ML4FG - Pre-proposal

Mike Fu (xf2209), Chinyere Ihuegbu (coi2002), Amritha Musipatla (sm3773)

February 2021

0.1 What biological question will your method answer/enable answering?

Single-cell omics (such as scRNA-seq, scATAC-seq, CITE-seq) employ high-throughput technology to caputure the genomic and proteomic contents at the cellular level. Phenotypes are a product of the protein content of the cell, and the ratio of RNA: proteins is not 1:1; and the ratio of proteins to a phenotye is not always 1:1 in health and disease states. Thus while sc-omics provides a lot of information concerning the genomics, proteomics and diversity of cell types in health and disease states, it is limited in providing accurate information of the phenotypes in health and disease states; and also being able to predict phenotyspe based omics data without the wet-lab experiments (which are costly, time consuming and also limited)

0.2 ML method(s)/model(s)

In the Claassen paper, they used a simple 1-layer ConvNet. We seek to expand the architecture by adding additional layers. We also will consider transfer learning techniques from other single cell RNA models, as well as other neural networks from other fields of study. We also want to explore whether applying dimension reduction to the mass cytometry data will improve performance.

0.3 Data

PBMC, AML, ALL datasets, NK cell dataset:
`https://imsb.ethz.ch/research/claassen/Software/cellcnn.html`

0.4 Relevant Papers

[1] 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`