DEVELOPING DEEP-LEARNING MODELS TO CREATE A SEARCH TOOL FOR GENOMIC REGION SETS

THE DETRIMENT OF RACIAL CATEGORIZATION IN GENOMICS RESEARCH AND FACTORS PROLONGING IT

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Biomedical Engineering

By
Peneeta Wojcik

October 27, 2023

Technical Team Members: Caitlyn Fay, Lily Jones, Zachary Mills

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kent Wayland, Ph.D., Department of Engineering and Society

Timothy Allen, Ph.D., Department of Biomedical Engineering

General Problem

How should biological samples in genomics research best be categorized in order to create personalized downstream therapeutics?

Machine learning (ML) algorithms provide an advanced method to analyze genomics data and inform healthcare decisions. Genomics research holds great promise in identifying therapeutic targets for disease, however, algorithmic results are highly dependent on the quality of the data. Inherent biases can be present in data, and one key bias is the use of race and ethnicity. Historically, researchers in genomics categorized human-derived cell samples in their studies based on race, pushing the flawed race-as-genetics idea that certain individuals were genetically disadvantaged because of their race. This type of classification was found to be biologically invalid after sequencing the human genome in 2001, when it was discovered that all humans had nearly identical DNA. Though race is now considered a social construct in contemporary studies, racial categorization still occurs in some studies to this day, resulting in incorrect generalizations across ethnic groups. There are known variations in disease across individuals with similar socioeconomic backgrounds, identifying a need for personalized treatments. Racial categorization should be avoided and quantitative methods of categorization are needed for effective personalized treatments.

The technical topic involves creating a genomic search tool that processes a user-entered genome region and returns closely related regions using a deep-learning ML model. This work utilizes high-throughput sequencing data at the cellular level with no prior assumptions made based on the ancestry or ethnicity of cell samples. The focus of the STS topic is to examine contemporary genomic studies and policies surrounding biomedical research which inadvertently prolong the categorization of cell samples based on ethnicity. The STS and technical topic encapsulate separate goals: the former discourages race-as-genetics approaches while the latter encourages using measurable biological metrics for classification. Both work in tandem to support and provide an alternative method of categorization to discover personalized therapeutic targets.

Developing Deep-Learning Models to Create a Search Tool for Genomic Region Sets

Can a search engine tool be made to identify and return closely related genome regions based on a user query?

After sequencing the human genome, a multitude of questions arose about cellular differentiation. Nearly all cells in the human body have identical DNA but can exhibit vastly different phenotypes and roles. Epigenomics is the study of external modifications to DNA that affect gene expression. DNA is packaged and organized in the nucleus by histones, which are cylindrical proteins that DNA is coiled around (Martire & Banaszynski, 2020). Acetyl, phosphate, or methyl groups can bind to histone proteins, modifying how tightly the DNA is coiled. These are called epigenomic modifications and can occur based on a variety of different environmental factors and time scales. Protein complexes responsible for transcribing genes can easily bind to unraveled DNA, causing increased expression of genes that are exposed and decreased expression of genes that are tightly wound. Epigenomic data is generated through experiments such as ATAC-seq. These data are stored as text files named Browser-Extensible Data (BED) files. Each line in the file represents a genomic region, which is a stretch of DNA specified with start and end coordinates.

The library of publicly available ATAC-seq data has exploded due to the rise of high-throughput sequencing. A single BED file can contain millions of individual genomic regions, making raw data analysis impossible. Multiple epigenomic modifications can occur simultaneously anywhere on the genome, which makes it challenging to determine closely related regions of DNA that are affected. One key task is to find these relationships between region sets, which requires pooling multiple BED files from different studies together. This is complicated due to extremely large data dimensionality, and conventional raw data analysis cannot capture biologically significant information (Dozmorov, 2017).

