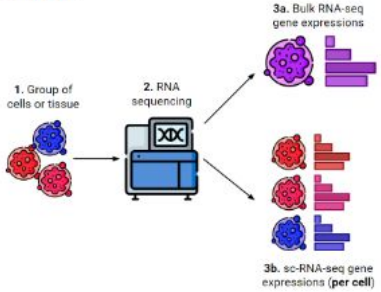
AS A CELL, IS IT BETTER TO BE SINGLE?

Alan Kuźnicki (a.l.kuznicki@student.tudelft.nl)
Supervisors: Prof. dr. ir. Marcel Reinders, Niek Brouwer

TU Delft

1. Introduction

Geneformer [1] is a model that can be used in a variety of biomedical applications, such as predicting the drug that a cell had been treated with. The model is trained and fine-tuned on gene expression data. Such data is obtained through RNA sequencing. When each cell is sequenced individually, the process is known as single-cell RNA sequencing (scRNA-seq.) The alternative, bulk sequencing, is applied to groups of cells. Bulk data is generally more widely available and easier to acquire; hence, it could be beneficial to find a way to use it to fine-tune Geneformer.



[2]

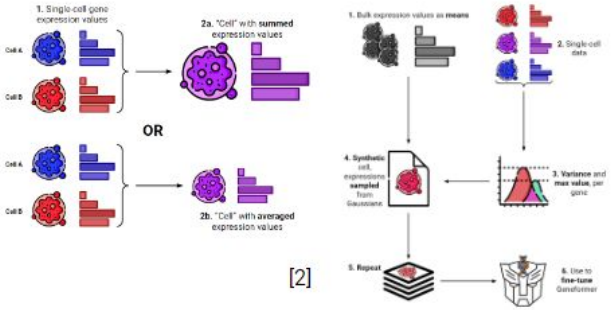
The aims are to ascertain whether bulk RNA-seq data can be used to fine-tune Geneformer, and if so, to what extent, to explore the feasibility of extracting more suitable data from a bulk dataset, and to verify whether both types of sequencing data can be used together to fine-tune more effectively.

2. Methodology

The experiments are divided into two main batches. In the first, pseudo-bulk data is generated through aggregation. In the second, synthetic single-cell data is obtained from bulk.

A representative 10% of the single-cell data is set aside as a benchmark/test dataset. The rest becomes the training dataset that is processed further.

An aggregation factor **k** is introduced, representing the average number of cells within a group that is aggregated into one data point within the pseudo-bulk dataset. Actual group sizes are normally distributed around **k**. Within each group, gene expression values are aggregated according to three strategies: summing, averaging with **k**, averaging with exact group size.

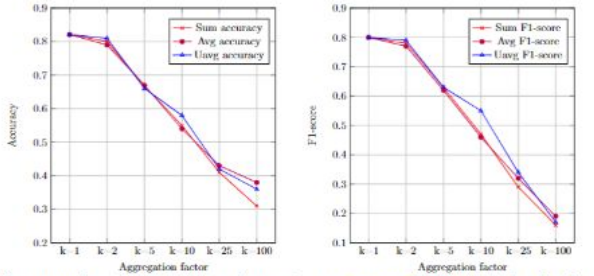


[2]

Bulk data is used as the basis for synthetic data points, generated by aggregating each label-dose pair into one data point. The variance and max value of each gene are calculated based on single-cell data. Synthetic cells are created by sampling gene expressions from a Gaussian with a bulk mean and single-cell variance. Resampled if below 0 or above max.

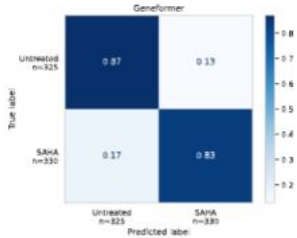
Dataset used: Sciplex2 [3]

3. Results



The results show that there is a steep decline in both the accuracy and F1-score of the model when fine-tuned on data of increasing bulkiness. Especially for the most bulky explored datasets, the validation scores during training were high, suggesting that the above performances are about the upper limit of what Geneformer can achieve by being fine-tuned on them. An attempt to introduce some single-cell data to the validation set did not improve these results meaningfully.

The generated synthetic data shows some potential in a simple two-label problem, where it succeeded in fine-tuning Geneformer effectively. This did not hold for more complex (e.g. five-label) problems, where accuracies did not exceed 0.45-0.50.



	Untreated	Nutlin	Dex	BMS	SAHA	Overall?
Baseline	0.68	0.73	0.94	0.70	0.92	
Trial 1	0.62	0.68	0.93	0.66	0.88	
Trial 2	0.64	0.72	0.94	0.73	0.90	
Trial 3	0.57	0.81	0.92	0.73	0.93	
Trial 4	0.61	0.79	0.94	0.76	0.91	?

Adding some generated data (BMS and Untreated) to a single-cell training set and fine-tuning Geneformer did not yield meaningfully different results on the five-label problem. Some changes in individual classes' accuracies can be observed, but these are inconsistent and most likely caused by the randomness of the fine-tuning.

4. Conclusions

Bulk gene expression data cannot be effectively used to directly fine-tune Geneformer for cell classification.

Generated synthetic data shows minor promise, but more sophisticated generation methods need to be explored to assess its true potential.

Mixing synthetic and real single-cell did not meaningfully affect performance, suggesting that this approach to using bulk data in fine-tuning may not be effective. Alternative approaches like augmenting the single-cell data based on bulk should be explored.

References

[1] C. Theodoris, L. Xiao, A. Chopra, *et al.*, "Transfer learning enables predictions in network biology," *Nature*, vol. 618, pp. 616–624, 2023. doi: https://doi.org/10.1038/s41586-023-06139-9.
[2] These figures were created using images from Flatikon.com
[3] S. R. Srivatsan *et al.*, "Massively multiplex chemical transcriptomics at single-cell resolution," *Science*, vol. 367, pp. 45–51, 2020. doi: https://doi.org/10.1126/science.aax6234.