The Application of Representation Learning for Generation of Genomic Region Data

Analysis of the Delay in Integration of Machine Learning in the Clinic

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
Zachary Mills

October 27, 2023

Technical Team Members: Lily Jones, Peneeta Wojcik, Caitlyn Fay

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

Dr. Rider Foley, Department of Engineering and Society

Dr. Nathan Sheffield, Department of Biomedical Engineering

Introduction

It is estimated by the World Health Organization that 10 in every 1000 people are affected by genetic disorders, corresponding to between 70 and 80 million people affected in the world (World Health Organization, 2020). This makes genetic research an important area with 3 billion dollars being spent on it each year (Hentschel, 2017). The amount of data from assay for transposase-accessible chromatin with sequencing (ATAC-seq) and chromatin immunoprecipitation (ChIP-seq) experiments has exploded over the past 10 years, increasing exponentially as sequencing technologies continue improving (Kodama et al., 2012). This increase in data can be attributed to the fact that, although the human genome has been fully sequenced, cells with the same genetic material can have be very different. (Alberts et al., 2002).

The difference between cells is not determined by the DNA code, but by their epigenomic states. Epigenomics is the study of modifications and associations of genomic sequences that are responsible for the differences between cells. Epigenomic modifications vary based on cell type and they bring up considerations to be made in analyzing genomic data. In fact, failing to adjust studies for differences can limit the accuracy (Qi & Teschendorff, 2022). This illustrates the importance of epigenomic data in the larger realm of genetic research.

Genomic region set data is becoming more prevalent in genomic research (Schwartzman & Tanay, 2015) because it gives a look at the epigenomic state to see which regions and genes are active in different physiological states and in distinct cell types. Even with the large mass of genomic region data being discovered, there still exist many cell and disease types with very little or no available data (Mitani & Haneuse, 2020). These data limitations exist because the conditions are rare or hard to access. For example, research into the rare autoimmune disease

alkaptonuria has stalled as there are not enough patients affected to perform high level analysis (Mitani & Haneuse, 2020).

For biomedical research to progress in finding treatments for these rare diseases, there needs to be development in the generation of data corresponding to them. This project proposes the idea of utilizing existing genomic region data to create a machine learning (ML) model capable of generating relevant genomic region sets to a user-entered search that can then be used in experiments for novel biomedical analysis. This has the potential to allow for more developed research into rare and hard to assess genetic diseases which could lead to improved patient treatment outcomes. However, the integration of machine learning and artificial intelligence into the medical industry has lagged behind other fields. Studies have shown that only 65% of medical practitioners are aware of clinical machine learning, while only 10-30% have reported using AI in the clinic (M. Chen et al., 2022). It is important to understand what causes this delay to streamline integration and improve healthcare in a way that is safe and efficient.

The current standard for storing genomic region sets is using browser extensible data (BED) files (UC Santa Cruz, 2022), which contain data from many genomic regions defined by chromosome name, start position, and end position. These regions were all generated in the same experiment and as such are biologically related. Typically, BED files are accompanied with meta-data annotations which contain text descriptions of the data, often including the cell line and antibodies associated with the experiment.

The goal of this project is to create an interface that allows generation of BED files from English language input. Currently, the only way to find relevant genomic data is to use a service like genemo.org (Zhang et al., 2016). This site utilizes pattern-matching to find similar BED files from an input BED file. It compares the input file to the entire Encyclopedia of DNA Elements

(ENCODE database), which is a comprehensive collection of data from the National Human Genome Research Institute (Abascal et al., 2020). While this system can accurately find similar BED files, it requires a BED file as input and only retrieves existing files rather than generating new data. This project seeks to improve on this in two ways: being able to use English language text as input and being able to generate novel data.

Representation Learning and Generative AI for Genomic Regions

To be able to accomplish the association from text to genomic regions, there must be a way to represent both entities as comparable dense vectors called embeddings. There already exists effective ways to represent text with technologies such as Google's word2vec (Mikolov et al., 2013). The project will utilize a system developed by researchers at the University of Virginia that can represent genomic region sets as 100-dimensional vectors that retain biological characteristics (Gharavi et al., 2021). This same team has also shown that it is possible to relate text to genomic region sets using vector embeddings although no generative work has been done (Gharavi et al., 2023).

