Fine-tuning Geneformer to distinguish Parkinson’s cells from normal control cells

Introduction

Geneformer [1] is a foundational biological model trained on 30 million single cell transcriptomes from various tissues and cell types. The model is pre-trained at scale to discern a representation of healthy cell transcriptomes across many different contexts. Transfer learning using foundational models like Geneformer has shown to be an effective method for using the prior information of contextual gene network structure for downstream tasks with limited task-specific data. To validate this claim, we fine-tuned Geneformer to distinguish between Parkinson’s-afflicted and normal control oligodendrocyte single cell transcriptomes.

Methods

We extracted single-cell RNA sequencing data from a Parkinson’s disease study profiled in the substantia nigra [2]. After filtering steps, our final dataset included ~143,000 cells split between Parkinson’s (n=7) and normal control (n=11) donors. We converted the dataset into the Anndata (.h5ad) format [3] for compatibility with the Tokenizer function specific to Geneformer. Once the single-cell data was tokenized, we fine-tuned Geneformer on the cell classification task of differentiating between Parkinson’s afflicted oligodendrocytes and normal control oligodendrocytes. We initialized the model with the pretrained Geneformer weights with the addition of a final task-specific transformer layer. We opted for hyperparameters recommended by the authors for a comparable task: learning rate: 8e-4; learning scheduler: polynomial; optimizer: Adam with weight decay fix; warm up steps: 1812; weight decay: 0.25; batch size: 12. According to the Geneformer author’s [1], the number of frozen layers should depend on how similar your fine-tuning task is to the original pre-training objective. Because our task was fairly dissimilar, we opted to freeze 2 of the 6 layers.

Results

Fine-tuned Geneformer showed an AUROC of 0.85 and area under the precision recall curve of 0.78 compared to simpler baselines of Logistic Regression (AUROC=0.5, AUPRC=0.58) and a neural network with two hidden layers (AUROC=0.5, AUPRC=0.61).

Discussion

We fine-tuned Geneformer to classify Parkinson’s afflicted transcriptomes. We were able to use the prior embedded knowledge regarding gene network structure in the healthy transcriptome of single cells to classify the diseased state even with limited task-specific data. Our fine-tuned Geneformer outperformed simpler baselines on the basis of AUROC and AUPRC. It should be noted that the baseline models were trained on padded 2048 gene set inputs and did not include any hyperparameter optimization. While fine-tuned Geneformer outperformed the baselines, the model still suffered from a high false positive rate and struggled with classifying normal control samples, with 68% of control samples from the test set being incorrectly labeled as Parkinson's. To increase Geneformer performance we may opt to run more rigorous hyperparameter optimization and increase the fidelity of the fine-tuning dataset. With a higher performing fine-tuned Geneformer model in the future, it may be interesting to run ablation experiments through the removal of genes from the input transcriptomes of healthy single cells. Through these experiments, we may be able to discover novel genes that drive the progression of cells from a healthy state to a Parkinson's state.

Literature Cited

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10949956/#SD2
[2] https://www.biorxiv.org/content/10.1101/2021.06.16.448661v1.full
[3] https://www.biorxiv.org/content/10.1101/2021.12.16.473007v1

Other Links

Geneformer repo: https://huggingface.co/ctheodoris/Geneformer
Parkinson's single-cell dataset:
https://singlecell.broadinstitute.org/single_cell/study/SCP1402/a-molecular-census-of-midbrain-dopaminergic-neurons-in-parkinsons-disease-preprint-data?cluster=Olig&spatialGroups=--&annotation=Cell_Type--group--cluster&subsample=100000#study-summary