Machine Learning in Genomics

Luke Irwin

BIOI1000-099: Final Paper

I. Introduction

Machine learning is an up-and-coming tool in science and is conducive to advancement in multiple fields of study, but it is especially important in bioinformatics. Machine learning is the study of computer algorithms that have the capability to learn from past data and can improve themselves. The idea of machine learning is that a computer attempts to simulate human learning capabilities and can improve and learn through experience [8]. This technology has so much potential and has been improving many fields of study, from medical to military. There is significant use for machine learning in fields of life sciences, especially biology and bioinformatics. Machine learning has been a part of bioinformatics for years, most notably with clustering, classification, and data analysis [4]. One of the main benefits of machine learning is how useful it is in early health applications and the prevention of diseases [1]. Genomics is a subfield of bioinformatics that focuses on the structural and functional aspects of genomes for many different species and organisms [1]. Machine learning assists scientists in going through the genomes of organisms and species and can help identify specific genes that can cause specific characteristics. It can be used to identify positive characteristics so that scientists can study what causes them or it can identify negative characteristics that may be harmful to an organism or species. This could include disease or identifying mutations. This can help scientists get a better understanding of the natural world. The real-world potential of machine learning in bioinformatics and genomics specifically is that doctors and scientists are enabled to find problematic genes and sequences that might be detrimental to humans or other organisms and, with the technology available, they could be able to cut out those genes from the organism. This could ultimately result in the eradication of certain genetic diseases. Another use for machine learning in genomics is genetic engineering. Genetic engineering is the process by which organisms' genomes are modified through changing and manipulating the organisms' genetic material. While it is currently deemed unethical for this process to be done on humans, this process can be seen in the genetic modification of plants to yield more food, or to become less susceptible to disease or parasites. The focus of my paper is the DNA sequencing aspect of genomics and genetic engineering and how machine learning plays a role in it. DNA sequencing is the process of understanding and determining the nucleotide sequences. The code that I will write will essentially create a digital model strand of DNA that is randomly filled with characters to represent base pairs. It will then create a complementary strand matching the base pairs. The program will create a strand of random DNA with as many base pairs as the user wants to input, up to the Java integer maximum value (2147483647). The user can then input any base pair sequence and the program will loop through the entire list of base pairs and count how many times the inputted sequence occurs within the represented strand of DNA. The main goal of this program is to give an example of how NCBI's random sequence

generator algorithm could work and give a visualization of what one might see when sequencing someone's genome with the output. This code is intended to be a learning tool used to visualize the process of how NCBI might come up with their random DNA sequences. The machine learning aspect of the project is the further applications after the sequence is acquired.

II. Background

Machine learning is still a relatively new tool in many sciences and is an important part of artificial intelligence [8]. Machine learning stems from computer science and using computers programs to replicate human learning and behavior. The main thing that makes machine learning different from just using algorithms is the idea of self-improvement through experience. The idea is that as the program runs, it gains experience and can spontaneously gain skills and knowledge through completing the tasks of the program [8]. There are two main kinds of machine learning: supervised and unsupervised learning. Supervised learning algorithms need external input from a user and can learn based on the data provided to the algorithm [8]. An example of supervised learning is a decision tree. A decision tree is a decision-making model that has nodes and branches that lead to different trees as depicted in **Figure 1** [12]. Unsupervised learning is a machine learning algorithm that does not usually learn from new data. It will apply past information to new data [8]. There are also neural networks that are useful tools with machine learning and they are neither supervised nor unsupervised [8]. Neural networks are like decision trees, but the decision nodes are all connected, more like a web [8]. The first-ever completely sequenced genome was that of a virus called the *phiX174 virus* [11]. This was accomplished by Frederick Sanger in 1977 [11]. Sanger's method of gene sequencing, now known as Sanger Sequencing, involves obtaining a specific piece of DNA and putting it through a chain-termination PCR reaction. Then it is run through gel electrophoresis, and then the gel is analyzed and the base pair length can be determined. The development of this technology has significantly contributed to the growth of genomics and helped pioneer the field. Genomics and specifically DNA sequencing is a central part of bioinformatics and involves the sequencing of an organism's genome. DNA sequencing involves identifying the specific nucleotide pattern and sequences that make up the genome of the organism. This is important because it allows scientists to be able to identify the specific DNA sequences code for specific proteins and identify the function of said protein. There is also the chance that there are some problematic genes in an organism that can cause fatal issues. In humans, an