To be able to train a generative model, a dataset of corresponding region set embeddings and text embeddings is needed. To do this, BED files from the ENCODE database, which aggregates data from thousands of different experiments, will be run through the existing model to source the region set embeddings, and the corresponding descriptions will be input into word2vec to create the text embeddings. This data will then be used to train four separate generative models using the PyTorch framework: a text to bed neural network, a direct encoder, a diffusion model, and a transformer. Each of these models will be attempted by a separate team member, with my focus being on the diffusion model. A diffusion model was chosen as it has

recently been shown to be productive in generating biodata such as protein backbones (Guo et al., 2023). These models are used to generate new data through an iterative process of noising and denoising data using a series of complex mathematical operations.

Once trained, these models will take in text embeddings and output a region set embedding to be decoded by the system designed by the UVA researchers. This process is illustrated in figure 1. The final system will only utilize one of these models which will be selected by determining which generates the most accurate BED files through comparison to existing BED files.

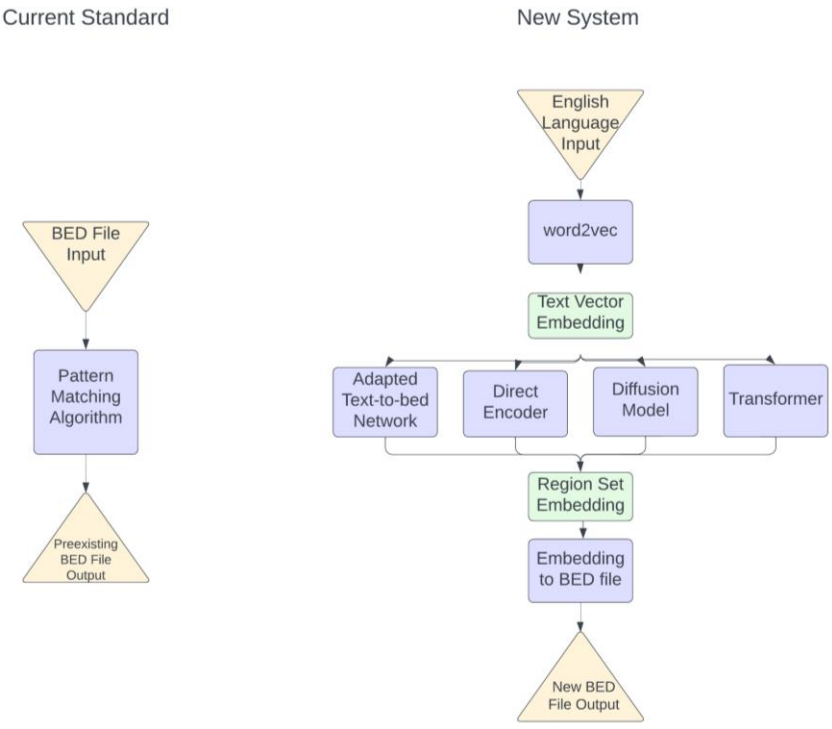

Figure 1: Comparison between the existing standard for finding BED files and the proposed new system. Note: Only one of the four models will be used in the final process all four are shown for completeness.

The final step is to develop a web interface, which allows users to input English language and generate BED files to allow easy access to researchers and other users. If successful, this capstone project would allow researchers to find relevant genomic region sets by inputting English language descriptions such as: “H34me3 histone modification on HCC1937 cell line.” The final product is a ML technology that will attempt to assimilate into the healthcare industry, which has been known to resist such integrations. This makes it important to analyze the effect that the product may have in the complex world of medicine as well as the development of the model and the collection of data.

Sociotechnical Analysis Using Infrastructure

There are various human and social factors to consider when developing this analysis. It is important to think about the impact of the many social groups connected to the project, especially researchers, physicians, and patients. One of the biggest concerns with any ML technology is data privacy. It has long been shown that patients are weary of allowing use of medical data as it contains sensitive information such as family history and health conditions (Sorani et al., 2015). For this reason, it is imperative that data is anonymized and protected. The data for this project is from an open-source collection of independent experiments. All the data is anonymous and contains only information needed for the analysis.

