Application and Implementation of Physical Trait Predicting Algorithms

(Technical Paper)

Attitudes Towards Trait Prediction via Genomic Samples

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science

University of Virginia | Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Paul Vann

Fall, 2022

Department of Computer Science

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

Signature ____ Date __________

Paul Vann

Signature ____ Date __________

Joshua Earle, Department of Engineering and Society

Introduction

Genomic science has been growing in popularity as new technologies are developed to analyze human DNA. Companies like 23andMe and Ancestry.com utilize these technologies to predict where a person's lineage comes from around the globe or locate a distant familial relative. The technologies for performing these tasks are only improving, revealing more information about a person or a sample. One such improvement is the identification of physical features based on a genomic sample. Researchers in the past have discovered specific segments of the genome that reveal a person's eye color and skin color with high probability. While these are big strides in the direction of trait prediction, genetic researchers propose that more complex physical traits such as jaw structure and others require a complex machine learning algorithm and a significant amount of data. This ability to determine an individual's physical features from their genome poses a major ethical question: At what point does an algorithm that has these features violate the privacy of those who it is used against?

My technical project tests the plausibility of such an algorithm being built, and tests the extent to which it can accurately identify physical features of a group of people. I plan to build a reinforcement learning algorithm that uses a large dataset of genomic data matched with physical traits and train the model to accurately predict the physical traits associated with genomic data not found in the training dataset. While this algorithm does have the potential to fail at this task, the goal of the technical project is to test its plausibility using the most advanced and widely studied form of machine learning today, reinforcement learning.

My STS project targets the ethical dilemma with a trait predicting algorithm as described above, which is the privacy of genomic data and whether it is ethical to be able to identify a person based on their genomic data or a DNA sample. This technology would allow for a company such as 23andMe to create a physical model of what a person looks like solely based on their DNA sample. This form of data is a very large asset for many big data companies, making this algorithm’s implementation a big privacy concern. I will focus on this privacy concern and attempt to answer the question: At what point does a tool like this become too much of a privacy concern? And how can its implementation and design be modified to avoid this privacy issue?

Overall my main goal in pursuing this project is to push the forefront of genomic research even further, adding a new technology to the field that has a lot of potential implications. For one, this type of algorithm can be used to identify a criminal based solely on a DNA sample, reducing the number of criminals that escape or are never arrested by a significant factor. Another possible use is learning more about organisms and people that came before us, and what they may have looked like. This thesis will cover the possibility of the algorithm's implementation, its potential uses, and the ethical concerns with it.

Technical Topic

Introduction

DNA and genetic data is traditionally viewed as a means of determining lineage or family trees, but what's rarely discussed is the potential to identify complex physical features from genomic data. In the United States there are over 60,000 murders unsolved each year, many of which have DNA from the perpetrator(s) on the scene (Hargrove, 2021). New fossils and evidence of organisms from hundreds to even thousands of years ago are also found every month, providing DNA and small fragments of genetic material from the past. While this technology does exist today, there are many limitations to it. For example, DNA from a crime scene can only be used if there is matching genetic data within a criminal database. In addition, little can be determined about an organism's physical features without a distinct fossil or bone structure. Physical trait identification via a reinforcement learning algorithm attempts to solve this problem.

For this technical project, I will use a reinforcement learning model to assist in identifying some of these traits. The model will target traits such as jaw structure, eye color, hair color, and skin color to put together a fairly sophisticated model of what a person's face looks like. There are numerous articles and reports highlighting the specific segments of a person's genome that define these features, however it is very difficult to perform manually. Therefore with a significant amount of genomic data corresponding to physical traits and research on correlations between genotypes and phenotypes, a highly accurate model could be engineered to accurately predict phenotypes (Nicodemus, 2009).

While I will not be building this tool at the scale where it can be used in a professional environment and be full proof, it will set up a lot of the necessary research and frameworks for producing a highly accurate version of this algorithm. A reinforcement learning algorithm, trained with much more data could pave the way to generating highly sophisticated models of a person's facial features from a genome sample.

Methods

The reinforcement learning model I will be using to tackle this problem is based off of the course material for CS 4501-008, Introduction to Reinforcement Learning. The content taught in this course includes numerous reinforcement learning models, what models are the best for certain situations, and how to properly tune each model. The problem at hand is identified as a multi-arm bandit problem. A multi-arm bandit problem can be defined as one where there are multiple options, each with a unique probability of "reward" (Kuleshov, 2014). I am choosing this model as the problem requires iterating through multiple genome sequences and determining which ones have the highest probability of predicting physical traits. The model I am specifically planning on using for this project is either a UCB (Upper Confidence Bound) model or a Thompson Sampling model, both of which are used to solve multi-arm bandit problems (Kuleshov, 2014).