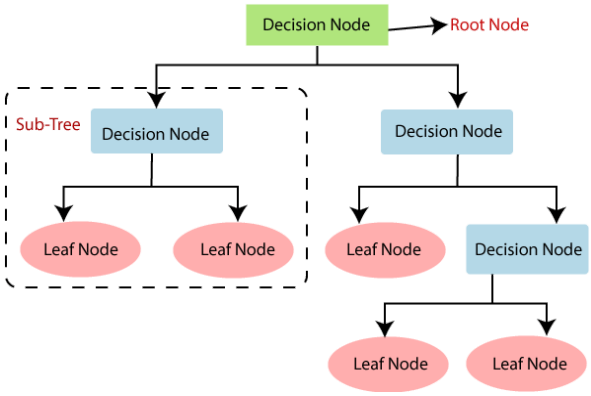

Figure 1. *Flow tree example.*

example of those genes could be cancer-causing genes and the genes that cause Cystic Fibrosis. A gene-editing tool, called CRISPR-Cas9, can be used to help treat those genetic disorders. CRISPR-Cas9 technology essentially gives scientists and doctors the ability to find a genetic sequence and make a specific guide RNA for the sequence that is problematic, and then they use the Cas9 protein to modify the target region by cutting the sequence out entirely, inserting a new sequence, or modifying the existing sequence. The CRISPR-Cas9 technology has so much potential for medical and biological advancements because of its ability to "quickly change the DNA of nearly any organism - including humans" [2]. Because of the utility that CRISPR-Cas9 technology has, scientists believe that this technology will be able to help treat and potentially cure genetic disorders. This technology is already being used for treatment of Cystic Fibrosis and multiple types of cancer. This technology is not exclusive to the medical field, though. CRISPR-Cas9 technology can also be used for farming and genetically engineering better crops and livestock. This technology grants scientists the ability to genetically modify crops. For example, scientists have been able to make wheat and rice disease-resistant [2]. Another application for the technology is a process called xenotransplantation in which an organ from a nonhuman organism is transplanted into a human. CRISPR-Cas9 technology is used in this instance to modify the organism's organs to be able to function more like a human organ, and recently, this procedure was accomplished successfully with a pig's kidney. This technology has fundamentally revolutionized gene editing because of how quick and cost-effective it is. While CRISPR-Cas9 is effective at cutting and replacing genes or sequences, the first step is identifying the gene that is problematic. DNA sequencing is how scientists are able to find and identify genes that are causing problems. When paired with computers and machine learning, the process of going through an organism's genome can be efficient and effective. Computers can sequence a genome and find a specific sequence significantly faster and more precisely than any human can, and that is one of the reasons machine learning and computers are so important in bioinformatics and genomics. Genomics focuses on an organism's entire genome; what my program is focused on specifically is DNA. DNA consists of a phosphate backbone, a pentose sugar, and a nitrogenous base. There are four nitrogenous bases that can make up DNA: adenine, cytosine, guanine, and thymine, which is exclusive to DNA. DNA sequencing is a part of genomics that focuses on the specific nucleotide sequence of DNA or RNA [5]. It is important because it gives scientists the exact genetic makeup of an organism and can reveal a lot about the organism and why it functions the way it does by showing the sequences that code for the proteins that determine specific characteristics. Neural networks are machine learning tools that are used in determining the structure of proteins, like with *Alfafold.* It can also be used when comparing gene sequences or genomes to identify mutations in genes. This can also be important in determining conditions in organisms caused by genetic mutations and can help scientists identify the causes of the expression of genotype or phenotypes. In this instance, supervised learning can be applied in the form of a decision tree to help detect and find mutations in a gene sequence or genome [8].

DNA sequencing in bioinformatics can be used specifically in molecular and evolutionary biology [5]. In molecular biology, DNA sequencing is important because it can reveal the reasons for the production of specific proteins and how they are made and thus can reveal a function of something in the body and can explain the expression of specific phenotypes [5]. It is used in evolutionary biology when comparing the genomes of organisms to see how organisms have evolved from their ancestors and developed certain traits [5]. Machine learning as a whole is involved with many different fields of science because of its capabilities. While it originally stems from computer science, it is used in many fields of science and research, and in interdisciplinary sciences, like bioinformatics.

