Gene Expression Inference with Deep Learning, Yifei Chen, et. al.

Paper Report I

BMES 543: Quantitative Systems Biology Tony Kabilan Okeke

Summary

The paper describes the development and testing of a deep learning method that infers the expression of target genes from the expression of landmark genes. The goal of the study was to develop a more accurate prediction method than the linear regression models being utilized by the NIH LINCS program. The paper describes the architecture of their novel deep learning method, D-GEX, as well as the various methods used to train and evaluate its performance. Following comparisons with other machine learning models (linear regression and k-nearest neighbors regression) the researchers found that D-GEX achieved more accurate predictions for target gene expressions than the methods currently used by the LINCS program.

Strengths

- The paper does a good job of explaining the architecture of D-GEX, as well as the different training techniques used to evaluate its performance. It also provides useful explanations of the various machine learning models evaluated in the study.
- The visualizations used in the paper made the comparisons between different models easy to make sense of.
- The training and validation methods used in the study, as well as the size of the training dataset increase the reliability of their findings.

Limitations

- The researchers only compared the performance of D-GEX to Linear and k-Nearest Neighbor regression methods. Other machine learning techniques that could outperform linear regression were not considered in their study.
- The datasets used to train the model came exclusively from Affymetrix microarrays. Given that D-GEX appeared to have been trained on platform dependent expression patterns for certain target genes, it would have been interesting to evaluate its performance on microarray data from other manufacturers (for example, Agilent and Illumina).

Potential Improvements

- Re-training D-GEX on the entire dataset - instead of training the model on smaller partitions – could result in better overall performance. This would require access to a more powerful GPU than was available to the researchers at the time of the study.
- A more in-depth exploration of so-called “hub units” that were identified while visualizing the major weights of the model could provide more insights into the relationships between the landmark genes and the target genes.

What have Authors Done to Extend the Study

- It does not appear that the authors have made significant extensions to this study.

Multiple Choice Questions

Question: What technique was used to identify the ‘landmark genes’?

a. KNN Regression
b. Deep Learning
c. Linear Regression
d. PCA [correct]

<u>Question</u>: Which D-GEX architecture was most effective on the RNA-seq datasets?

a. 3000x1
b. 9000x2 [correct]
c. 6000x2
d. 9000x3

<u>Question</u>: Why was KNN-GE more accurate for some target genes than D-GEX?

a. KNN regression is biased by duplicate samples in the data
b. Inconsistent k nearest landmark genes for each target gene
c. Cross platform invariance of expression patterns [correct]
d. KNN regression is a non-parametric, instance-based method