Deep Learning for Predicting Gene Expression from Histone Modifications

The Ethical Ramifications of the Commodification of DNA

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 Computer Science

By

Meghan Anderson

October 17, 2021

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

Sean Ferguson, Department of Engineering and Society

Yanjun Qi, Department of Computer Science

Introduction

In recent years, companies like 23andMe and Ancestry which allow customers to submit DNA samples for analysis have contributed to the massive amounts of genomic data available to researchers. What once was difficult data to obtain has now been commodified as consumers willingly give up their DNA destined to be sold to other companies (Pandya, 2019). The ever-increasing stores of DNA data have the potential to revolutionize the field of biomedicine as this influx of information coincides with a boom in the development of machine learning algorithms. The DNA data analysis made possible by these developments in computer science may possess key answers for the treatment of both genetic diseases and cancer.

Various contemporary research projects are in progress which implement machine learning algorithms to predict gene expression based on histone modifications. The goal of such research projects is to generate efficient, accurate models for biomedical researchers to further their work with histones and gene expression. One particular research project in this subject involves the implementation of a Convolutional Neural Network (CNN) as a predictive model for gene expression. I have the privilege of participating in this research project as a code developer. My role will initially involve learning about the file types and data manipulation tactics employed in bioinformatics. Once I have amassed enough knowledge, I will begin improving the current program to ensure it may be utilized effectively by research scientists.

Beyond these technical responsibilities, as a computer scientist with a concentration in bioinformatics, it is imperative I have a deep understanding of the structures that make work in this field possible. Specifically, knowledge of where the data being handled comes from and who stands to benefit from the models constructed is crucial. In this paper, I will analyze the

blossoming relationships between DNA testing companies and pharmaceutical companies as human genes are commercialized.

Deep Learning for Predicting Gene Expression from Histone Modifications

The DeepChrome program to predict gene expression based on histone modifications was first developed by Ritambhara Singh, Jack Lanchantin, Gabriel Robins, and Yanjun Qi and was published by the Oxford University Press (Singh *et.al.*, 2016). In order to situate the significance of this research in the field of biomedicine, a short explanation of the role of histones in gene regulation is necessary. Histone proteins hold strands of DNA together. These proteins are prone to a chemical process known as methylation which may lead to changes in spatial arrangement, and consequently changes in binding and gene regulation. These modifications can be detrimental to transcription pathways and are sometimes associated with the proliferation of cancerous cells. What is particularly interesting about these types of gene modifications is that unlike mutations they may be reversible. Discovering how to predict histone modifications on gene expression may open the door for developing new techniques to treat certain types of cancer by reversing adverse histone modifications (Neganova *et.al.*, 2020).

The pursuit of deep learning algorithms with the ability to predict the effects of non-coding variants of genes is anything but novel. In the past, researchers have implemented algorithms such as Support Vector Machines (SVM) (Cheng *et.al.*, 2011). Others made use of a Random Forest algorithm to produce predictive models for histone activity (Dong *et.al.*, 2012). The researchers in the study cited earlier took a different approach and made use of

Convolutional Neural Networks (CNNs) to develop their model. Through testing, the researchers found their predictive model performed better than the previously tested models and provided a more optimized program to analyze the data (Singh *et.al.*, 2016).

For the most part, the program has been fully developed already to analyze histone data. However, the program lacks some sophistication and is highly disorganized and difficult to follow. The current tasks of the research team led by Yanjun Qi involve refining the DeepChrome program with the intention of creating a system useful to other researchers. To achieve this goal the program currently exists as a GitHub repository, so public users may clone and edit on their own. This allows for outside groups to assess and test the code and make suggestions for improvements. The team has set out a plan for each of the undergraduate researchers involved in the project to analyze the program by testing it and reading through the code individually before adjustments are made to it based on conclusions the group draws after individual research.

