Machine Learning Prediction of CITE-seq Protein Expression from scRNA-seq Data

Kevin Hoffer-Hawlik[1], Quentin Chappat[1], Zachary Abessera[1], Crystal Shin[1]

[1]Department of Biomedical Engineering, Columbia University.

Abstract

Motivation: Recent advancements in single-cell genomics have enabled scaled multi-modal genomic measurements, but a major challenge we still face is predicting variability of downstream biological processes from upstream genomic data, such as cell surface protein expression (CITE-seq) from single-cell RNA expression data (scRNA-seq). In this report, we present machine learning models that can predict CITE-seq from scRNA-seq alone, which may improve the power of CD34+ HSPC subpopulation analysis from scRNA-seq without paired CITE-seq data.

Results: We initially constructed four machine learning model types incorporating feature selection using principal component analysis or selection of top correlated genes and with input data either from all four patients or from per-patient data. Our final model used singular value decomposition for feature selection and a LightGBM design, with input data from all four patients. Our models were trained and tested on paired scRNA-seq/CITE-seq data from CD34+ HSPCs from four healthy donors. Our final model predicted CITE-seq for 140 target proteins with a custom correlation score 0.878 and mean squared error of 2.75. Comparison to top performing models will require testing on a held-out dataset to be published in the future.

Availability: All code and analysis is available in the public GitHub repository: https://github.com/qchappat/ECBME-4060-2022-Project-Kaggle-Open-Problems-Multimodal-Single-Cell-Integration

1 Introduction

Recent advancements in single-cell genomics have enabled us to take simultaneous measurements of multiple genomic modalities in single cells. However, data analysis methods to synthesize data and uncover dynamic biological processes still need to be studied further. One of the major challenges that we face is predicting variability of cell circuitry and downstream biological processes from upstream genomic data. When analyzing single-cell data, it is important to not only focus on individual feature spaces, but also to study the shared and unique variations between different modalities and batches (Lähnemann, 2020).

The overarching goal of the referenced Kaggle challenge (Open Problems in Single-Cell Analysis) was to collectively devise methods to map genetic information across layers of cellular state and predict one genetic modality using another modality (i.e., transcriptomic from genomic, and proteomic from transcriptomic data), which could lead to a better understanding of complex regulatory processes in cells. We focused on the second challenge, which was to predict how single cell RNA and cell surface protein measurements co-vary as bone marrow stem cells mature into various types of blood cells. The challenge used a dataset consisting of CD34+ hematopoietic stem and precursor cells, a rare self-renewing cell type that can differentiate into every type of mature blood cell with a well-defined hierarchy (AbuSamra, 2017). Since these are transcriptionally similar cells, understanding both RNA and protein expression data (CITE-seq) is vital to deconvolve their cell identities. In this report, we present multiple machine learning models that can predict CITE-seq protein expression (100 – 200 features per cell) from scRNA-seq alone (>10,000 features per cell) for a given cell type, which may subsequently improve the power of cell subpopulation analysis from scRNA-seq without paired CITE-seq data.

2 Methods

2.1 Model Designs:

In creating a method to predict CITE-seq using scRNA-seq, we desired to evaluate whether a generalizable machine learning model could be applied across individual patients or across all patients for different proteins, and which approach would have greater performance and, therefore, greater biological significance. Additionally, we desired to incorporate genes that did not encode for a given protein as those transcripts could still have predictive power on protein expression (e.g., protein translation programs upstream of a given protein, or proteins/genes with related biological pathways). However, given the massive number of scRNA-seq features and the expected impact on required training time, we incorporated feature selection into each model through principal component analysis (PCA), selection of top correlated genes, or singular value decomposition (SVD).

