PROJECT REPORT

SingleCell Multiomics Data Analysis: Developing Predictive Models for CITEseq Data

Prepared by: Bibit Waluyo Aji

Submitted to Dr. Boxiang Liu, Yong Loo Lin School of Medicine, NUS

1. Introduction

The field of single-cell genomics has undergone significant advancements over the past decade, allowing researchers to measure and analyze DNA, RNA, and protein expressions at the single-cell level. This granularity has provided unprecedented insights into cellular behavior and the intricate processes governing cell differentiation, development, and disease progression. Single-cell multiomics, which involves simultaneous measurement of multiple genomic modalities in the same cell, is particularly powerful as it captures the complex regulatory mechanisms underlying cellular functions.

Among the many applications of single-cell genomics is the study of hematopoiesis—the process by which hematopoietic stem and progenitor cells (HSPCs) in the bone marrow develop into various types of blood cells. Understanding the regulation and differentiation pathways of HSPCs is crucial for advancing treatments for blood disorders, improving bone marrow transplants, and developing targeted therapies for hematological malignancies.

In this project, we aim to develop predictive models to infer RNA expression from chromatin accessibility (DNA) and protein levels from RNA expression using CITE-seq data. Specifically, the task involves using a dataset of CD34+ HSPCs collected at different time points from four human donors. This predictive modeling is challenging due to the inherent sparsity and noise in single-cell data, the complexity of biological regulatory networks, and the temporal dynamics of cell differentiation. We compare the performance of various models, including K-Nearest Neighbors (KNN), LightGBM, XGBoost, and CatBoost, to determine the most effective approach for this task.

2. Objectives

The primary objectives of this miniproject are as follows:

2.1. Developing Predictive Models for CITEseq Data.

2.2. Compare the performance of various models, including KNearest Neighbors (KNN), LightGBM, and XGBoost.

3. Model Descriptions

This section provides a detailed description of the machine learning models and preprocessing techniques utilized in this project, including K-Nearest Neighbors (K-NN), LightGBM, XGBoost, and Catboost. Additionally, Principal Component Analysis (PCA) was employed as a dimensionality reduction technique to enhance model performance.

Before feeding the data into the machine learning models, PCA was applied to reduce the dimensionality of the dataset. PCA is a statistical procedure that transforms the original features into a set of linearly uncorrelated components, ordered by the amount of variance they capture from the data. For this project, PCA was configured to reduce the dataset to 100 components.

K-Nearest Neighbors (K-NN) is a straightforward, non-parametric method used for both classification and regression tasks. In this study, the K-NN model was set up to consider the 5 nearest neighbors, using the Manhattan distance metric to measure proximity. This model works by identifying the k-nearest data points to a given input and predicting the output based on the majority class (for classification) or average (for regression) of these neighbors.

LightGBM (Light Gradient Boosting Machine) is a highly efficient and scalable implementation of gradient boosting, particularly designed for high-performance, large-scale data processing. In this project, the LightGBM model was configured with 1000 estimators and a learning rate of 0.1. Additionally, the model used the Mean Absolute Error (MAE) metric for evaluation, was initialized with a seed value of 42 for reproducibility, and included regularization parameters with alpha set to 0.0014 and lambda to 0.2. The model was also set to use 80% of features for each tree (colsample_bytree), a subsample rate of 50%, a maximum depth of 10, 722 leaves per tree, and a minimum of 83 samples required in each leaf.

XGBoost (Extreme Gradient Boosting) is a robust gradient boosting framework optimized for speed and performance. For this study, the XGBoost model was set up with 1000 estimators and a learning rate of 0.1. It was configured to use the squared error as the objective function for regression tasks and the MAE for evaluation. The model parameters included an alpha regularization term of 0.0014, a lambda regularization term of 0.2, and a feature subsample rate of 80%. Additionally, it was set to use 50% of the data for each boosting round (subsample), a maximum depth of 10, and a minimum child weight of 83.

CatBoost (Categorical Boosting) is a gradient boosting library tailored to handle categorical features without extensive preprocessing. In this project, the CatBoost model was set to run for 1000 iterations with a learning rate of 0.1. The model was initialized with a random seed of 42 and included a regularization term (l2_leaf_reg) set to 0.2. Other parameters included a maximum depth of 10, an 80% feature subsample rate (rsm), a 50% subsample rate, a minimum of 83 data points in each leaf, a bagging temperature of 1, random strength of 1, a one-hot encoding threshold of 2, and used the Newton method for leaf value estimation.

4. Result and Discussion

This section evaluates the performance of four machine learning models: KNearest Neighbors (KNN), LightGBM, XGBoost, and Catboost. These models were chosen for their potential to handle multioutput regression tasks, crucial for predicting gene expression and protein levels from chromatin accessibility and gene expression data.

Table 1. Training model performance on citeseq data

Model	R2	MSE
KNN	0.3883	2.0286
LightGBM	0.9811	0.0535
XGBoost	0.9520	0.1476
CatBoost	0.8927	0.3367
Public Benchmark (LightGBM)	-	0.0203

From the results, it is evident that gradient boosting models, particularly LightGBM and XGBoost, significantly outperform KNN and CatBoost in predicting protein levels from RNA expression data. The LightGBM model achieved the highest R^2 score of 0.9811 and the lowest MSE of 0.0535, indicating a strong fit to the data. XGBoost also performed well with an R^2 score of 0.9520 and an MSE of 0.1476, showcasing its capability in handling complex data structures. CatBoost, while designed to handle categorical data efficiently, did not perform as well as LightGBM and XGBoost, achieving an R^2 score of 0.8927 and an MSE of 0.3367. The KNN model showed the least performance with an R^2 score of 0.3883 and an MSE of 2.0286, demonstrating the limitations of non-parametric methods in this context.

The public benchmark model, a LightGBM configuration, had an MSE of 0.0203. While our LightGBM model did not quite reach this benchmark, with an MSE of 0.0535, it nonetheless demonstrated substantial predictive capability, outperforming other models tested in this study. The slight performance gap might be attributed to differences in hyperparameter tuning, feature engineering, or preprocessing steps that were not identical to those used in the benchmark model.

Overall, the analysis demonstrates that gradient boosting methods, particularly LightGBM, are highly effective for predicting RNA expression from chromatin accessibility data. LightGBM's ability to handle high-dimensional data efficiently and its sophisticated boosting algorithms make it a standout choice for bioinformatics applications. XGBoost also offers robust performance and serves as a strong alternative, while CatBoost remains a viable option depending on the nature of the data. In contrast, K-NN’s performance highlights its limitations in handling complex, high-dimensional datasets common in biological research.

5. Conclusion

The study demonstrated that advanced gradient boosting models, particularly LightGBM and XGBoost, significantly outperformed KNN and CatBoost in predicting RNA expression from chromatin accessibility (DNA) and protein levels from RNA expression using CITE-seq data. LightGBM achieved the highest performance with an R^2 score of 0.9811 and an MSE of 0.0535, while XGBoost also performed well with an R^2 score of 0.9520 and an MSE of 0.1476. These results underscore the efficacy of gradient boosting techniques in handling the complexity and noise inherent in single-cell data, thereby providing robust predictive frameworks for single-cell multiomics analysis.