms14825

Appendix

Summary table of patient cell counts

patient	Number of cells
Healthy1	194185
Healthy2	1003294
Healthy3	2552004
Healthy4	662174
Healthy5	81935
UPN1	857516
UPN1-Relapse	10330
UPN10	62436
UPN10-Relapse	158774
UPN11	353739
UPN12	856763
UPN13	356801
UPN14	357327
UPN15	720760
UPN16	471291
UPN17	1054732
UPN18	216344
UPN19	701822
UPN2	310516
UPN20	640674
UPN21	162456
UPN22	34637
UPN22-Relapse	42027
UPN23	508632
UPN24	481026
UPN25	768576

patient	Number of cells
UPN26	351252
UPN27	782173
UPN28	340810
UPN29	722647
UPN3	161258
UPN30	765328
UPN31	377982
UPN35	3525
UPN35-Relapse	14228
UPN4	622846
UPN45	138423
UPN45-Relapse	61082
UPN47	255890
UPN48	306940
UPN49	468008
UPN5	659056
UPN50	388890
UPN51	298401
UPN52	370408
UPN53	564090
UPN54	545150
UPN55	500697
UPN56	401562
UPN57	134172
UPN58	226294
UPN6	752164
UPN60	121538
UPN60-Blood	197192
UPN61-Blood	111542
UPN62	243358
UPN62-Blood	61112
UPN63	214145
UPN63-Blood	51455
UPN64-Blood	12919
UPN65-Blood	50700
UPN67	62025
UPN68	79601
UPN69	227090
UPN7	908214
UPN8	730373
UPN9	870053
UPN90	447251
UPN90-Relapse	154082
UPN91	361951
UPN92	34814
UPN93	128414
UPN94	67792
UPN95	15756
UPN95-Relapse	173120
UPN96	921890
UPN97	47313
UPN98	353178