In the end, we created five separate model types with different designs and inputs, with the final model informed by results from the first four:

(1) Models taking input scRNA-seq features consisting of a) genes encoding proteins of interest and b) top principle components (PCs) learned from genes not directly encoding proteins of interest; and trained on data from all patients ("PCA All")

(2) Models using a Pearson correlation matrix to select top correlated genes for each protein; and trained on data from all patients ("Pearson-R All")

(3) Models taking similar input scRNA-seq features as "PCA All" (encoding genes and top PCs from non-encoding genes), but trained on per-patient datasets ("PCA Per-Patient")

(4) Models taking similar input scRNA-seq features as "Pearson-R All" (selecting for top correlated genes), but trained on per-patient datasets ("Pearson-R Per-Patient")

(5) LightGBM models taking input scRNA-seq features consisting of a) genes directly encoding to proteins of interest and b) top features learned from SVD on non-encoding genes; and trained on data from all patients ("SVD + LightGBM")

For the first four model types, we leveraged AutoGluon, an automated machine learning model toolkit allowing for quick prototyping of three distinct machine learning models (LightGBM, XGBoost, and CAT regression), and to rapidly optimize these model architectures with preliminarily optimized hyperparameters (Erickson, 2020). For the fifth model type, we specifically trained LightGBM models, as the first four AutoGluon models had greatest efficiency trade-offs with LightGBM designs (i.e., highest result accuracy compared to training time). We implemented LightGBM models with 10 max depth, 100 max leaves in a tree, and minimum child samples of 250 (Guolin, 2017).

For each of the PCA models, the top 500 PCs were used for feature selection. For the Pearson correlation matrix models, the top 10 correlated genes were chosen for feature selection. For the SVD model, we set the output dimensionality to 512. The choice of PCA dimensions, top number of correlated genes, and SVD dimensions was informed by the top entries in the Kaggle competition.

2.2 Training Methodology

For the all-patient model types ("PCA All" and "Pearson-R All"), 140 models were trained using every patient's data, with each model predicting CITE-seq levels for one given protein (for 140 CITE-seq predictions in total). For the per-patient model types ("PCA Per-Patient" and "Pearson-R Per-Patient"), 140 models for each protein were trained for three patients, with each model predicting CITE-seq levels for one given protein for one given patient (for 420 models across the three patients, and 420 CITE-seq predictions in total). The four AutoGluon model types were trained using AutoGluon's "Best Quality" preset to optimize for most hyperparameters (e.g., learning rate). Root mean squared error was used as the loss function for training. We chose an 80/20 training/testing split of our dataset; for all-patient models we used 8-fold cross-validation, and for per-patient models we used 4-fold cross-validation.

For the "SVD + LightGBM" model types, 140 models were trained using every patient's data. L2 loss (squared error loss) was used as the loss function for training. Learning rate was set to 0.1. We chose a 85/15 training/testing split of our data and 3-fold cross-validation, with the folds based on donor identities.

2.3 Data and Feature Selection

Our dataset consisted of paired scRNA-seq and CITE-seq data from 70,988 unique CD34+ hematopoietic stem and precursor cells (HSPCs) derived from four healthy donors, from the original Kaggle competition dataset of ~280,000 HSPCs. Our input data consisted of scRNA-seq (22,050 gene expression features for each cell), and our prediction data was the paired CITE-seq (140 protein expression features for each cell), although our input dataset was scarce with ~78% of scRNA-seq features having zero values. Nonetheless, few preprocessing steps were required for the data as it was already cleaned and normalized, with no missing values. The only step performed was removing scRNA-seq features with non-existent gene expression (i.e., genes that were not expressed in any cells in the train or test datasets and thus have limited predictive utility), or around 449 of the scRNA-seq features.

Prior to developing our models, we performed exploratory data analysis on the full, unpaired scRNA-seq dataset (~120,000 unique HSPCs) (**Fig. 1**) and the CITE-seq dataset. The HSCPs were evenly distributed across donors. Slightly more samples were collected on the third collection day. Lastly, cells were labeled mostly as hematopoietic stem cells; very few were labeled as B cell progenitors or monocyte progenitors. Analysis of ~20 genes' expression level distributions showed that most genes had normally distributed expression across cell samples. However, analysis of ~20 proteins' expression level distributions suggested that protein expression had diverse distributions, with many normal, some multimodal, and several noisy outliers.

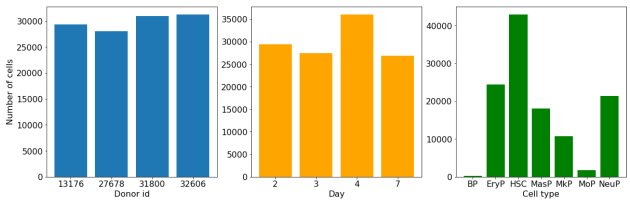

Fig. 1. Distribution of CD34+ hematopoietic stem and precursor cells by metadata variables. (Left) Number of cells by donor identity. (Middle) Number of cells by day of sample collection. (Right) Number of cells by annotated cell type. Note: EDA was performed on the full scRNA-seq dataset, although a subset (~71,000) of cells were used for model purposes given the requirement of paired CITE-seq data.

