Exploratory Machine Learning Algorithm Development with a Real World Alzheimer's RNA Data Set

PURDUE UNIVERSITY | The Data Mine

Josh Aflleje, Arjun Agrawal, Racheal Arkorful, Yash Ashtekar, Fabiha Tasnim Aroni, Jerome Brown, Nadia Camacho, Safal Ghimire, Mahee Jain, Prince Kankam, Candice Koo, Tochi Obinma, Elizabeth Pardhe, Sachi Sayal, Leo Si-Yang Liu, Numa Vhora, Cheyenne Ward

TranslationalHDPM

Introduction

About TranslationalHDPM

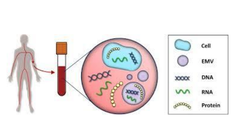
TranslationalHDPM is a start-up biotechnology company that uses Biomedical Data Science to gain insight into the underlying pathological mechanisms of neurodegenerative diseases to inform Network Medicine Therapeutics and Diagnostics.

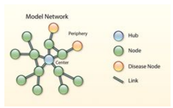
Background Motivation Objectives

Background: The appreciation of the connectedness of biological processes into networks has led to an improved understanding of health and disease. The use of 'clinical-grade' approaches including software coding has decreased the time to and increased likelihood of innovation progressing to improved patient care. The opportunity to interrogate ordinarily unavailable tissues, like the brain, has been accomplished by sampling cell free RNA (cf-RNA).

Motivation: Gain insight into underlying pathology of Alzheimer's Disease.

Objectives: Perform i) cf-mRNA data preprocessing of a previously published Alzheimer's Disease study, ii) use correlation analysis to identify network topology and iii) develop machine learning models leveraging analyses.

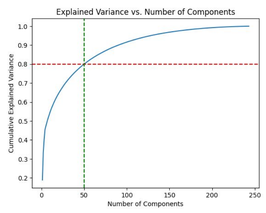




Materials

Analyses were conducted in Python and R. Data preprocessing and manipulation were performed using Pandas and NumPy. PCA and logistic regression were implemented with Scikit-Learn libraries: sklearn.decomposition.PCA and sklearn.linear_model.LogisticRegression. Visualization of results, including ROC curves, was generated using Matplotlib.

Feature Reduction



Data: Using CF-mRNA data from Toden et al. from Non-Cognitively Impaired (NCI) and Alzheimer's Disease (AD) where subjects served as proof-of-principle study.

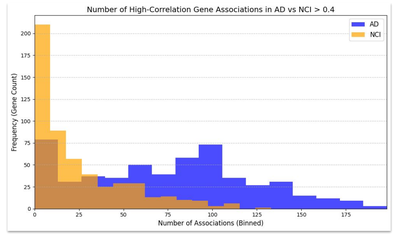
Feature Reduction: The dimensionality of the RNA-seq data was reduced using two methods: Principal Component Analysis (PCA) and aggregation of hub genes. PCA reduced 534 genes into 50 principal components that explained 80% of the variance.

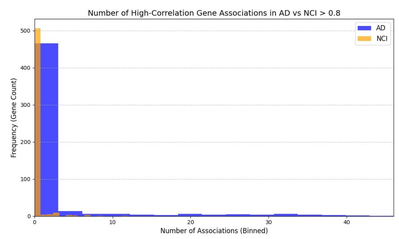
The hub gene approach produced two datasets. The first is Hub Genes alone that resulted in 7 features and the second is Clustered Hub Genes (hub genes and their associates) which also resulted in 7 features.

Correlation Analysis and Association Network

Pairwise Spearman correlations were computed on AD and NCI datasets separately to identify gene pairs with varying magnitudes of positive correlations (Spearman's $\rho > 0.4$ to 0.8). A threshold of 0.8 was ultimately selected to focus on biologically meaningful relationships while minimizing false positives. We consider hub genes to be genes with a correlation greater than 0.8 in AD and less than 0.8 in NCI.

An association network was constructed using genes as nodes and associations as edges. This allowed us to identify how many genes each hub gene was associated with in gene-gene correlation pairs.





Threshold Analysis

Threshold	Number of Hub Genes	Final Accuracy	Cross-Validation Accuracy	AUC
0.8	66	84%	84% +-4%	0.92
0.7	180	45%	83% +- 3%	0.72
0.6	341	59%	87% +- 3%	0.76
0.5	458	53%	88% +- 3%	0.67
0.4	503	55%	88% +- 2%	0.69

- Consider hub genes to be genes with:
  - correlation > threshold in AD and < threshold in NCI
- Higher threshold levels result in more relevant genes being kept (less noise)
- Model performance is near random for lower thresholds

Model Development

All Logistic Regression models used a 80/20 training and testing split. To ensure the model is not sensitive to the large values in the data, we perform a log2 transformation. Input to the models included the dataset with original 534 dimension, PCA projected data and a combination of hub genes and low correlations.