As a member of this research team, my responsibilities will include reformatting and developing the program in ways that make it more readable, accessible, and improve its performance. Debugging portions of the program is also a crucial task. Considering the general complexity of biological concepts the program depends, the size of the datasets, and the complexity of the code itself, I must first master usage of the bioinformatics file types and gain an understanding of how the program works. FASTQ files are files that store a biological sequence and corresponding values associated with that sequence in a text-based format. They are just one of many file formats I will need to familiarize myself with for this project. The majority of the fall semester in 2021 will consist of educating myself on all the material necessary to modify the program. This may also include adding docstrings to the code to help

guide other developers through complicated functions or methods. In the spring semester of 2022, I will contribute to the project by researching ways to improve the deep learning model by comparing DeepChrome to other machine learning models which incorporate CNNs. Additionally, I will be making changes to the existing code to improve its overall performance and usability based on results of testing metrics I will generate by collaborating with my capstone project leader Dr. Yanjun Qi.

Eventually the model generated will be efficient and accurate enough to be incorporated into the methodology of biomedical research groups.

The Ethical Ramifications of the Commodification of DNA

Henrietta Lacks was diagnosed with cervical cancer and sample cells were taken from her cervix during a routine examination. These cells were studied and her DNA was sequenced without her knowledge. There was something unique about her cells which enabled them to grow and replicate in a culture outside the body. This phenomenon had not been witnessed prior to the study of cells that came to be known as “HeLa cells”, making them of particular interest for researchers. Her cells were widely distributed for study without her consent and in 2013 her DNA was posted publicly. Her family continued to live in poverty, experienced low access to healthcare, and were never able to receive financial benefits for the contributions of HeLa cells to modern science (Skloot, 2010). The exceptional case of Henrietta Lacks serves as the jumping off point for the discussion of equity in the use of human biospecimens in research (Beskow, 2016). Such a discussion is timely because as DNA data being collected by genetic sequencing companies increases so too does the potential for the abuse of power seen in the Henrietta Lacks

case occurring on a mass scale. Through the lens of sociotechnical systems with a focus on techno politics, I will analyze the collaboration of DNA testing companies and pharmaceutical companies as DNA has become an increasingly powerful commodity in U.S. markets. Where does the individual's right to reap the benefits of research conducted using their DNA fall in the for-profit business model?

Shobita Parthasarathy's extensive work with patents explores the way patents and biotechnologies coalesce. Parthasarathy's *Patent Politics: Life Forms, Markets, and the Public Interest in the United States and Europe* employs a technopolitical framework to analyze the influence of patent systems in the United States and Europe on their respective markets in biotechnology. She contends current U.S. court rulings have favored the patent system for altered ("not natural") genes which in turn has complicated the fight for legal rights surrounding genetic data initiated by the ACLU and other institutions (Parthasarathy, 2017). Rather than having a unilateral policy preventing what are essentially monopolies over biospecimens for research, agencies created to protect the populace must go through legislative means instead. This is no small feat as regulatory legislation may take years to be created and even longer to get the necessary support to be passed. The rulings accentuate a power dynamic in which access to resources for research are concentrated in already powerful establishments. Parthasarathy's work has guided my own work in examining DNA data's fate through the technopolitical structures that govern genetic research.

Identification of the key players acting in the human genome market in the United States is the first step in assessing the destiny of consumers and their DNA. The premise for this paper's analysis rests on the notion that pharmaceutical companies are currently purchasing DNA data from private companies that offer genome sequencing services. The growing

allegiances between DNA data collectors and the pharmaceutical companies that bid on them can be visualized in the graphic of Figure 1 (Roland, 2019).