Finally, we implemented feature selection to mitigate computational costs required to train models at scale for each protein. We expected most of the ~22,000 genetic features do not encode for the 140 protein targets, although we hypothesized that a subset of scRNA-seq genes may be directly linked to CITE-seq target proteins. This would allow us to preserve putative encoding genes while performing feature selection on putative non-encoding genes. Specifically, to identify encoding genes, we identified the set of genes for which any of the 140 protein names were contained within the gene name (149 genes in total). We subsequently defined all other genes as non-encoding genes to be used in feature selection steps for each model.

3 Results

The Kaggle competition scored model performance using a correlation score, defined by the average of each sample's Pearson correlation coefficient $\rho_{\hat{Y},Y}$ and calculated by:

$$\rho_{\hat{Y},Y} = \frac{cov(\hat{Y},Y)}{\sigma_{\hat{Y}}\sigma_{Y}} \quad (1)$$

where $cov(\hat{Y},Y)$ is the covariance between the predicted and ground truth CITE-seq expression level, and σ is the standard deviation of the predicted and ground truth expression. However, as the final test dataset (corresponding to a later sample collection timepoint) was not publicly available at the time of the project, we cannot directly compare correlation scores to the full leaderboard, although we can get a general sense of comparative performance. Additionally, we internally evaluated our models using a mean-squared-error (MSE) metric.

We first assessed our models' performances using MSE and the correlation score calculated on the withheld test dataset (**Fig. 2**). MSE was lowest for "SVD + LightGBM" model type (~2.75), and lower for per-patient model types compared to all-patient model types (2.79 and 2.81 vs. 2.88 and 2.90). However, while correlation score was comparable between the "SVD + LightGBM" model type and the all-patient model types (0.878 vs. 0.870 and 0.875), it was substantially lower for per-patient model types (0.362 and 0.411). We hypothesize per-patient model correlation scores suffer significantly because the number of observations in the training datasets is effectively quartered when creating per-patient models (despite the number of features staying the same). Taken together, it is difficult to conclude how patient-specific genetic information is in the context of protein expression prediction; further experiments should control for training data size to conclude whether a single model across all patients or one model per patient is more accurate.

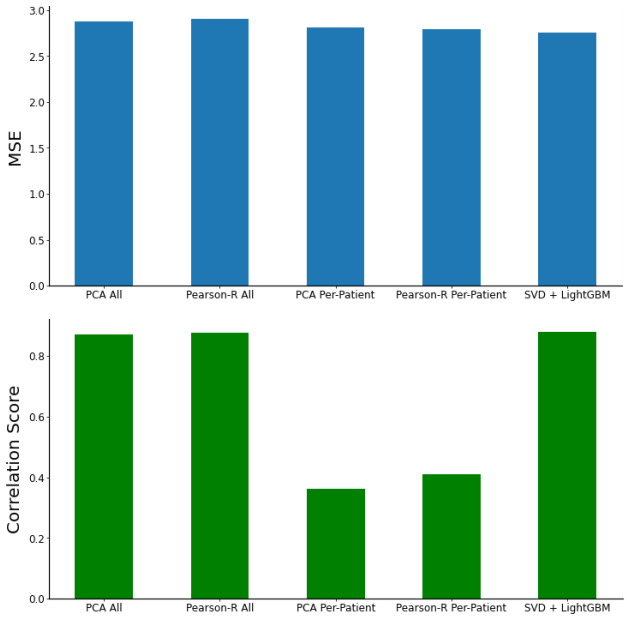

Fig. 2. Comparison of model performance evaluated on test data. Note: correlation scores from the public Kaggle leaderboard are not plotted as our models tested on a different test dataset.

