Mais 202 Project Deliverable #1

Cole Killian

October 1, 2019

1. **Dataset**

   In recent years there has been rapid growth to the publically available methylation data. My project aims to predict the age, sex, and ethnicity of a person by comparing his methylation data to other methylation profiles where the age, sex, and ethnicity is already known. I will use the ”Illumina Human Methylation 450 Platform” made available by the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534).

   I chose this datasets because it contains many annotated samples (over 90000) and is free. Even if a relatively large percentage of the samples are unusable, I will have a lot to work with.

2. **Methodology**

   a) **Data Preprocessing.**

   While the dataset contains many samples, they are submitted by a variety of sources and therefore are formatted in slightly different ways. The most useful data will be those samples that have consistent annotation types. Preprocessing will involve going through the samples (ideally with a script) and identifying all those that are compatible with one another / usable. At this point I will combine the samples into a single dataset. The most useful information will be the methylation profile itself, and the annotations for age, sex, and ethnicity. Unfortunately I will not be able to make use of samples that do not have these annotations and will have to remove them.

   b) **Machine Learning Model.**

   I plan on developing a better understanding of what can be predicted along the way, but right now the goal would be to predict the age, sex, and ethnicity of a person given their methylation data. I plan on experimenting both with supervised learning classification and unsupervised learning via principle component analysis. My reasoning is that methylation profiles will coorelate age, sex, and ethnicity, and therefore that these techniques will be able to categorize these traits.

   c) **Final conceptualization.**

   My choice for the final project is an application. My goal would be to have a webapp that allows anyone to upload a methylation profile and be given an estimation as to the person’s age, sex, and ethnicity.