Homework 6. DL Reading

Mikhail Ushakov

a. What is the problem the article's trying to address?

Authors are trying to solve the critical problem of predicting cancer drug responses at the single-cell level in the context of cellular heterogeneity, when different cancer cell subpopulations can respond differently to the same treatment.

b. What are the related works in the field and why there is still a need to propose new solutions?

It is mentioned that existing drug-response prediction methods for bulk-data are not suitable for single-cell data. Development of deep learning tools for scRNA-seq data expanded our capabilities and achieved high performance in gene expression imputation, cell clustering, batch correction, and similar tasks. Previous studies utilized DTL for cancer drug response prediction based on bulk data sources. However, DTL effectiveness in transferring bulk-level knowledge to the single-cell level is unexplored.

Thus, there is no effective enough methods for drug repsonse prediction at the single-cell level. That is why we need new solutions.

c. What do the authors propose? Describe their solution, input data, processing, metrics etc.

Authors propose novel approach, which integrates bulk and scRNA-seq data.

If I understand correctly, scDEAL utilizes bulk RNA-seq data from two large databases – GDSC and CCLE. We provide scDEAL with the patient's scRNA-seq data and information about the drugs for which response predictions are needed. scDEAL first extracts bulk gene features from these databases using a denoising autoencoder (DAE) to reduce noise and dimensionality. It then performs prediction of drug response for each bulk cell line using these extracted features.

Next, single-cell features are extracted from the patient's scRNA-seq data, also using a DAE. Two separate DAEs are used to extract low-dimensional gene features from the bulk and single-cell RNA-seq data independently. During the train process, the Domain-adaptive Neural Network (DaNN) adjusts the parameters of the encoders and other parts of the model to make the bulk and single-cell features as similar as possible.

The Maximum Mean Discrepancy (MMD) metric is used to compare the single-cell and bulk features, ensuring they are comparable. The model tries to minimize this metric during learning to harmonize the features.

The final step is drug response prediction. The patient's scRNA-seq data is transformed into low-dimensional features and then fed into the predictor model, which generates drug response predictions for each individual cell.

d. What are the major results and achievements of the proposed solution? How do they relate to the existing methods and what are the limitations?

scDEAL was tested on 6 public scRNA-seq datasets treated by five drugs. Authors tested the tool with and without transfer learning step on each dataset, observing an average 19% increase in F1-score for the approach with transfer learning. Also they checked that combining data from GDSC and CCLE databases indeed results in significant F1-score increase compared to a single database approach.

In a case study involving leukemia cells treated with a BET inhibitor scDEAL accurately predicted drug responses with 97.1% of predicted drug-resistant cells and 95.8% of predicted drug-sensitive cells matching the original labels.

I didn't find any information on whether they tested other tools within the same benchmark on the same data. It seems that authors assume that scDEAL performance without transfer-learning step is the same as or higher than that of other existing methods.

Among limitations of scDEAL there are:

- scDEAL relies on the quality and availability of bulk gene expression data
- Lack of experimentally validated drug response scRNA-seq data

e. What are the conclusions?

The study introduces scDEAL, a deep transfer learning framework designed to predict single-cell drug responses by integrating bulk and scRNA-seq data. Testing on multiple datasets showed improvements in prediction accuracy compared to using bulk data alone, but no direct comparisons with other existing tools were made. Case study prediction results look promising; however, the tool is still limited by the contents of bulk gene expression databases and there is a need for more validated drug response scRNA-seq data.