In addition to the model, I will also need to collect a large amount of accurate and comprehensive genomic data, in order to properly train the model. I have identified multiple sources for this data, one of which is John Hopkins University. They provide a source with over ten different sources for genomic data, which could very likely be suitable for this project (John Hopkins Medicine, 2021). Another reputable source, known for having datasets for almost any machine learning related project is Kaggle.com, which I intend to utilize at least for some part of the dataset. Finally, if a suitable dataset is not found from one of the two listed sources, I will reach out to published authors who have performed similar research and discuss the possibility of them assisting in putting together a dataset for the model.

With the timeframe for this project in mind, as well as the level of expertise I have, I plan to put together a functional model that, at a basic level, is able to identify these physical traits. I also intend to fully document the algorithm and the dataset I am provided, with the goal of assisting in future research. If time permits, I plan to include a UI (User Interface), which will make the tool much easier to use and test.

STS Topic

Introduction

To go hand in hand with the technical project described above, I plan to focus on the ethical implications of a trait predicting algorithm in respect to privacy. Genetic data is very sensitive, making it a very large privacy issue if big data corporations or groups are able to more readily access and gain knowledge from DNA samples. As Kay writes in her book “Who wrote the book of Life”, while genomic projects can have a significant impact, the human genome is not just an information system or a language to be processed by an algorithm, and should not always be treated as such (Kay, 2000). A trait predicting algorithm allows for knowledge to be gained about an individual by enabling individuals and groups with access to genetic information to create models and/or images of people based on their DNA sample. This applies to many organizations, including but not limited to 23andMe and Ancestry.com. A trait predicting algorithm also allows for this potential violation of privacy simply by providing another technology that utilizes genetic information and data, opening a door for more genetic data to be obtained, analyzed, and released (Naveed, 2015).

STS Framework

I will be focusing on two main STS methodologies. These include Technological Determinism and Social Construction of Technology. Technological Determinism is a theory that suggests that technology and the way it is designed and constructed has an impact on societal and cultural development. In direct relation to this project, I intend to analyze and research the impact that both genetic algorithms and genetic privacy laws can have on society as technology in the genetic field continues to improve. Furthermore, I will heavily consider the social and cultural implications of an algorithm like the one described in the technical project section. As

Nelson discusses in his book, “The social life of DNA”, genetic data and DNA can in some cases be used either by an individual or an algorithm in a manner that has a negative impact on a certain population (Nelson, 2016). This is something that I intend to avoid as I pursue this project.

Furthermore, social construction of technology is a theory that suggests the use and development of technology is driven by social and cultural factors, causing some technologies to be very socially-positive and others to be very socially-negative. In the context of this project, I intend to research and analyze unique methods for managing the use of a genetic algorithm including control access mechanisms, de-identification algorithms, and policies dictating the use of a genetic algorithm. Furthermore, I intend to analyze the risks of implementing an algorithm similar to the one discussed in the technical project section, and determine viable methods for reducing these risks.

Methods

There are two research methods I intend to pursue for this portion of the project. The first method is performing research on the privacy policies surrounding the sale and distribution of genetic data, as well as the policies of companies that actually handle this type of data. I will do this by researching both state and federal legislation on genetic data, and what is legally allowed and not allowed. I will then look at a set of ten specific companies and groups that handle genetic data, and view their policies on what they are allowed to do with the genetic data they obtain. Finally, I will connect the research on the legislation and policies back to the characteristics of a trait predicting algorithm, and use that to determine what impacts this algorithm may have, and what risks there are.

Secondly, I will perform a poll on a group of students at UVA asking questions such as: On a scale of 1-10 how uncomfortable would you be if your DNA sample was shared? How likely on a scale of 1-10 would you be to share your DNA sample today? And How likely on a scale of 1-10 would you be to share your DNA sample if a trait predicting algorithm existed? This poll will allow me to gauge how people feel about the privacy of their DNA, as well as compare their thoughts on sharing a sample of their DNA with and without a trait predicting algorithm. This comparison can be used to argue whether or not a trait predicting algorithm actually makes a difference in genetic information privacy and how people feel about it.

The data from this poll as well as the information retrieved from the policies and legislations discussed above, can be used to fairly accurately gauge the response to a trait predicting algorithm, and how it affects people's feelings on privacy.

Conclusion

In conclusion, both the technical component and the STS component of this project serve a crucial role in determining the feasibility of creating a trait predicting algorithm, both technologically and socially. As discussed previously, this project is not a full proof method for predicting traits from a genome, nor does it give a direct answer on whether or not a tool like this could exist ethically. However, it should answer some of the questions presented in this thesis such as: Is it technologically feasible to predict physical traits from a genomic sequence? Is a reinforcement learning model the correct approach to this problem? And is it ethically plausible for a technology like this to be built and used?