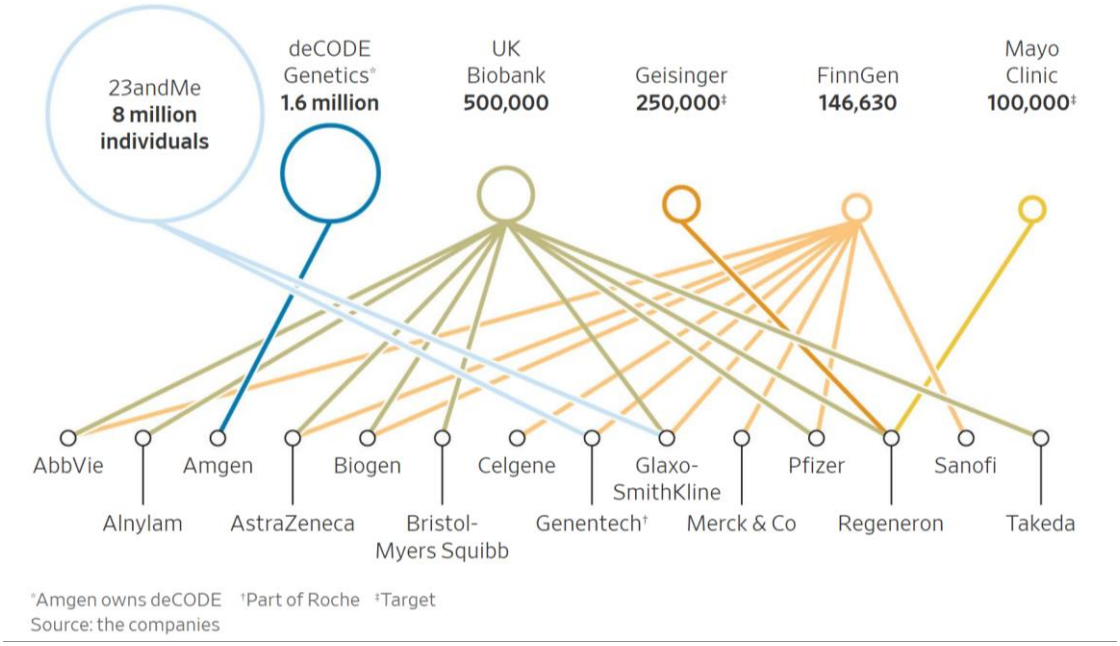

Figure 1. Selected genetic databases and their industry partners.

The partnerships are a natural progression as companies with stores of consumer DNA data want a way to profit off of the DNA data they collect while pharmaceutical companies are compelled to purchase DNA data for drug research. The general dynamics of these partnerships and their technopolitical origins can be traced back to neoliberalism and its impact on biotechnology in the United States. Neoliberalism is defined as a political approach with an emphasis on the deregulation and minimization of governmental bodies that allow for a partially free market system to take place. Neoliberalism, as the dominant political ideology in the U.S., has resulted in a dynamic wherein the government prioritizes the interests of the biotechnical private sector in lieu of communities. The United States' techno-positive culture and desire for global prestige only stands to exacerbate this phenomenon (Meghani, 2017). Consequently, these types of

alliances between corporations that share consumer data are not prohibited and inherently favor corporate incentives over individuals.

Pharmaceutical companies, like practically all private firms, are entities whose main incentive is increasing the profit margin. This calls into question the altruism of their research efforts surrounding the human genome. In the research being conducted, should profit incentives take precedence over universally beneficial outcomes?

In order to conceptualize what is currently at stake in the presiding sociotechnical system of the United States consider the following scenario:

John decides to submit his DNA to be sequenced by Company A out of sheer curiosity. Company A then proceeds to sell his DNA in a dataset to a pharmaceutical company, Company B. In their research procedure, Company B makes a slight modification to John's DNA that enables them to find a viable gene therapy method in the treatment of cancer. Company B may patent and profit millions off of a slight change made to John's DNA. Meanwhile John gets his sequencing results back, discovering he is destined to get cancer. John may never know how influential his DNA has been and may pay hundreds of thousands of dollars on a treatment that only exists because of his miraculous DNA. The practices of both Company A and Company B are particularly exploitative and should be concerning to everyone, not just those who willingly decide to submit their DNA to private entities.

Considering current trends, individuals who submit their DNA to a company collaborating with a pharmaceutical company are more likely to be unbeknownst to any usage of their DNA, than they are to receiving any direct benefit for their contributions. Ultimately, the

search for ways to produce genetically based knowledge that is collectively beneficial without being extractive must be pursued.

Next Steps

Looking ahead at the months leading up to graduation, the table below details some of the objectives I hope to have accomplished by the end of my undergraduate career.

