Protein Expression in Down Syndrome Mice

CSC 621 – Machine Learning

Han Saim Jeong

Dr. Ruth Cheng

Introduction:

Down Syndrome is a condition caused by the trisomy of chromosome 21 in humans. This means that there is an extra chromosome resulting in three copies of chromosome 21. This overexpression of genes caused by the extra chromosome results in a variety of abnormal and detrimental conditions. Some of those conditions include heart defects, immune system issues, blood disorders, developmental delay, mental retardation, increased risk of Alzheimer’s disease, and epilepsy. This genetic abnormality affects around 1 in 1000 live births worldwide and more than 200,000 cases annually in the United States. Due to the high incidence of occurrence of Down Syndrome, there is a great deal of interest in researching both pharmaceutical relief for symptoms of Down Syndrome and the mechanisms behind the symptoms caused by Down Syndrome.

In studying the mechanisms behind Down Syndrome’s variety of symptoms, studying human patients is not always a feasible option. Animal modeling allows for a lot of invasive tests to be conducted in a cost efficient, and time efficient manner. Mice are similar to humans in terms of their physiology and anatomy. In fact, mice have a brain architecture that is composed of many of the same brain cell types as humans. Mice also have a similar genome. Mice are useful as research subjects as well due to their relatively cheap cost and ease of care. Mice are very prolific and have a short turnaround between pregnancies. Additionally, they have a short lifespan allowing for longitudinal studies to occur over a short period of time. Lastly, genetic manipulations are simple to make in mice allowing for the research of specific receptors or conditions such as the trisomy in Down Syndrome.

For this study, a mouse model of Down Syndrome was used. Ts65Dn mice are genetically manipulated to express trisomy of mouse chromosome 16, and thus express trisomy for around two-thirds of the genes orthologous to human chromosome 21. These mice display neural cognitive deficits and behavioral abnormalities. They also show spatial learning defects and slow developmental delay. These symptoms correspond to the same symptoms shown in human Down Syndrome patients.

Besides manipulating the genotype of the mice, the study involved two other independent variables as factors for research. The first involved running a test of context fear conditioning on the mice. In fear context conditioning, mice are introduced to a novel context/environment and then either: shocked and then allowed to investigate their surroundings or allowed to investigate and then shocked after. This test assesses the associative learning ability of the subjects. Research shows that when the shock precedes the exploration of the novel context, mice do not learn to associate the environment with the shock. However, when the context precedes the shock, mice learn to associate the context with the shock and will display fear behavior patterns prior to the shock application. Studies in Down Syndrome mice show that mice with Down Syndrome are inhibited in this associative learning.

The second of the additional variables is the use of a drug called Memantine. Memantine is a NMDA receptor antagonist that is currently used in Alzheimer’s Disease treatment. While Memantine does not affect learning in control mice, or mice with no genetic manipulations, it does rescue learning in Ts65Dn mice. This means that Down Syndrome synonymous mice, when injected with Memantine, will regain the ability to learn association in the context fear conditioning. When studying mice with a combination of these three conditions, genotype, behavior, and drug treatment, the proteins behind the mental disability of Down Syndrome can be investigated and identified.

The data set utilized in this project consisted of the measurements of protein expression of 77 different proteins/protein modifications taken from the nuclear fraction of the cortex of the brain. 38 control mice and 34 trisomic mice were utilized in the experiment. For the experiment, 15 measurements of each protein were registered per sample. This resulted in a total of 570 measurements for the control mice and 510 measurements for the trisomic mice. Each measurement was taken as an independent sample. The classes of mice were separated based on the genotype, treatment, and behavior of the mice. This resulted in 8 classes:

c-CS-s: control mice, stimulated to learn, injected with saline (9 mice)  
c-CS-m: control mice, stimulated to learn, injected with memantine (10 mice)  
c-SC-s: control mice, not stimulated to learn, injected with saline (9 mice)  
c-SC-m: control mice, not stimulated to learn, injected with memantine (10 mice)  
t-CS-s: trisomy mice, stimulated to learn, injected with saline (7 mice)  
t-CS-m: trisomy mice, stimulated to learn, injected with memantine (9 mice)  
t-SC-s: trisomy mice, not stimulated to learn, injected with saline (9 mice)  
t-SC-m: trisomy mice, not stimulated to learn, injected with memantine (9 mice)

By investigating this data set, we can identify specific proteins in the brain that are associated with Down Syndrome, the learning deficit involved, and the effects of Memantine in Down Syndrome mice. Through the proteins identified, further studies into the pathways and mechanisms involved with those proteins can be investigated for future knowledge and potential cures.

