Prospectus

Similarity Score for Epigenomic Region Sets
(Technical Topic)

Commercial Genetic Testing and the Golden State Killer
(STS Topic)

By

Aaron Gu

11/25/19

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.

Signed: ____

Approved: _____________________________________ Date _______________________
Benjamin Laugelli, Department of Engineering and Society

Approved: _____________________________________ Date _______________________
Aidong Zhang, Department of Computer Science STS Prospectus

Introduction

Twelve years ago when the human genome was first sequenced, only 1.5% of the genome could be annotated as protein encoding, whereas the rest of the genome was thought to be mostly "junk." Today, however, with large advances in the field of epigenomics, nearly half of the genome is predicted to play a role in protein synthesis through biochemical processes and regulatory activity. This alludes to the main question of epigenomics: how the same genome gives rise to over 200 different types of cells (Rivera & Ren, 2013). As the pace of research has increased, the number of epigenomic datasets has grown exponentially in recent years ("ENCODE Project Overview", 2019). Though there are online data catalogs, there has not been a standardized way to quickly compare data.

Currently, researchers write programs ad hoc, which wastes time and may not be accurate. The pace of research is slowed by this, and some relationships between epigenomic data remain undiscovered. My technical research project aims to solve this problem by creating a similarity score for epigenomic datasets. The score first involves converting epigenomic data into vectors, which enables fast comparison so that researchers can query massive amounts of data at once. However, a purely technical solution would not address the problem fully. There are social factors to consider too, such as how different groups can potentially use the technology to their benefit.

To help analyze complications around epigenomic data analysis, a parallel can be drawn to genomic data by looking at DNA testing services offered by companies like 23andMe. The way these services work is that customers send in a spit sample through the mail, which is then sequenced and compared to millions of other customers' DNA profiles (Regalado, 2019). DNA testing can help everyday consumers find long lost relatives and discover ancestry, but recently it has been helping a different group of users - the police. In my science, technology, and society (STS) research project, I will use Kline and Pinch's theory of users as agents (1996) to examine how police came to use the DNA testing services to solve crimes, which was unexpected for the original creators of the technology.

Addressing the sociotechnical nature of the research problem will have many benefits. For example, if my technical research project is successful in developing a similarity score for epigenomic data, the same issues from DNA testing could arise when unexpected user groups begin sending in epigenomic profiles to compare. By studying the case of DNA testing companies and the police, these kinds of problems would be prevented.

For my technical project, I propose a novel similarity scoring method for epigenomic datasets, which will allow search queries to be quickly executed. In my STS project, I analyze how an unexpected user group, the police, began to use DNA testing services. With both of these research projects completed, an epigenomic data comparison service could be implemented that will give researchers and consumers more knowledge about epigenomic data.

Technical Problem

Epigenomic data is often represented as regions which have different significances depending on the experiment they were produced by. For example, ChipSeq produces regions of protein interaction, and DNaseSeq produces regulatory regions (Rivera & Ren, 2013). In a standard data file, each region is represented by a chromosome number, start position, and stop position. A research study may produce hundreds of these data files, called region sets, each containing thousands of regions of interest. With these files, a researcher may want to look at previous experiments that produced a similar region set or examine which of the regions in the file are significant. In order to find similar region sets or regions, a similarity score is needed to compare different files.

So far, there has been no widely accepted method that produces a similarity score. One method, developed by Sheffield and Bock (2018), provides a ranking of similar files rather than a definitive score. Their method uses a Fisher's exact test on three different kinds of region sets: a user-specified query and universe, and a database. The query regions are compared to the universe and database for matches. Other methods to compute scores are even more basic. An algorithm called AIList simply totals up the number of overlaps between regions from the query and universe and divides by the total number of regions in the query (Feng, Ratan, & Sheffield, 2019).

Simply ranking similar files does not give enough information for a researcher to be confident that the files are actually a good match. The top result could be 40% similar or 99% similar, but the researcher would have no way of knowing the exact percentage. There are issues with the AIList algorithm too. Regions with a small overlap of only a few positions would be over-signified, and regions with a large overlap would be under-signified in the similarity score.

Developing an accurate similarity score would lead to faster research progress in the field of epigenomics. Researchers could easily upload their data to a website and obtain accurate matches in seconds. Finding similar results on the same cell type would provide verification of experimental integrity. Moreover, finding similar results across different cell types might open up even more interesting research questions, such as examining whether regions of the DNA sequence in heart cells are also expressed in lung cells. Without a similarity score, epigenomic data relationships would be harder to uncover.

My research proposes a new method to compute similarity scores between region sets involving ideas taken from the natural language processing field. The first step is to create a standardized “vocabulary” of regions called a segmentation, which would be a collection of regions that most accurately encompasses all possible regions. Then, other region sets could be translated into regions defined in the segmentation and leverage existing techniques like word2vec to convert regions into vectors. These vectors could then be compared using linear algebra techniques like cosine similarity to find out which regions are most similar or occur together frequently. Furthermore, the region vectors could be used to construct a vector representing the entire region set. Vectors are advantageous because computations can be performed on them very quickly.

