DS 340W Group Project Proposal

Group 10

Zheng Zhang
Department of Statistics, Eberly College of Science
Pennsylvania State University

Madison Novak
Department of EECS, College of Engineering
Pennsylvania State University

I. Introduction

The Encyclopedia of DNA Elements (ENCODE) Consortium is organizing a challenge to accurately impute biochemical data associated with functional genomic elements in a variety of cell types. The consortium has performed many experiments to characterize biochemical activity associated with regulatory activity across the entire human genome in diverse cell types and tissues. However, many combinations of assays and cell types remain to be performed. These experiments are expensive, and technical challenges may prevent comprehensive characterization of all marks in all cell types, so computational methods capable of predicting the output of these assays are potentially valuable.

II. Problem Statement

The goal of this project is to integrate the machine learning model with the computational algorithm developed by the-Mahony-Lab to improve the imputation accuracy further. In this problem, we're facing two issues that needs to be resolved. First, how to develop and use the computational method to gain insight and learn from the human epigenome data to predict the data for the unknown human cell types and histone marks. The other challenge is the size and sparsity of the training data we have. Since there are over three billion positions (rows) for each data, the overall size can be massively large so that even a simple loop could be time-consuming.

III. Previous Work

Since the human epigenome is characterized by many different measurements it becomes difficult to understand how to use the data for computational methods. Previously, the best method used to address these challenges have been by using the deep neural network tensor factorization method, also known as Avocado. The Avocado method is able to combine the enriched epigenomic data and form it into an accurate portrayal of the human genome. This is needed in order to be able to get a more concise look at the already data packed data set. This has been found to be the most accurate method for investing the data from the Roadmap Epigenomics Consortium which is the data that we will be using for this project. In addition to being the most accurate method, it has been shown that this method will allow for the machine learning models to perform better than if the models were to be used directly on the epigenomic data.

IV. Approach and Methodology

In order to solve this problem, we plan to take two approaches to this problem. First, we're going to wrap the computational algorithm developed by the-Mahony-Lab into a more efficient distributed/parallel algorithm to reduce the computing time. Second, we're going to apply large-scale machine/statistical learning algorithm and integrating the results from the IDEAS epigenome annotation system (also developed by the-Mahony-Lab) to gain insight from the data and make predictions to improve the accuracy further.

V. Evaluation Metrics

In regard to evaluating our results, we will be following the Prediction Scoring Metrics that has been given on the ENCODE Imputation Challenge website. Our groups findings will be compared against the measured values that were found in the experimental stage of this challenge. The predicted (our groups predicted data) and true (determined by previous experiments) values will be compared in order to measure performance. Our predicted data will be reviewed by the squared error, Pearson correlation, and Spearman correlation and compared to the true analysis. In addition to these, the evaluation metrics will be looked across multiple aspects of the genome.

VI. Milestones

To keep on top of our project we will be creating milestones to follow in order to be able to achieve in the given time frame. By week 6, we want to have the EDA completed as well as trying different machine learning algorithms to see which has the best outcome. In addition to this, we want to try different types of data normalization methods to see how these compare against one another. Once we have experimented with the machine learning algorithms and data normalization methods, we want to further investigate the results of IDEAS epigenome annotation system. For week 9, we want to use the results from the annotation systems in order to improve the prediction results. Along with this, we want to start working on the pipeline of the project and implementing our new methods while working through any problems that arise.