III. Code and Documentation

https://github.com/lukeirwin03/BioI-Genome-Sequence

IV. Discussion

The main thing that I found is how beneficial machine learning is to the world of bioinformatics and specifically, genomics. Having the technology to simulate human learning alleviates scientists of many mundane and repetitive tasks. It can also complete tasks that would ordinarily take a lot of time and precision significantly faster and more precise. One of these tasks is sequencing DNA. DNA sequencing is a part of many things and can tell scientists a lot about specific organisms. Looking at the DNA of an organism can tell scientists things about potential genetic disorders, explain why a specific phenotype is expressed over another, and much more. In the case of genetic disorders, it can be paired with CRISPR-Cas9 technology to help treat the disorder and hopefully cure it eventually. This is the case for disorders like Cystic Fibrosis and several types of cancer. And with this technology evolving every day with the influx of data and findings that happen every day, the technology keeps getting better and better. I think that if this kind of technology keeps advancing at the rate it's going, scientists and doctors will be able to treat more genetic disorders and potentially cure them. This is a huge thing in the world of science because of all of the benefits that it has. Another use for it is with genetic engineering of crops and making them better which has the potential to feed more people and contribute to the efforts against world hunger. And, more recently, the scientific community has discovered that with machine learning and CRISPR-Cas9 technology, genetic engineering for xenotransplantation is possible. This is such a monumental advancement because we now know that it is possible and opens up many options for the future of genetic engineering. Machine learning has so much potential for good because of the area for growth in the technology. Overall there are so many benefits to using this technology. Machine learning alone will not accomplish this, but it will aid in the process by helping pinpoint genes from a DNA sequence that can be manipulated for good. Being able to manipulate genes for good can contribute to the fight against world hunger, in the genetic engineering of crops, and also contribute towards finding a cure/aiding in treatment for specific cancers. There is so much

potential for good and as this machine learning technology improves, the scientific community as a whole gets closer to these goals. That is why machine learning is so important across a lot of sciences and specifically biology and genomics. There is so much area for growth as machine learning is still a relatively new technology and with improvements to this technology, scientists and doctors could be able to use this technology with DNA sequencing and CRISPR-Cas9 technology to treat and potentially cure so many genetic diseases. There is so much potential behind this technology, but because it is still relatively new, there are some limitations. The main limitations are the inability to completely replicate human learning. There have been many efforts to replicate human thinking, but at the moment, it isn't possible to completely replicate human thinking.

V. References

[1] B. Bagiroz, E. Doruk, and O. Yildiz, "Machine Learning in Bioinformatics: Gene expression and microarray studies," *2020 Medical Technologies Congress (TIPTEKNO)*, 2020.

[2] H. Ledford, "CRISPR, the Disruptor," *Nature*, vol. 522, no. 7554, pp. 20–24, Jun. 2015.

[3] R. Barrangou and J. A. Doudna, "Applications of CRISPR technologies in research and beyond," *Nature Biotechnology*, vol. 34, no. 9, pp. 933–941, 2016.

[4] Y.-T. Tsai, "An Overview of Machine Learning and HPC in Open Sources for Bioinformatics*," *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018.

[5] P. Dixit and G. I. Prajapati, "Machine Learning in Bioinformatics: A novel approach for DNA sequencing," *2015 Fifth International Conference on Advanced Computing & Communication Technologies*, 2015.

[6] B. Yimwadsana and P. Artiwet, "On optimizing DNA sequence design for DNA logic and circuit," *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2018.

[7] M. Zhang, C. L. Sabharwal, W. Tao, T.-J. Tarn, N. Xi, and G. Li, "Interactive DNA Sequence and Structure Design for DNA Nanoapplications," *IEEE Transactions on Nanobioscience*, vol. 3, no. 4, pp. 286–292, 2004.

[8] Y. Cai, Q. Dong, and A. Li, "Application and research progress of machine learning in bioinformatics," *2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, 2020.

[9] J. Cai, L. Wei, K. Zeng, and G. Xiao, "Identification and prediction of key nucleotide sites using machine learning in Bioinformatics: A brief overview," *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, 2019.

[10] S. Shakhari, P. Ghosal, and M. Sarkar, "A provably good method to generate good DNA sequences," *2016 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS)*, 2016.

[11] Smith, Yolanda. "History of Genomics." *News Medical*, 26 Feb. 2019, https://www.news-medical.net/life-sciences/History-of-Genomics.aspx.

[12] "Machine learning decision tree classification algorithm - javatpoint," *www.javatpoint.com*. [Online]. Available: https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.