The project will be implemented in the Python programming language. To test the accuracy of the similarity score, a perturbation tool must be developed because there are no previous similarity scores or test datasets to compare to. The perturbation tool will perform a number of alterations on existing region sets and produce an expected similarity score between the altered region set and the original, which can be compared to the actual similarity score produced by our algorithm. The similarity score can also be applied to known relationships between regions to see if it can reproduce the findings.

STS Problem

“Find out what your DNA has to say about you and your family” (23andMe, 2019). Today, DNA testing has become a common service through companies like 23andMe and

Ancestry. The process is simple - a customer orders a kit for spit sample collection, then mails it to the company. Their DNA is sequenced and uploaded to the company's database, enabling comparisons against millions of other users to find relatives, check for disease-causing genes, and trace ancestry to a country of origin (Regalado, 2019). For many people, genetic testing has helped them discover something new. There is a whole page on the website of 23andMe dedicated to customer stories, ranging from people who were alerted of an increased risk of celiac disease to those who reconnected with an extended branch of their family (23andMe, 2019).

However, some genetic testing companies are encouraging customers to send in their DNA for a different reason. "Give us your DNA. Help catch a criminal," is the message from FamilyTreeDNA, which is one of the companies now allowing police to scan their databases for DNA matches with crime suspects. This new use for DNA databases has become increasingly common due to a high profile case where police used the website GEDMatch to arrest the Golden State Killer (Zhang, 2019). In this case, which had been unsolved for 40 years, police uploaded DNA from the crime scene to the GEDMatch database and found 10 to 20 distant relatives of the killer. They then traced the lineage back to a common ancestor - great-great-great grandparents from the 1800s - and began narrowing down suspects to present day offspring using genealogy techniques (Jouvenal, 2018).

Because of the widespread coverage of the Golden State Killer case, police use of the DNA databases is only going to increase. If companies do not recognize that there is a new user group, then they may not be prepared for when police use their DNA databases. The technology may not be completely suited for the police's needs, even though they now comprise part of the user base. In addition, everyday consumers will not have an accurate understanding of who is participating with them in these DNA testing services.

I argue that the use of DNA testing companies' services has expanded beyond the technology's original purpose due to groups of criminal investigators using DNA databases to solve crimes and that DNA testing companies are not prepared for the ramifications of the new use case. I will use the "Users as agents of technological change" framework to analyze the issue (Kline & Pinch, 1996), which focuses on how new user groups begin to use a technology in a way that differs from the creators' intended use, resulting in unforeseen consequences and the shaping of the technology. By using this framework, I will explain that DNA testing was a relatively stable and straightforward technology for consumers, but is now being interpreted and used in a different way by police.

Conclusion

The amount of epigenomic data continues to grow, and in order to utilize it properly, more tools need to be developed. My research will create a similarity score for comparing region sets that will allow scientists to discover new relationships between cell types and specific regions. I will also consider the potential social implications of this technical research by studying DNA testing services and new user groups, such as the police, using the framework of users as agents. Research in both the technical and social projects will allow faster developments in epigenomics and predict how new groups will use this kind of epigenomic comparison technology.

Word count: 1713

References

23andMe. (n.d.). DNA genetic testing & analysis. Retrieved from https://www.23andme.com/.

ENCODE Project Overview. (n.d.). Retrieved October 30, 2019, from https://www.encodeproject.org/help/project-overview/.

FamilyTreeDNA. (n.d.). DNA testing for ancestry and genealogy. Retrieved from https://www.familytreedna.com/.

Feng, J., Ratan, A., & Sheffield, N. C. (2019). Augmented interval list: A novel data structure for efficient genomic interval search. *Bioinformatics*. doi: 10.1093/bioinformatics/btz407

Jouvenal, J. (2018, May 2). To find alleged Golden State Killer, investigators first found his great-great-great-grandparents. Retrieved from https://www.washingtonpost.com/local/public-safety/to-find-alleged-golden-state-killer-investigators-first-found-his-great-great-great-grandparents/2018/04/30/3c865fe7-dfcc-4a0e-b6b2-0bec548d501f_story.html.

Kline, R., & Pinch, T. J. (1996). Users as agents of technological change: The social construction of the automobile in the rural United States. *Technology and Culture, 37*(4), 763-795.

Regalado, A. (2019, February 18). More than 26 million people have taken an at-home ancestry test. Retrieved from https://www.technologyreview.com/s/612880/more-than-26-million-people-have-taken-an-at-home-ancestry-test/.

Rivera, C. M., & Ren, B. (2013). Mapping human epigenomes. *Cell*, *155*(1), 39–55. doi: 10.1016/j.cell.2013.09.011

Sheffield, N. C., & Bock, C. (2015). LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*, *32*(4), 587–589. doi: 10.1093/bioinformatics/btv612

Zhang, S. (2019, April 2). A DNA company wants you to help catch criminals. Retrieved from https://www.theatlantic.com/science/archive/2019/03/a-dna-company-wants-your-dna-to-catch-criminals/586120/.

Zhang, S. (2019, October 1). The messy consequences of the Golden State Killer case. Retrieved from https://www.theatlantic.com/science/archive/2019/10/genetic-genealogy-dna-database-criminal-investigations/599005/.