Key Texts

Erlich, Y., Williams, J. B., Glazer, D., Yocum, K., Farahany, N., Olson, M., … Kain, R. C. (11 2014). Redefining Genomic Privacy: Trust and Empowerment. PLOS Biology, 12(11), 1–5. doi:10.1371/journal.pbio.1001983

This paper introduces a unique take on a method for protecting the genetic information of individuals compared to other approaches. While other approaches have focused on zero-trust models and de-identification of genomic data, this approach focuses more on a trust-centric approach. The main goal of this is for there to be more transparency for researchers who are attempting to develop novel solutions and guaranteed trust and privacy for individuals who are sharing their genomic data. This is relevant to this project in that a method like this could be very beneficial to implement when building the model described in the technical project. Furthermore, this is another approach that can be discussed in the STS portion of my thesis on how to protect genetic information.

Nicodemus, K. K., & Malley, J. D. (2009, May). Predictor correlation impacts machine learning algorithms: implications for genomic studies. Bioinformatics, 25(15), 1884–1890. doi:10.1093/bioinformatics/btp331

This paper focuses on the use of large-number predictors for predicting phenotypes using genomic data. Furthermore, this paper discusses a machine learning technique for being able to handle a large dataset and process it thoroughly to make accurate predictions based on Genomic data. The paper includes full results and implementation details for their project, as well as background information highlighting their decision making for this implementation. This will be very valuable for my technical project in that my technical project performs a similar task and also will be processing a large amount of information and a large number of predictors. Furthermore, I can use the implementation and methodology components of the paper to assist in preparing my algorithm.

Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J.P., Wang, X. (2015). Privacy in the Genomic Era. ACM Comput. Surv., 48(1). doi:10.1145/2767007

This paper, similar to others in this bibliography, discusses the issues with genetic privacy in today's day and age as genetic information is being used more and more in the medical and research fields. However this paper also discusses other privacy issues outside of personal identification with genomic information including being able to determine if an individual is at risk for certain medical issues, determining familial relationships, and more. The paper goes on further to discuss methods for protecting against these forms of genetic privacy attacks, going over the current existing methods today. This will help significantly with writing the STS portion of my thesis in that it discusses multiple state-of-the-art solutions for protecting genetic privacy commonly used today.

Pośpiech, E., Teisseyre, P., Mielniczuk, J., & Branicki, W. (2022). Predicting Physical Appearance from DNA Data—Towards Genomic Solutions. Genes, 13(1). doi:10.3390/genes13010121

The main argument of this article is that genomic data found in an investigation or a crime scene has the potential to be used to predict physical phenotypes or visible traits. The article further mentions that there are limitations to this however and that this trait prediction is currently only possible for traits that are influenced by fewer genes.

However, using machine learning and a very large dataset there is potential for a very accurate predictor to be created and more sophisticated traits to be identified from genomic data. This is relevant to the project at hand because the technical portion of this project focuses directly on creating such a machine learning model. This article also details implementation techniques and training techniques for building such a model which will be a very useful tool when attempting to create the algorithm.

References

Clayton, E. W., Evans, B. J., Hazel, J. W., & Rothstein, M. A. (2019, May). The law of genetic privacy: applications, implications, and limitations. Journal of Law and the Biosciences, 6(1), 1–36. doi:10.1093/jlb/lsz007

Hargrove, T. (2021, May). Cold case homicide stats - project: Cold case. Project Cold Case. Retrieved December 16, 2022, from https://projectcoldcase.org/cold-case-homicide-stats/

John Hopkins Medicine. (2021, July). Welch Medical Library Guides: Data Resources for Health and Medical Research: Genomic data. - Welch Medical Library Guides at Johns Hopkins University-Welch Medical Library. Retrieved December 16, 2022, from https://browse.welch.jhmi.edu/data-resources/genomic-data

Kay, L. E. (2000). Who wrote The book of life?: A history of the genetic code. Stanford University Press.

Kuleshov, V., & Precup, D. (2014). Algorithms for multi-armed bandit problems. doi:10.48550/ARXIV.1402.6028

Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J.P., Wang, X. (2015). Privacy in the Genomic Era. ACM Comput. Surv., 48(1). doi:10.1145/2767007

Nelson, A. (2016). The social life of Dna: Race, reparations, and reconciliation after the genome. Beacon Press.

Nicodemus, K. K., & Malley, J. D. (2009, May). Predictor correlation impacts machine learning algorithms: implications for genomic studies. Bioinformatics, 25(15), 1884–1890. doi:10.1093/bioinformatics/btp331

Pośpiech, E., Teisseyre, P., Mielniczuk, J., & Branicki, W. (2022). Predicting Physical Appearance from DNA Data—Towards Genomic Solutions. Genes, 13(1). doi:10.3390/genes13010121