Another concern with ML in healthcare is the exasperation of disparity and inequity in the industry. Evidence has been found that ML algorithms regularly underperform on women and racial minorities due to unrepresentative data sets (I. Y. Chen et al., 2021). A further consideration is how ML is advertised to patients and research participants. Oftentimes they are marketed as “democratizing healthcare” and “personalized medicine,” but these labels are often

overselling the product. Studies have shown that marketing like this is purposefully misleading and has been used to coerce patients into giving access to data with purposefully vague promises (Roth & Bruni, 2022).

It is also important to examine how researchers and physicians will interact with ML technologies. Doctors are known to be a stubborn crowd, and do not like to adjust their routines (Roberts et al., 2021). Researchers are also likely to want to incorporate any new system with the ones they already use. This makes it important that conventions such as file type, interface, and accessibility are followed to ensure models are compatible with existing systems.

When considering the effects that these factors have on this technology, Susan Leigh Star's framework in The Ethnography of Infrastructure gives insight into this relationship (Star, 1999). In this article she lays out several aspects of infrastructure that relate technologies to society. The first aspect, built on an installed base, means that systems are not completely built from scratch; rather, they use existing technology to build upon in new ways to gain functionality. Essentially, old technologies are used as the basis for new ones. In the context of this project this can be seen in the adaptation of the existing word2vec to create vector embeddings for genomic regions (Gharavi et al., 2021). She also includes the idea of embeddedness, which means that a system is fit into other systems hiding their innerworkings from the eye of the public. In essence, new technology is sunken into existing systems and structures to seamlessly integrate with the infrastructure. The final search engine embodies the idea of embeddedness in that it hides the inner model behind a usable interface. The aspect of links with conventions of practice means that a technology or system must comply with certain standards or conventions whether legal or common practice. In other words, a new technology must fit within the common ways that the community utilizes similar technologies to

accommodate users. This applies to the project in that it will cooperate with conventions of practice such as the use of BED files to ensure that users are willing to adopt the new system.

Research Question and Methods

In recent years the amount of research into machine learning for medical applications has dramatically increased. There are now technologies that exist that can outperform physicians such as algorithms that detect eye diseases (Abràmoff et al., 2018) and apps that can diagnose skin cancer using only a cell phone camera (Freeman et al., 2020). Despite this, if you were to walk into the nearest doctor's office or hospital, it is unlikely that you would see anyone using a ML technology. In fact only 12% of hospital CIO's listed utilizing AI as an initiative in improving care (Stoltenberg Consulting, 2023). So, what has caused the integration of machine learning to be slower in the medical industry compared to other fields?

To answer this question, I will gather information from medical professionals employed at UVA health facilities utilizing my connections with research faculty. This will include conducting interviews with several doctors and other professionals through people I know within the health system. Data collected from the interviews will include answers to the questions: do they currently use machine learning in their workflow, are they opposed to machine learning in medicine, and whether they would be willing to integrate a new technology if it meant they had to change their standard procedures among others.

Analysis of this data would give insight into the current use of ML at UVA health, whether UVA health employees are willing to adopt new technologies, and what reasons are being given for reluctance. The responses will be analyzed in the context of Star's infrastructure framework to discover where integration goes wrong and what aspects are falling short. Once

data is analyzed, steps can be made to address concerns and shortcomings of technologies to increase the likelihood of adoption into practice.

Conclusion

As genomic region data continues to grow, many rare and inaccessible cell conditions remain without viable data. It is necessary to find a way to generate data for these conditions, so that more research can be done to increase knowledge and potentially develop life saving treatments. The aim of this project is to create an interface where such data can be generated using a machine learning model to rectify data deficiency. New machine learning technologies are being made every day, yet their full effects have not been realized. Analysis of data from healthcare professionals can bring insights into why this is the case and how these technologies can be designed and marketed in ways that make them more acceptable.

References

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Ai, R., Aken, B., Akiyama, J. A., Jammal, O. A., Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A. A., Barnes, I. H. A., … Weng, Z. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, *583*(7818), 699–710. https://doi.org/10.1038/s41586-020-2493-4

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Npj Digital Medicine*, *1*(1), 39. https://doi.org/10.1038/s41746-018-0040-6

Alberts, B., Johnson, A., & Lewis, J. (2002). *Molecular Biology of the Cell* (4th ed.). Garland Science.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, *4*(1), 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

