Philip Rinaldi and Molly Lyons

CMPS 3160

Professor Mattei

29 September 2020

Milestone 1

We are currently looking at a dataset about mice protein expression (https://www.kaggle.com/ruslankl/mice-protein-expression). This dataset contains information on proteins that are expressed in several conditions in mice, including when mice are treated with the drug memantine (a drug used to treat Alzheimer's disease). We want to use the data to determine the effect memantine on the mice. Also, memantine is NMDA receptor antagonist like the drug ketamine, which is being researched as a fast-acting antidepressant, so we were thinking of finding another dataset related to depression and Alzheimer's to see if there is some correlation between NMDA receptors, Alzheimer's, and depression. It might be difficult to find datasets that have exactly what we are looking for, which would be proteins that are over or under expressed in major depression, during ketamine use, and in people with Alzheimer's disease. The Psychiatric Genomics Consortium has data on major depressive disorder, though it is unclear what form the data is in (entire genome, proteins expressed, etc.). We also must apply to download the data from the Psychiatric Genomics Consortium, which poses the risk that we might not be able to get access to the data. There is a microarray analysis on Alzheimer's disease on Kaggle (https://www.kaggle.com/andrewgao/alzheimer-microarray-analysis), which could be of use to us if we decide to take this direction for our final tutorial. Particularly, the mice protein expression data contains 1080 measurements of expression per protein, for 72 mice. The data has already been neatly categorized by the experimental conditions to which the mice were subjected and their biological traits (trisomic or control gene, injected/not injected with memantic, etc.)

Two other datasets we found are ones pertaining to structural proteins and genes related to pancreatic cancer. (https://www.kaggle.com/shahir/protein-data-set and https://www.kaggle.com/abhiparashar/cancer-prediction, respectively. It was a bit difficult for us to formulate reasonable queries, as these are enormous sets of data and are intended for a wide

range of research. Though these datasets are broad, we could look for overall trends that might help point us in a more specific direction. Additionally, it might be very difficult for a pair of undergraduates, one of whom does not study Biology, would be able to carry out any project of such magnitude without proper expertise and guidance. We are both in agreement that we want to work with a biological dataset, so even if our previously chosen datasets do not work out we are confident that we will be able to find another one that works.

We met on Zoom once to set up the github.io public page, share access to it, and establish a google doc to write up our milestone. We plan on zooming again next week, and will likely continue to meet bi-weekly for the remainder of the project. We will also establish a private github repository where we can collaborate on code.