Original Data

The original dataset had dimension 234 x 534. Logistic Regression was run on this dataset to show motivation for using dimension reduction methods and the exploration of collinearity.

Principal Component Analysis (PCA)

Principal Component Analysis is traditionally used as a method to reduce data in high dimension to low dimension. In our research, we look at the minimum number of principal components needed to maintain 80% variance in the data. We then projected the original dataset onto these components which resulted in a new dataset of 50 features.

Hub Genes

Unique genes in gene-gene pairs that have a correlation above the threshold in AD and below the threshold in NCI are considered hub genes. The genes that these hubs are connected with are associations. We perform logistic regression on hub genes alone, hub genes and their associations as well as hub genes and low correlations.

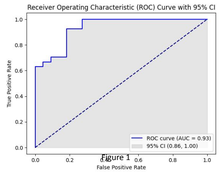


Figure 1

Figure 2

Figure 3

Logistic Regression Using…	Final Accuracy	Cross-Validation Accuracy	ROC	Number of Features	Overfitting/ Underfitting?
PCA (with log2 transformation)	86%	81% +- 4%	0.94	50	Yes
Original Dataset (w/o log transformation)	73%	80% +- 6%	0.92	534	Yes
Log2 transform on reduced features (hub genes)	86%	77% +- 5%	0.88	7	No
Log2 transform on hub genes and low correlations	84%	84% +- 4%	0.92	66	No
Clustered Hub Genes	84%	79% +- 3%	0.93	7	No

Conclusions

Our study explores different ways to process and analyze gene expression data for early Alzheimer's detection.

When transforming the original dataset with PCA, we found that we were able to keep around 80% variance in the dataset while also reducing the number of dimensions. When logistic regression was run on this reduced dataset we found the model performed with an accuracy of 86% however it was overfitted and did not generalize well to unseen data (refer to *Figure 2*).

The next method was to explore correlated genes using a combination of highly correlated genes (hub genes) and low correlated genes. Only hub genes with a correlation greater than 0.8 in the AD dataset were kept. When logistic regression was performed on hub genes alone the model's accuracy was 86% with no overfitting.

Our last method was to cluster hub genes and associated genes. That is hub genes and the genes they connect with at least 10 times. Doing so allowed the dataset to be reduced to 7 dimensions and the model to perform with an accuracy of 84% with no overfitting (refer to *Figure 3*) signaling that the model generalizes well to unseen data.

Future Studies

Future research will maximise Alzheimer's prediction by comparing Random Forest (RF), Support Vector Machines (SVM), and Decision Trees (DT). RF detects subtle interactions and excludes overfitting, SVM prefers high-dimensional non-linear data, and DT lends itself to explaining biomarker selection. Feature amalgamation, engineering hybrids, and network-based paradigms will also be evaluated to maximise model performance as well as interpretation. Integrating these approaches will improve predictive efficacy and machine learning in Alzheimer's disease.

Acknowledgements

We thank Dr.Sninsky and Dr.Khoury of SuperfluidDX/TranslationalHDPM for their valuable counsel and assistance in facilitating this project. We appreciate Dr. Ward, Yash, Mahee, Arjav, and the entire Data Mine team for their constant support. A special thank you to Dr. Marko Samara and Dr. Steffen Eikenberry for heading the ASU team and Dr. Drew Lazar for heading the BSU team.This material is based upon work supported by the National Science Foundation under Award No. 2123321. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors thank our colleagues at the American Statistical Association for administrating the activities of the grant, entitled *NSF HDR DSC: National Data Mine Network*.

References

Barabasi et al. Nat Rev Genetics 12, 56 (2011).
Cao et al. Frontiers in Mol Neuro 16, 1 (2023).
De Sota et al. Expert Rev Mol Med 26, 1 (2024).
de la Fuente Trends in Genet 26, 326 (2010).
Ideker and Krogen Mol Sys Biol 8,565 (2012).
Mitra et al. Sys Biol App 10, 50 (2024).
Pandey and Loscalzo Nat Rev Nephology 19, 463 (2023).
Spies et al. Clin Chem 70, 1334 (2024).
Toden et al. Science Advances 6, eabb1654 (2020).