Project Proposal

● Ungraded

Group

Riti Bhatia
Sanchitha Kuthethoor
Talha Ahmad Khan
...and 1 more
View or edit group

Total Points

- / 1 pts

Question 1

Project Proposal 1 pt

Question assigned to the following page: 1

Leveraging Landmark Genes for Genome-Wide Gene Expression Prediction: A Comparison of Machine Learning Methods

Riti Bhatia, Sanchitha Kuthethoor, Sumeet Kothare, Talha Khan

Problem Statement and Questions to Address

The goal of this project is to predict the expression levels of a set of genes based on the expression of a smaller subset of landmark genes, using different machine learning methods in a multi output regression setting. Previous research has focused on predicting expression for large numbers of genes (such as 1000 landmark genes scaled to 21,000 genes). In our project, keeping in mind the computational and resource constraints we may face, a few ideas we had as workarounds were:

a. Chunking the problem: Using smaller disjoint subsets, such as using the 1000 landmark genes to predict the expression of a set of 100 genes at a time, iterating this process to predict a larger overall set of genes.
b. Using PCA for dimensionality reduction on the input genes

We believe this is an interesting and relevant problem because gene expression data plays a crucial role in understanding cell state, cell type, disease progression, and other important biological inferences which have implications in research and medicine. Sequencing and analyzing the expression of levels of tens of thousands of genes can be computationally and resource intensive. This is why, if the expression levels of all genes can be determined by a considerably smaller subset of landmark genes, we can simplify the process and reduce the storage, resource, and computational demands. This will allow us to efficiently make gene expression predictions while still retaining valuable biological insights.

Dataset

The dataset is compiled from multiple sources such as Gene Expression Omnibus (GEO), GTEx and the 1000 Genomes project. GEO provides 130,000 expression profiles across 22,000 probes. GTEx gives us tissue or cell type specific expressions using Illumina RNA-seq technology. The 1000 genomes expression dataset gives us an expression profile of lymphoblastoid cell lines using RNA-seq technology. As these databases operate on different units, the merged dataset is quantile-normalized to allow maximum information exchange between the three resources. Overall, our final preprocessed dataset gives us an expression catalog of 1000 landmark and 21000 target genes.

Literature From the Field

Question assigned to the following page: 1

Chen et al. (2016) in their study on gene expression inference with deep learning demonstrated the effectiveness of deep learning models in predicting gene expression by outperforming traditional methods like linear regression. Building on this, we implement k-Nearest Neighbors (k-NN) and Multivariate Linear Regression (MLR) as baseline models in our study, similar to the non-parametric methods explored by Chen et al. While their work highlights the potential of deep learning in capturing complex, non-linear relationships between gene expressions, our project aims to evaluate a broader range of machine learning methods, including neural networks and ensemble models, to predict the expression levels of a large set of genes using a smaller subset of landmark genes. By comparing these models, we seek to understand the trade-offs between model complexity and prediction accuracy in gene expression analysis.

- Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, Xiaohui Xie, Gene expression inference with deep learning, Bioinformatics, Volume 32, Issue 12, June 2016, Pages 1832–1839

Our project involves predicting gene expression levels from a subset of landmark genes, which requires handling high-dimensional data. Although our focus is on bulk gene expression, methods like scVI, which use deep generative models for single-cell transcriptomics, highlight the potential of advanced modeling techniques in gene expression analysis. We can use approaches similar to scVI's approach to dimensionality reduction and probabilistic modeling to enhance our own methods, to possibly improve the accuracy and robustness of our predictions.

- Lopez, R., Regier, J., Cole, M.B. et al. Deep generative modeling for single-cell transcriptomics. Nat Methods 15, 1053–1058 (2018). https://doi.org/10.1038/s41592-018-0229-2

Methods

a. ***k*-Nearest Neighbors (*k*-NN)**: This will be our baseline non-parametric method. This model will be based on the assumption that genes with similar expression profiles relative to the 1000 landmark genes will have similar expression levels across the full set of 21,000 genes being predicted. This method will not be able to model complex relationships, but will be a starting point to use as comparison for more sophisticated models that we will employ.
b. **Multivariate Linear Regression**: To build on the baseline model, we will implement Multivariate Linear regression. This will assume linear relationships between the landmark gene expression and the expression of the target genes. This will offer a more structured and parametric approach and will be able to capture simple, linear patterns in our data.
c. **Neural Networks**: To capture more complex, non-linear relationships between genes, we will explore neural networks, which should be able to capture more intricate patterns

Question assigned to the following page: 1

that may not be highlighted using traditional machine learning methods. This could help improve prediction accuracy.

d. **Model from a Package** (Random Forests, XGBoost): Finally, we plan to implement Random Forests using established machine learning packages. This is a method that is well-suited for handling high-dimensional data and capturing nonlinear interactions.

By starting with more simple models and advancing to mode complex models that are able to capture more complex relationships, we aim to evaluate the trade-offs between model complexity and performance/accuracy in predicting gene expression.

Model Evaluation and Analysis Metrics

We aim to split the data into an approximate 80:20 fraction to create the training and test datasets. After training, we will evaluate the model on the test dataset and employ evaluation metrics such as Accuracy, Precision, Recall, and F1 Score to study the model's performance. Depending on the evaluation metrics, we may employ cross-validation by splitting the dataset, for example, into 5 random fractions and train and test the model progressively; this will allow us to gauge the increment in performance due to cross-validation. Overall, we will compare the different models' accuracies and metrics to paint a holistic picture of model fitness.