	Capstone Research	STS Topic
November 2021	• Increase familiarity with bioinformatics file format • Add docstrings to the DeepChrome program to aid others (and myself) in following code's functions	• Turn in the final prospectus paper • Receive feedback on the final prospectus and get signatures
December 2021	• Present what I've learned over the past semester to the capstone advisor to create a plan for the spring semester	• Present to my STS course peers on my STS topic and research
January 2021	• Begin coursework in biological computing to improve my skills for the Capstone project	• Contact Shobita Parthasarathy for a short interview to get an expert's perspective on DNA research in the hands of pharmaceutical companies
February 2021		• Research gene patenting and corporate ties globally • Research U.S. healthcare system as a player in

		sociotechnical system for genetics market
April 2021	• Present an overview of my contributions to the DeepChrome project	• Turn in the STS portfolio

The nature of my capstone research is somewhat open-ended and will be dependent on how my advisor and I decide to proceed at the end of this semester. I have provided as much detail as is possible at the moment. All in all, I am hoping I can contribute a significant amount of work to the project that will enable the program's usage on a broader scale.

References

Beskow, L. M. (2016, August 31). *Lessons from Hela cells: The Ethics and policy of Biospecimens*. Annual review of genomics and human genetics. Retrieved October 17, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5072843/.

Cheng, C., Yan, K.-K., Yip, K. Y., Rozowsky, J., Alexander, R., Shou, C., & Gerstein, M. (2011, February 16). *A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets*. Genome Biology. Retrieved October 30, 2021, from https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-2-r15.

Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigó, R., Birney, E., & Weng, Z. (2012, September 5). *Modeling gene expression using chromatin features in various cellular contexts*. Genome Biology. Retrieved October 30, 2021, from https://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-9-r53.

Lave, R., Mirowski, P., & Randalls, S. (2010, September 28). *Introduction: STS and neoliberal science - researchgate*. Sage Journals. Retrieved October 5, 2021, from https://journals.sagepub.com/doi/abs/10.1177/0306312710378549.

Meghani, Z. (2017, November 21). *Genetically engineered animals, drugs, and neoliberalism: The need for a new biotechnology regulatory policy framework*. Journal of Agricultural and Environmental Ethics. Retrieved November 3, 2021, from https://link.springer.com/article/10.1007%2Fs10806-017-9696-1.

Neganova, M. E., Klochkov, S. G., Aleksandrova, Y. R., & Aliev, G. (2020, August 16). *Histone modifications in epigenetic regulation of cancer: Perspectives and achieved progress*. Seminars in Cancer Biology. Retrieved October 30, 2021, from https://www.sciencedirect.com/science/article/pii/S1044579X20301760.

Pandya, J. (2019, April 17). *The rise of genetic testing companies and DNA data race*. Forbes. Retrieved October 30, 2021, from https://www.forbes.com/sites/cognitiveworld/2019/04/01/the-rise-of-genetic-testing-companies-and-dna-data-race/?sh=6f8c41c72afb.

Parthasarathy, S. (2017-02-21). Confronting the Questions of Life-Form Patentability. In *Patent Politics: Life Forms, Markets, and the Public Interest in the United States and Europe.* University of Chicago Press. Retrieved 3 Nov. 2021, from https://chicagouniversitypressscholarship.com.proxy01.its.virginia.edu/view/10.7208/chicago/9780226437996.001.0001/upso-9780226437859-chapter-003.

Roland, D. (2019, July 22). *How drug companies are using your DNA to make new medicine*. The Wall Street Journal. Retrieved October 5, 2021, from

https://www.wsj.com/articles/23andme-glaxo-mine-dna-data-in-hunt-for-new-drugs-11563879881.

Singh, R., Lanchantin, J., Robins, G., & Qi, Y. (2016, August 29). *Deepchrome: Deep-learning for predicting gene expression from histone modifications*. OUP Academic. Retrieved October 17, 2021, from https://academic.oup.com/bioinformatics/article/32/17/i639/2450757?login=true.

Skloot, R. (2010). *The Immortal Life of Henrietta Lacks*. Broadway Paperbacks, an imprint of the Crown Publishing Group, a division of Random House, Inc.

Zs. (2021, January 27). *DNA-based data is a hot commodity, and pharma is buying*. ZS. Retrieved October 5, 2021, from https://www.zs.com/insights/dna-based-data-is-a-hot-commodity-and-pharma-is-buying.