Utilizing Neural Network Autoencoders for Unsupervised Feature Learning from RNA-Seq Differential Expression Analysis Towards Gene Ontology Term Prediction

Tony Kabilan Okeke[1], Ahmet Sacan[1]

[1]Drexel University, Philadelphia PA

4437

Introduction

- **Differential Expression Analysis** is used to identify genes that have different levels of expression between >2 samples/conditions.
- **Gene Ontology Enrichment** analyzes functional annotations of differentially expressed genes.
- Extracting biologically significant information from genomic data is a difficult task.
  - The high dimensionality and complexity of genomic data makes it hard to build accurate and interpretable prediction models for protein function
- **Autoencoders** are a type of neural network that are used to learn efficient data encodings in an unsupervised manner.
  - Trained to minimize reconstruction error
  - They include 2 components:
    - **Encoder:** compress input into latent space representation
    - **Decoder:** reconstruct original input from latent space

Objective:

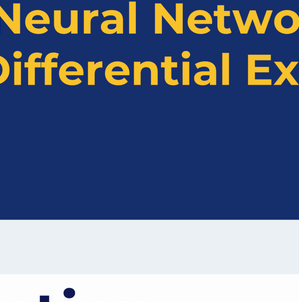
- ✓ Develop an autoencoder model to learn a compressed representation of differential expression data (i.e., log2 fold change values).
- Use the latent representation of the autoencoder to predict GO terms.

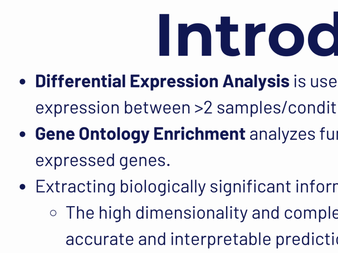
Methods

Dataset: DEE2

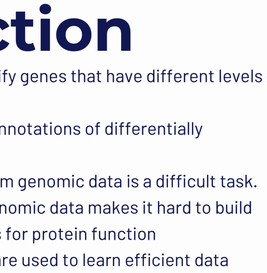
- 7,130 RNA-seq datasets from mice
- 48,878 gene transcripts

- Identify distinct groups

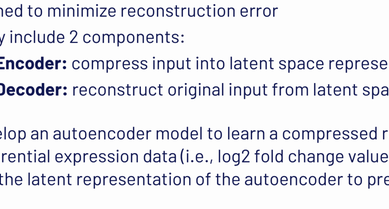




- Perform Differential Expression Analysis
- Compute Fold Changes



- Identify enriched GO terms



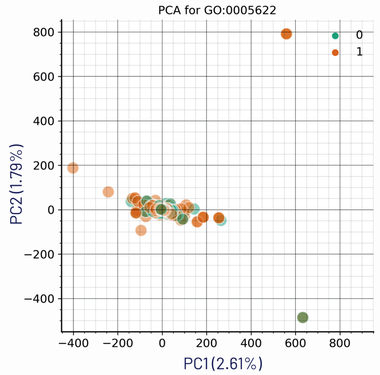
Train Autoencoder to reconstruct fold-change values.

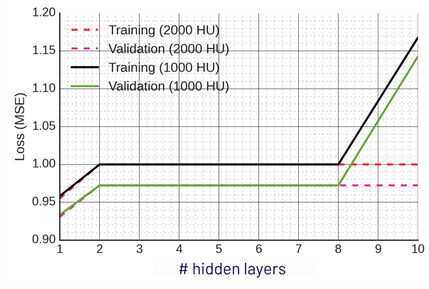
- Used **Hyperparameter tuning** to identify optimal # neurons and # hidden layers
- ReLU activation, MSE loss, Adam Optimizer

Results

Dimensionality Reduction

- Used PCA to visualize the ~21k genes in 2-dimensions
- Grouped samples by enrichment of GO:0005622 - 'intracellular anatomical structure'
  - Most common GO term in dataset



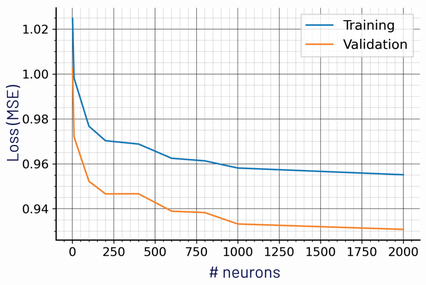


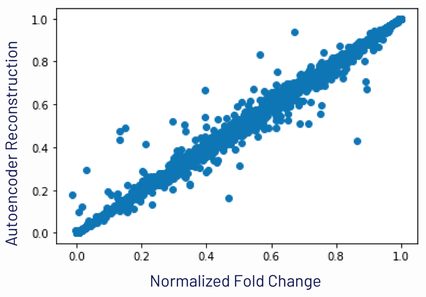
Hyperparameter Tuning: # hidden layers

- Trained & validated the model with a varied number of hidden layers
- Kept # neurons fixed at 2000
- Neurons are distributed evenly between the encoder and decoder
- 1 hidden layer performed the best

Hyperparameter Tuning: # neurons

- Trained & validated the model with a varied number of neurons
- Used only a single hidden layer
- Neurons are distributed evenly between the encoder and decoder
- 2000 neurons performed the best





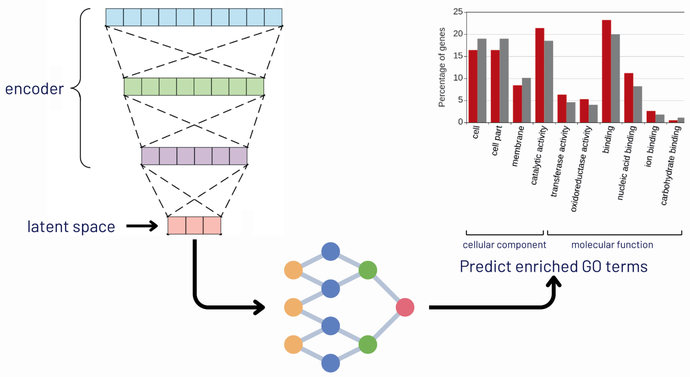
Cross-Validation

- Trained autoencoder with optimal hyperparameters over 50 epochs
- Model achieved an average cross-validation loss of 0.9632

Discussion

- The autoencoder was able to reconstruct the input gene expression data well, capturing the underlying structure of the data.
- However, due to memory constraints, we were only able to include 21,953 of the ~48k genes in the dataset. We were also bottlenecked by GPU compute, and could not explore more combinations of hyperparameters.
- More advanced autoencoder architectures like sparse and variational autoencoders could potentially improve performance.
- While the study provides a foundation for using autoencoders for unsupervised feature learning from gene expression data, further research is needed to address limitations and evaluate the approach on predicting biological annotations.

Future Work



References

- M. Ziemann, A. Kaspi, and A. El-Osta, "Digital expression explorer 2: a repository of uniformly processed rna sequencing data," Gigascience, vol. 8, no. 4, p. giz022, 2019.
- C. M. Koch, S. F. Chiu, M. Akbarpour, A. Bharat, K. M. Ridge, E. T. Bartom, and D. R. Winter, "A beginner's guide to analysis of rna sequencing data," American journal of respiratory cell and molecular biology, vol. 59, no. 2, pp. 145–157, 2018.
- Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, "Gene expression inference with deep learning," Bioinformatics, vol. 32, no. 12, pp. 1832–1839, 2016.
- M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for rna-sequencing and microarray studies," Nucleic acids research, vol. 43, no. 7, pp. e47–e47, 2015.

Want to learn more?

Contact the author for more information.