Project Proposal: MotifSeeker

Justin Hii, A16682841
Tusha Karnani, A17339806
Numa Yazadi, A16837717

Name of Tool: MotifSeeker

Description: We would like to build a tool in Python to perform motif enrichment analysis on ChiP-seq datasets. The goal of this tool is to allow us to identify overrepresented sequence motifs in specific regions. THis tool compares to an analysis tool called Analysis of Motif Enrichment (AME). This tool can help evaluate the accuracy of the tool we are building.

Implementation: Using Python, we will use specific libraries such as Biopython to facilitate sequence manipulation and enrichment analysis algorithms. The tool will use Chip-seq files and reference genomes as its input. The output will be enriched motifs with the statistical analysis scores. We will include options to adjust parameters like motif length and statistical significance thresholds.

Plan to Benchmark: We plan to benchmark our tool against HOMER, which is a tool for analyzing ChIP-seq datasets. We will be comparing the runtime of both programs in order to calculate the speed of both tools. Additionally, we will be checking if our program is as accurate as HOMER by comparing its results.

Public Dataset: The public dataset that we are using is found on Encyclopedia of DNA Elements (ENCODE) dataset. Using this, we will focus on CHIP-seq data for transcription factors in human cells. The data comes with CHIP)seq peaks and the metadata that goes along with it. So it suits our model for testing and evaluating motif enrichment analysis tools. An example of a dataset we can use is the CHIP-seq data for transcription factor (CCCTC-binding factor) in human cells. We will also be using JASPAR motif databases in order to identify motifs.

Data Set Link: https://www.encodeproject.org/