Genetic Basis of Exploratory Behaviors in Mice (FY1821)

Cynthia, Josh, Kewei, Srinidhi, Zach

With Prof.Andres Bendesky & Data Science Institute

Contents

Overview

The report briefly discusses the ethical implications of the capstone project(GSAS FY1821): The Genetic Basis of Exploratory Behaviors in Mice

Privacy and Security

The project uses pre-captured aerial video data(top view) of mice exploring an elevated plus maze and genetic marker data of their genomes. Mice were not removed from their natural environment and used for the study, rather, they were bred in laboratory conditions for the purpose of performing scientific research. Hence, this resulted in no known effect on the existing ecological balance in the wild. Normally, in experiments capturing data related to humans, sensitive data can create privacy concerns when the subjects creating the data have not been informed of data acquisition or of their consent towards data acquisition. Considering our project's subjects were animals, there was no concern in acquiring consent from the subjects. However, although there were no privacy concerns, the animals still needed to be ethically treated during the study. In accordance with IACUC guidelines, all animals were housed safely and treated humanely for the duration of the experiment.

Biases and Limitations

Data for the project consisted of raw videos and genetic markers for two sister species of mice. The Mice were introduced to the center of the elevated plus maze through a lift and the video was captured aerially without disturbing the subject. The genetic data was acquired before our project began using biological methods.

The video data, captured at 3̃0 frames per second, proved to be an excellent analog of explaining where the mouse was at any point during the trial, allowing for feature engineering to be performed easily. If there was a limitation concerning the video data, it could be in its accuracy. We relied on the video to inform us where the mouse was using pixel values, so any discrepancies in the quality of the video may have affected the results in the processing script. This was counteracted by visual inspection of randomly selected trials to confirm the results were valid. We also assumed the genetic markers provided were reliable and consistent across the animals, which allowed us to discover genetic regions associated with the observed behaviors in the video data. Further biological analyses such as a replication study could confirm the validity of the genetic data.

We obtained desired data through a processing script that took the video as input and produced values for desired features as output. This script assumed that the mouse followed a continuous path. The mouse outline that was detected from the video is motion based. Hence, sometimes, the area the mouse occupied was calculated to be zero(when the mouse did not move). This meant care had to be taken to filter out these invalid mouse regions so as not to skew the features(some cleaning was also done by smoothing all the features which took care of the 'discontinuities' in data).

Reproducibility and Transparency

Because our project consisted of data from a scientific study performed for the purpose of furthering basic research in the biological domain, reproducibility was a necessity. The process of advancing scientific understanding relies on reproducing results in order to confirm that the findings are genuine. From a data science perspective, this meant that all of our code should be accessible and able to be used by the lab. Because of this, the entire project is available on our GitHub repo.

We tried to make the data pipeline as understandable as possible (Figure 1). To reproduce the project, one would have to extract frame by frame positional data of the Mice from the raw videos(not included in the Repo owing to its size), calculate aggregated features by referring to our code, conduct two-factor significance tests on pure species(+Sex) to select features and finally, run QTL analysis on the selected features using genetic data from pure species & their F1, F2 hybrids.

As for transparency of the workflow, detailed explanation of each step in the data pipeline and interpretation of outcomes can be found in our final report. Video data was fed through a processing script to create the CSV, which contains the feature values of each mice. Feature Engineering decisions involved discussions with the head of the lab (project mentor) to determine if the results were comparable to known/observed tendencies of the species. The evidence for when to perform quantitative trait loci (QTL) analyses was determined by looking for statistical significance between pure species for the specific feature. For QTL thresholds, we consulted with our project mentor for the ideal LOD score that wouldn't exclude too many features while also trying to reduce noise.

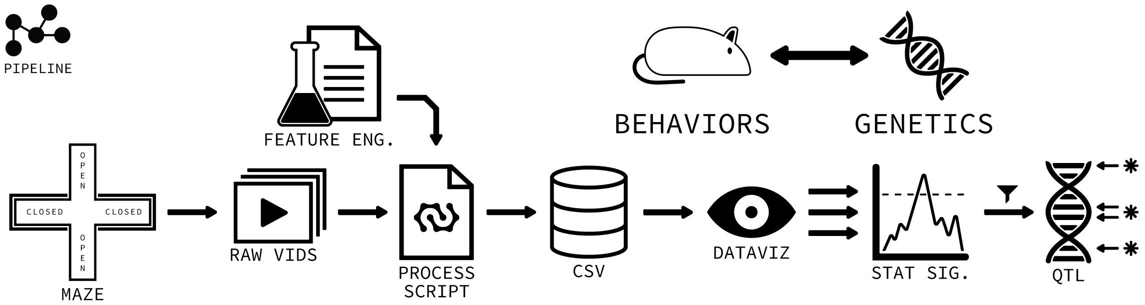

Figure 1: Pictorial representation of project pipeline