Lastly, we briefly and indirectly compare our models' correlation scores to the Kaggle leaderboard. At the time of this report, the top entries in the public leaderboard have scores 0.815 - 0.816, while our correlation score is marginally higher at 0.878 for our top-performing "SVD + LightGBM" model type. As stated before, our test dataset was not the same as the private test dataset used for Kaggle submissions. Our correlation score may be higher because our test dataset is likely more similar to the training dataset than the private test dataset, given the private test dataset on Kaggle consists of samples collected at a later fifth timepoint. Thus, we cannot conclude whether our model types perform better than leading models on Kaggle. Nonetheless, 0.816 top correlation score for the public dataset and 0.776 top correlation score for the private dataset suggests that further investigation is needed to improve methods to accurately map genomic information across layers of single-cell data.

Discussion

We successfully developed five different machine learning models to predict CITE-seq from scRNA-seq data and aimed to evaluate 1) whether models using data from all patients or from one patient at a time would have greatest predictive performance (i.e., the importance of patient-specific genetic information) and 2) different feature selection strategies such as PCA, top correlated genes, and SVD. Our final model type incorporated a LightGBM design with feature selection using SVD, with hyperparameters and feature selection parameters informed by experiments with our first four models and the models submitted on the original Kaggle competition. We trained and evaluated our models using a CD34+ HSPC dataset with paired scRNA-seq and CITE-seq data and compared MSE and a custom correlation score metric to determine which models had the best performance. The five models had similar MSEs, though MSE for the LightGBM/SVD model was lowest. However, correlation scores were significantly higher for those models that have trained on data from all patients, possibly due to per-patient models training on significantly smaller datasets to perform per-patient CITE-seq predictions. As the dataset used to test the Kaggle leaderboard submissions was not publicly available at the time of this report, we could not directly compare accuracy, although our models had marginally higher correlation scores (likely due to greater similarity between our training and test datasets, compared to our training dataset and the private test dataset on Kaggle). Our findings suggest that, while it could be further improved, our machine learning method can learn both direct and indirect relationships between genetic features and downstream cell-surface protein expression, with indirect relationships learned after feature selection.

Beyond additional validation, we suggest further investigation into improving feature selection and model tuning. For example, a more nuanced feature selection method (e.g., assigning scores to genes using known protein-protein, gene-gene, or protein-gene relationships, and then selecting for top scores) could alleviate computational costs to train models but still preserve biologically significant relationships between target proteins and non-encoding genes, although this approach would require a comprehensive review of literature describing the protein targets in our CD34+ HSPC single-cell dataset. Additionally, while AutoGluon does a reasonably good job of optimizing hyperparameters for a given machine learning method, future rigorous experiments could better optimize hyperparameters for the LightGBM model we developed.

Although our study focused on predicting CITE-seq from scRNA-seq derived from CD34+ HSPCs, knowledge gained from this project and similar studies could be applied to research in other areas of genomic biology. One relevant topic could be to address the first challenge of the original Kaggle competition, to predict gene expression scRNA-seq from chromatin accessibility ATAC-seq using model designs similar to our developed models. More generally, this project underscores the value of machine learning in the field of biology and genomics – even with limited computing resources and dataset sizes, we were able to obtain meaningful insights into predicting variability of cell circuitry and biological mechanisms from genomic information. In sum, our machine learning models show promise in addressing the crucial challenge of uncovering and mapping biological processes across genomic layers, which would be of great scientific value to the field of genomics and computational biology as single-cell study data continues to expand.

Acknowledgements

We would like to thank Prof. Wei-Yi Cheng for his support throughout the project, as well as his guidance throughout the entire ECBM E4000 class. All code and analyses are available in the public Github repository: https://github.com/qchappat/ECBME-4060-2022-Project-Kaggle-Open-Problems-Multimodal-Single-Cell-Integration

References

AbuSamra, D. (2017) Not just a marker: CD34 on human hematopoietic stem/progenitor cells dominates vascular selectin binding along with CD44. *Blood Adv.* https://doi:10.1182/bloodadvances.2017004317

Erickson, N. (2020) AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint.* https://arxiv.org/abs/2003.06505

Gry, M. (2009) Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics*. https://doi.org/10.1186/1471-2164-10-365

Guolin K. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Lähnemann, D. (2020) Eleven grand challenges in single-cell data science. *Genome Biol*. https://doi.org/10.1186/s13059-020-1926-6

Open Problems in Single-Cell Analysis - Multimodal Single-Cell Integration Competition. https://www.kaggle.com/competitions/open-problems-multimodal