Chen, M., Zhang, B., Cai, Z., Seery, S., Gonzalez, M. J., Ali, N. M., Ren, R., Qiao, Y., Xue, P., & Jiang, Y. (2022). Acceptance of clinical artificial intelligence among physicians and medical students: A systematic review with cross-sectional survey. *Frontiers in Medicine*, *9*, 990604. https://doi.org/10.3389/fmed.2022.990604

Freeman, K., Dinnes, J., Chuchu, N., Takwoingi, Y., Bayliss, S. E., Matin, R. N., Jain, A., Walter, F. M., Williams, H. C., & Deeks, J. J. (2020). Algorithm based smartphone apps to assess risk of skin cancer in adults: Systematic review of diagnostic accuracy studies. *BMJ*, m127. https://doi.org/10.1136/bmj.m127

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

Gharavi, E., LeRoy, N. J., Zheng, G., Zhang, A., Brown, D. E., & Sheffield, N. C. (2023). *Joint representation learning for retrieval and annotation of genomic interval sets* [Preprint]. Bioinformatics. https://doi.org/10.1101/2023.08.21.554131

Guo, Z., Liu, J., Wang, Y., Chen, M., Wang, D., Xu, D., & Cheng, J. (2023). *Diffusion Models in Bioinformatics: A New Wave of Deep Learning Revolution in Action* (arXiv:2302.10907). arXiv. http://arxiv.org/abs/2302.10907

Hentschel, R. (2017, March 2). *Why it might be time to reconsider the money spent on genetics research*. Melbourne School of Population and Global Health. https://mspgh.unimelb.edu.au/centres-institutes/centre-for-health-equity/news-and-events/archived-news/why-it-might-be-time-to-reconsider-the-money-spent-on-genetics-research

Kodama, Y., Shumway, M., Leinonen, R., & on behalf of the International Nucleotide Sequence Database Collaboration. (2012). The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, *40*(D1), D54–D56. https://doi.org/10.1093/nar/gkr854

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mitani, A. A., & Haneuse, S. (2020). Small Data Challenges of Studying Rare Diseases. *JAMA Network Open*, *3*(3), e201965. https://doi.org/10.1001/jamanetworkopen.2020.1965

Qi, L., & Teschendorff, A. E. (2022). Cell-type heterogeneity: Why we should adjust for it in epigenome and biomarker studies. *Clinical Epigenetics*, *14*(1), 31. https://doi.org/10.1186/s13148-022-01253-3

Roberts, M., Zeitzer, S., & Hohensee, A. (2021, April 15). *Common Reasons Doctors Don't Adopt New Health Tech & How to Address Them*. Health Connective. https://www.healthconnectivetech.com/insights/common-reasons-doctors-dont-adopt-new-health-tech-how-to-address-them/

Roth, P. H., & Bruni, T. (2022). Participation, Empowerment, and Evidence in the Current Discourse on Personalized Medicine: A Critique of "Democratizing Healthcare." *Science, Technology, & Human Values*, *47*(5), 1033–1056. https://doi.org/10.1177/01622439211023568

Schwartzman, O., & Tanay, A. (2015). Single-cell epigenomics: Techniques and emerging applications. *Nature Reviews Genetics*, *16*(12), 716–726. https://doi.org/10.1038/nrg3980

Sorani, M. D., Yue, J. K., Sharma, S., Manley, G. T., Ferguson, A. R., Cooper, S. R., Dams-O'Connor, K., Gordon, W. A., Lingsma, H. F., Maas, A. I. R., Menon, D. K., Morabito, D. J., Mukherjee, P., Okonkwo, D. O., Puccio, A. M., Valadka, A. B., & Yuh, E. L. (2015). Genetic Data Sharing and Privacy. *Neuroinformatics*, *13*(1), 1–6. https://doi.org/10.1007/s12021-014-9248-z

Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326

Stoltenberg Consulting. (2023). *11th Annual Health IT Industry Outlook Survey Report*. Stoltenberg Consulting. https://www2.stoltenberg.com/11thHITOutlookReport

UC Santa Cruz. (2022). *Genome Browser FAQ*. Genome Browser.

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

World Health Organization. (2020). *Genes and human diseases*. World Health Organization. https://www.who.int/genomics/public/geneticdiseases/en/index2.html

Zhang, Y., Cao, X., & Zhong, S. (2016). GeNemo: A search engine for web-based functional genomic data. *Nucleic Acids Research*, *44*(W1), W122-127. https://doi.org/10.1093/nar/gkw299