Data Processing:

The data set utilized had several null values for measurements that were missing. To analyze the data properly, first these null values would need to be replaced. In order to replace the null values, first the protein expression data was separated into groups by their class. Then the null values were replaced by the mean of the expression data based on which group it belonged to.

Text

Description automatically generated

Then, the data needs to be normalized. Thus, we run min-max scaling normalization on the data set so that all the protein expression data is within the range of [0, 1].

A computer screen capture

Description automatically generated with low confidence

After this step, all the preprocessing required to analyze the data set is finished.

Approach:

Now, we can begin analyzing the data. To begin with, k-means clustering was utilized to view the data set and if any clear differences were present. However, as the data set consisted of 77 different protein expression data, the dimensionality of the data set needed to be reduced before clustering was run. To reduce dimensionality, principal component analysis (PCA) was run.

Shape

Description automatically generated

Principal component sum was graphed against the # of components.

The number of components used for clustering analysis must account for 70-80% of the variance

Next, to calculate the number of clusters optimal for running k-means on the data set, the within cluster sum of squares (WCSS) was calculated for the principal components calculated. By graphing the WCSS against the number of clusters, we can determine how many clusters would be optimal to use as k for k-means clustering.

A picture containing shape

Description automatically generated

Now that we have a reduced dimensionality data set and the k value for k-means clustering, we can run the k-means clustering algorithm and plot the results to visualize the data set.

Chart, scatter chart

Description automatically generated

Here we can see that there are clear clusters in the data set to investigate. However, the plot does not show which groups are within those clusters nor the critical proteins. Thus a Welch’s test was run on the data set, comparing first the two genotypes against each other. Then the two behavior conditioning groups, and last the two treatment groups. From these Welch’s tests, the proteins that were significantly different between groups were identified and isolated. From these isolated proteins, we then ran through the same process of PCA, WCSS, and k-means clustering for each condition: genotype vs. treatment, behavior vs. treatment, genotype vs. behavior, and genotype vs. treatment vs. behavior. (Shown below in order from left to right)

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated  
From these clusters, parallel coordinates plots were graphed and used to identify which proteins exactly were implicit as critical proteins between conditions. (Shown below in the same ordering as the clusters plotted above)

A picture containing text, tree

Description automatically generatedA picture containing text, writing implement, pencil, stationary

Description automatically generated

A picture containing text, pencil, curtain

Description automatically generatedA picture containing text, tree, outdoor

Description automatically generated

From these parallel coordinates plots, the proteins were determined visually to have significant impact on the condition groups as follows.

|  |  |  |
| --- | --- | --- |
| **Drug vs. Genotype** | **Drug vs. Behavior** | **Rescued Learning** |
|  |  | **pCREB** |
| **pMTOR** |  | **pMTOR** |
|  |  | **P38** |
|  | **CaNA** |  |
|  | **pGSK3B** |  |
| **S6** | **S6** | **S6** |
| **ARC** |  | **ARC** |
|  | **pPKCAB** |  |
|  | **pS6** | **pS6** |
| **pPKCG** |  |  |
| **AcetylH3K9** |  |  |
| **Tau** |  |  |

Conclusion:

There are certain proteins in the cortex that can be identified as critical to affecting differences in groups between control and trisomal mice, learning and not learning, and memantine injected and saline injected. By looking into the proteins identified, we can begin to investigate the pathways in the brain affected by these specific proteins. However, due to the high dimensionality of the data set as well as the data being measured, the method of clustering by k-means clustering algorithm is difficult. The reliability of the results found is not confirmed. Especially as the proteins found different were determined visually through the parallel coordinates plots. Next time, a density-based clustering algorithm such as DBSCAN or OPTICS would be preferable to k-means clustering.