PH240C Final Project

Protein Expression across Disease Groups: Analysis of the Human Brain Proteome

Ramasubramanian Balasubramanian, John Canty, Annan Deng, William Krinsman

December 6th, 2018

Abstract

Contents

1 Introduction

Patients with neurodegenerative disorders such as Alzheimer's disease (AD) and Parkinson's disease (PD) share common physiological and symptomatic similarities, however the molecular pathways linking these diseases are incompletely understood. Recently, a comprehensive quantitative proteomic dataset of the human brain in patients with AD, PD, and AD/PD comorbidity has been publicly released. We will perform exploratory data analysis and modelling in order to identify if there are disease-specific differential gene expression patterns that are associated with AD and PD neuropathologies. Furthermore, we will then perform Weighted Correlation Network Analysis (WCNA) of the identified genes to assess if there coexpression relationships between subsets of these genes. Finally, using the PANTHER gene ontology classification system, we will assess whether these genes exhibit functional relationships.

2 Dataset

The dataset[3] that we will be evaluating was obtained by performing Tandem Mass Tag (TMT) isobaric mass spectrometry on brain tissue obtained from individual human donors. Brain tissue samples were obtained from 40 individual patient samples across two separate brain regions, the Frontal Cortex and the Anterior Cingulate Gyrus). The dataset obtained by using a factorial experimental design, where there are 5 experimental batches, with two samples of each phenotype (Control Patient, Alzheimer's Patient, Parkinson's Patient, Alzheimer's/Parkinson's Patient) within each batch. For each batch, a control sample was generated by pooling fractions of all samples in order to generate a global internal standard (GIS) expression measure. Individual data entries are displayed as base-10 log-transformed ratios of the obtained peptide counts for the ijth sample covariate of interest and the ith covariate of the GIS standard. From the Frontal Cortex, $10,000$ unique protein groups were identified, while from the Anterior Cingulate Gyrus, $10,695$ protein groups were identified.

3 Exploratory analysis and preprocessing

In the initial portion of this project, we will assess data quality and phenotype and inter-batch variability by box-plot visualization. We will then compare different normalization procedures such as loess, full-quantile, and upper-quantile when applied between samples. Next, we will apply PCA and MDS methods in order to determine if there is clustering of samples based on the disease phenotype.

1. **Data quality and variability**
2. **Preprocessing**
3. **Exploratory analysis**

4 Linear modelling to identify differential expression

We will utilize linear models using samples as covariates (with parameter coefficients for the control, AD, PD, and AD/PD samples) in order to determine the strength of association between each gene and the disease phenotypes. Finally, we will identify differentially expressed genes relative to the control samples for each phenotype using volcano-plots.

1. Model definitions

2. Volcano plots to identify differential expression

5 Weighted correlation network analysis

Using the candidate differentially expressed genes for the PD, AD, and PD/AD disease phenotypes, we will apply weighted correlation network analysis in order to determine if the candidate genes exhibit interesting coexpression patterns. First, using the available WGNA R-package[1], we will determine whether there are interesting coexpression clusterings between differentially expressed genes for each phenotype. Next, we will assess the clustering size/structure and compare these properties between phenotypes.

1. Coexpression clusterings for each phenotype

(a) Coexpression clusterings for Alzheimer's

(b) Coexpression clustering for Parkinson's

(c) Coexpression clusterings for both Alzheimer's and Parkinson's

2. Analysis of coexpression clusterings

(a) Size and structure of clusterings

(b) Comparison of clusters between phenotypes

6 Biological functions of differentially expressed genes

Using the PANTHER gene ontology database, we identify the functional classifications of the differential expressed genes. We will perform this analysis between all phenotypes in order to comment on genes that may have unique biological roles related to PD, AD, and PD/AD disease states.

1. PANTHER gene ontology database

The PANTHER project, which we used to identify the biological functions of the differentially expressed genes, has the following purpose[2]:

> The PANTHER (protein annotation through evolutionary relationship) classification system (http://www.pantherdb.org/http://www.pantherdb.org/) is a comprehensive system that combines gene function, ontology, pathways and statistical analysis tools that enable biologists to analyze large-scale, genome-wide data from sequencing, proteomics or gene expression experiments. The system is built with 82 complete genomes organized into gene families and subfamilies, and their evolutionary relationships are captured in phylogenetic trees, multiple sequence alignments and statistical models (hidden Markov models or HMMs). Genes are classified according to their function in several different ways: families and subfamilies are annotated with ontology terms (Gene Ontology (GO) and PANTHER protein class), and sequences are assigned to PANTHER pathways.

2. Functional classifications of differentially expressed genes

(a) Biological differences between Alzheimer's only and controls

(b) Biological differences between Parkinson's only and controls

(c) Biological differences between controls and combined Alzheimer's and Parkinson's

(d) Biological differences between Alzheimer's only and Parkinson's only

(e) Biological differences between Alzheimer's only and Alzheimer's with Parkinson's

(f) Biological differences between Parkinson's only and Parkinson's with Alzheimer's

References

[1] P. Langfelder and S. Horvath, *Wgcna: an r package for weighted correlation network analysis*, BMC Bioinformatics, 9 (2008), p. 559.

[2] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, *Large-scale gene function analysis with the panther classification system*, Nature Protocols, 8 (2013/07/18/online).

[3] L. Ping, D. M. Duong, L. Yin, M. Gearing, J. J. Lah, A. I. Levey, and N. T. Seyfried, *Global quantitative analysis of the human brain proteome in alzheimer's and parkinson's disease*, Scientific Data, 5 (2018/03/13/online).