Thesis Project Portfolio

Deep Multimodal Representation Learning to Integrate Natural Language Processing with Genomic Interval Data for Tailored Biomedical Discovery

(Technical Report)

Tracking the History of Race and Ethnicity Data Collection in Genomics Research

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Peneeta Ann Wojcik

Spring, 2024

Department of Biomedical Engineering

Contents of Portfolio

Executive Summary

Researchers use DNA and RNA samples in studies to understand the human genome and diseases resulting from genetic dysregulation. It is vital to understand that these samples originated from an individual despite seeming disembodied. When samples are collected, researchers must abide by federal regulations instantiated by agencies such as the Food and Drug Administration (FDA) and the National Institutes of Health (NIH). Race and ethnicity data is collected for cell samples and patients in clinical studies to ensure that population groups are equally represented during trials. This is largely contested because there is established evidence that race and ethnicity are not valid biological constructs. Since this is the case, how can human genetic variability be quantitatively measured? The technical topic focuses on developing a method to identify relevant genomic regions associated with a search query, while the STS topic tracks the history of federal regulations on racial and ethnic data collection in biomedical research. Together, these topics will explore distinct areas related to genomics. One involves developing a tool using only raw biological data ignoring race and ethnicity, while the other unveils the trajectory of federal race and ethnicity data collection standards.

Genomic data has exploded since the development of high-throughput genetic sequencing. One area of interest in genomics is epigenomics, or the study of external modifications to DNA that affect gene expression. These modifications can occur during disease states or regular cellular processes. Epigenomic data is stored in text files called BED files, where each row contains coordinates representing a genomic region. Due to the nature of this data, raw data analysis is impossible. Hundreds of thousands of genomic regions can be present in a single file, making it challenging to determine closely related regions of DNA that are affected by epigenomic modifications. The focus of the technical topic is to train four distinct deep learning models to

translate a user-entered search query to a list of relevant genomic regions in the form of a BED file. These models are a direct encoder, Text2Bed neural network, diffusion model, and transformer. The goal is to eventually create a genomic search tool to aid researchers. The model with the best performance will be used to return a list of genomic regions to a search query.

The STS project explores the regulatory aspect of biomedical research. All research in the United States is supported by federal institutions such as the NIH and the FDA. These institutions have guidelines on how cell samples are collected, which includes the collection of race, ethnicity, and ancestry (REA) data. Current arguments supporting REA data collection in research are that it is the best method to estimate socioeconomic impact, and there is no standardized method to quantifiably measure human genetic diversity. The STS research attempts to understand how federal guidelines for REA data collection prolong existing racial biases in research and clinical settings. To do this, three policies were analyzed chronologically: United States Office of Management and Budget (OMB) Directive 15, NIH policies on the reporting of REA data, and FDA REA collection guidelines. The latest update to OMB Directive 15 shows promise to remove REA data collection with the addition of a clause stating the collection of REA data is not required by any agency, but it is still too early for other agencies to respond.

Though successfully designing each model for the technical project, unexpected issues arose. For preliminary model training, the dimensionality of each BED file was too large to input into the model, so a small subset of regions was sampled. Two of our models successfully generated BED files, accomplishing the task of translating from a query to a BED file. Future work for the technical project would be to develop a method to computationally validate generated genomic regions and include a method to select the most relevant genomic regions from each file before

training to preserve the biological information lost from sampling random regions. From the STS research, REA data is the best estimator of socioeconomic status, which plays a large role in disease onset and progression. It would be interesting to analyze current methods of determining human genetic variability and the feasibility of their widespread implementations. Conducting a survey for clinicians and genomic researchers to take regarding the use of REA data would highlight differing objectives between the two groups and recommend solutions to replace REA data with a more biologically relevant factor.