Representation learning is a field in machine learning that has been successfully used to extract relationships from genomic data. In a recent study, a Region2Vec model was created to encode high-dimensional genomic region data into lower-dimensional vectors called embeddings. This is based on Word2Vec, a neural network model developed by Google to learn word associations. From this model, linear algebra methods can be used to determine the similarity between vectors to quantitatively identify how closely they are related. Region2Vec, compared to prior models, showed an increased ability to separate genomic region sets that were different and cluster region sets that were similar in a perturbation experiment (Gharavi et al., 2021). Novel deep learning architectures have not been tested for this purpose, providing an avenue to compare the performance of Region2Vec to other models.

Four deep-learning models will be developed to expand on prior work: a text-to-BED file neural network, direct encoder, diffusion model, and transformer. All models involve a neural network architecture with variations in layer organization. Each model will be evaluated using a Cluster Tendency Test (CTT) and a classification task with a threshold of 95% accuracy to determine how well each model can separate different cell types and whether cell types can be classified as the correct group. The optimal model will be incorporated into a search tool that returns closely related genomic regions based on a user-entered search. This work will significantly contribute to genomics research by creating a powerful, publicly available search tool to extract relationships from a large corpus of ATAC-seq studies. It will also increase the discovery of related DNA sequences that could serve as therapeutic disease targets.

The Detriment of Racial Categorization in Genomics Research and Factors Prolonging It

What regulatory groups are prolonging racial categorization in contemporary biomedical studies?

Background and Theoretical Framework

After the Human Genome Project, a paradigm shift occurred in thinking about race and ethnicity, which are defined as differences in physical appearance between individuals. All humans have 99% of their genomes in common, with phenotypic differences attributed to 0.1% of the variation (International Human Genome Sequencing Consortium, 2001). This revealed major flaws in relating biology to race. There are known variations in the frequency and severity of diseases among population groups. The term “population group” is defined as a group of individuals that are affected by similar socioeconomic or environmental impacts. Many medical conditions demonstrate this, with one example being cardiovascular disease (CVD). A study that analyzed the impact of social determinants on the development of CVD found that it was more prevalent in groups with higher socioeconomic burdens. These burdens include higher stress levels, increased exposure to smoking, inadequate medical care, and poor diet and exercise

(Kreatsoulas & Anand, 2010). Some forms of disease may present more aggressively in these individuals than others, presenting a need for personalized treatment. The question lies in how to best categorize patients to maximize personalized treatments for similar individuals.

Scientific racism in genetics is the idea that different races can be separated by gene expression or another factor of DNA. This idea has been prolonged incorrectly in many scientific studies. It is important to acknowledge that ethnicity and race are social constructs, not genetic. These extremely vague social constructs have been applied to biomedical research for the sake of simplicity, as humans are pattern-seeking and often classify groups based on prior biases or conclusions (Goodman, 2000). This sets a dangerous precedent and could cause untrue conclusions to be made that certain racial groups are fundamentally disadvantaged due to genetics when instead a multitude of other external factors are contributing to the results.

One recent paper examines the limitations and perspectives of using race and ethnicity in biomedical research, critiquing the lack of specificity in racial data collection (Gombault et al., 2023). The authors mention standardization guidelines created by the Food and Drug Administration (FDA) on racial and ethnic data collection. These were created in hopes that it would allow for personalized treatments to be developed (Office of the Commissioner, 2023), but the collection of racial data without paired social factors prolongs harmful biases in a field that should remain objective. The authors mention policies enacted by scientific journals, however, do not examine policies in depth, leaving a gap in the literature to be filled.

The goal of this project is to determine how regulations in research journals and the government have the inherent issue of racial categorization embedded into policy. A key idea that will be used to analyze this issue is value-neutrality thesis (VNT), or the theory that technology is never morally and politically neutral (Pitt, 2001). In this case, genomics represents technology. The human genome is objectively neutral, however historical biases and incorrect assumptions have bled into science and ultimately genomics, prolonging incorrect beliefs that ethnicity is an accurate human classifier. Social groups that play a large role are researchers and funding institutions, which will specifically be examined.

Another phenomenon related to this issue is responsible research and innovation. This is the idea that engineers and researchers must address current issues in their devices and be aware of issues the devices may cause in the future (Stilgoe et al., 2013). This specifically relates to researchers, who must always be aware of any biases present if using racial data in their studies and highlight the importance of social factors in conjunction with racial data in publications.

Methods

A policy analysis will be conducted to determine how regulatory bodies such as the FDA affected the trajectory of this issue (Patton et al., 2015). It is vital to examine current policies and regulations in scientific journals to identify how racial data is collected and classified in scientific studies. In supplement, a qualitative content analysis (QCA) will be conducted on existing literature in this area, as an extensive collection of articles exist about this issue. This analysis will allow for research and opinion articles on racial categorization to be condensed into key themes based on interpretation (Mayring, 2000). Both analyses will be mutually beneficial. The QCA will unveil responses and reactions from policies found from the policy analysis, while also uncovering reasons that guided policies to being enacted.

Conclusion

The findings from the STS project will provide insight to what laws and regulations exist in peer-reviewed journals and in government that inadvertently prolong racial categorization in research and what alternative features can be used to create personalized therapeutics. The goals of the technical project are to create a search tool that returns genomic regions based on a user-entered query using a deep learning model. The technical project is coupled with the STS project by providing an alternative classification metric genomics search tool that does not involve any assumptions based on ethnicity. This work will create awareness of policy shortcomings that lead to biased human classification and provide a quantitative way to categorize cell samples.

References

Dozmorov, Mn. G. (2017). Epigenomic annotation-based interpretation of genomic data: From enrichment analysis to machine learning. *Bioinformatics*, *33*(20), 3323–3330. https://doi.org/10.1093/bioinformatics/btx414

Gharavi, E., Gu, A., Zheng, G., Smith, J. P., Cho, H. J., Zhang, A., Brown, D. E., & Sheffield, N. C. (2021). Embeddings of genomic region sets capture rich biological associations in lower dimensions. *Bioinformatics*, *37*(23), 4299–4306. https://doi.org/10.1093/bioinformatics/btab439

Gombault, C., Grenet, G., Segurel, L., Duret, L., Gueyffier, F., Cathébras, P., Pontier, D., Mainbourg, S., Sanchez-Mazas, A., & Lega, J. (2023). Population designations in biomedical research: Limitations and perspectives. *Hla*, *101*(1), 3–15. https://doi.org/10.1111/tan.14852

Goodman, A. H. (2000). Why genes don't count (for racial differences in health). *American Journal of Public Health*, *90*(11), 1699–1702.

International Human Genome Sequencing Consortium. (2001). *Initial sequencing and analysis of the human genome*. https://www.nature.com/articles/35057062

Kreatsoulas, C., & Anand, S. S. (2010). The impact of social determinants on cardiovascular disease. *The Canadian Journal of Cardiology*, *26*(Suppl C), 8C-13C.

Martire, S., & Banaszynski, L. (2020). *The roles of histone variants in fine-tuning chromatin organization and function*. Nature Reviews Molecular Cell Biology. https://www.nature.com/articles/s41580-020-0262-8

Mayring, P. (2000). Qualitative Content Analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(2), Article 2. https://doi.org/10.17169/fqs-1.2.1089

Office of the Commissioner. (2023, August 10). *Collection of Race and Ethnicity Data in Clinical Trials*. FDA. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/collection-race-and-ethnicity-data-clinical-trials

Patton, C., Sawicki, D., & Clark, J. (2015). *Basic Methods of Policy Analysis and Planning* (3rd ed.). Routledge. https://doi.org/10.4324/9781315664736

Pitt, J. C. (2001). Thinking Through Technology: The Path Between Engineering and Philosophy. *International Studies in Philosophy*, *33*(2), 147–149. https://doi.org/10.5840/intstudphil200133228

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). *Developing a framework for responsible innovation*. Science Direct. https://www.sciencedirect.com/science/article/pii/S0048733313000930