# DNA ANALYSIS AND ILLNESS INFERENCE

Enes Civan

Computer Engineering, Faculty of Engineering and Nature Sciences, Kütahya Health Sciences University, Kütahya, Türkiye

# ABSTRACT

This project aims to create a program to help researchers and medical personal to read whole genome files and work on them with various categorization and filtering options provided to the user. The project is going to save so much time and computing power by easily clipping the data to what the user wants to actually work on so individual patients can hold only the necessary data and no more on the database and make comparisons to get similarity rates of the needed part. After this phase 1 ends, we are planning to work on machine learning to integrate a genome test by training the program with healthy and corrupt genes thus it can automatically have an idea by rating the sickness possibility.

## INTRODUCTION

Genome researches have been trending up with so much potential to give humanity cure diseases or make medicine to fight individual bacteria or virus by interpreting its genome and make an antidote to target them. This kind of potential led to huge investments on research, letting the industry grow. According to [National Human Genome Research Institute](https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science), genomics projects will generate 40 exabytes of data in the next decade. This caught our attention and as we dug deeper trying to understand this field, we kept finding more things to learn. As working with genome requires computers, I and my team decided to make a project to help pushing the innovation forward or at least create an alternative to pre-existing software and learn more along the way so we could create new useful things. The team consists of me, Mehmet Serdar Koz, Mert Can Kabakçı, and Bilal Emre Taydaş with separate reports to present so we all can provide more details about our parts.

## MATERIALS AND METHODS

First stages our work was merely doing research. We wanted to find all the sources to help working on our project. Even though we took a look at many articles and platforms, it all seemed to be referencing one single source, [National Center for Biotechnology](https://www.ncbi.nlm.nih.gov/) Information from [National Library of Medicine](https://www.nlm.nih.gov/).

While trying to write a function to do genome comparisons and give a similarity rate output, I’ve noticed that the function from [BioPython](https://biopython.org/) is much more advanced than my ideas and research about my own diff algorithm as it aligns the genes and gives a much more accurate output. I ended up integrating the necessary functions to align and compare the genome parts of need.

## National Center for Biotechnology (NCBI)

This is a center within the National Library of Medicine which is located in National Institutes of Health, letting us get data about Genes, Genomes and provides much more. We downloaded a [human genome data](https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml) file to work on from this source and found so much useful information about it.

## BioPython

BioPython provides functions to help write code about biological work field. I’ve used Seq, PairwiseAligner and SeqIO functions from Bio library. Seq function creates an object from the chosen part of the genome that can be aligned with another Seq object. SeqIO basically does the same thing but it doesn’t require a string input, it pulls the data directly from a file. I wrote saveGene function to save the wanted part of the gen to a new file so it can directly be pulled from the file with SeqIO. Lastly PairwiseAligner creates an aligner object to get two Seq objects and aligns them. We get the similarity rate with the score method and visual alignments with the align method.

## THE FUNCTIONS

I made around a dozen of functions that work together to search, filter, clip, save, and compare the genome. They work with [FASTA](https://www.ncbi.nlm.nih.gov/genbank/fastaformat/) files with .fna extension. For simplicity’s sake, we limited our work on 21st chromosome as the fasta files can get so big so the development can be easier. As the work is still in progress, the file names, function names and even the outputs may change over time so this report won’t cover the detailed usage of each function.

## RESULTS AND CONCLUSION

By the time graphical interface, database and encrypting work is done, the program will be ready to use. Users are going to have a login and register screen; a panel that shows the patient list with tools to edit, add or remove patients. The list will provide individual patient profiles with genome info as well as user notes. User will be able to compare a location of genome with another location of another genome and get a similarity score. The user will also be able to see gene alignments of the comparison.

## REFERENCES

* Eric W Sayers, Richa Agarwala, Evan E Bolton, J Rodney Brister, Kathi Canese, Karen Clark, Ryan Connor, Nicolas Fiorini, Kathryn Funk, Timothy Hefferon, J Bradley Holmes, Sunghwan Kim, Avi Kimchi, Paul A Kitts, Stacy Lathrop, Zhiyong Lu, Thomas L Madden, Aron Marchler-Bauer, Lon Phan, Valerie A Schneider, Conrad L Schoch, Kim D Pruitt, and James Ostell -Nucleic Acids Res. 2019 Jan 8; 47(Database issue): D23–D28. Published online 2018 Nov 5. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323993/>
* About the National Library of Medicine https://www.nlm.nih.gov/about/index.html
* A Brief History of NLM <https://www.nlm.nih.gov/about/briefhistory.html>
* Genome assembly GRCh38.p14 <https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/>
* Human Genome Resources at NCBI <https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml>
* Chromosome Map <https://www.ncbi.nlm.nih.gov/books/NBK22266/>
* Genomic Data Science <https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science>
* BioPython <https://biopython.org/>
* FASTA Format for Nucleotide Sequences <https://www.ncbi.nlm.nih.gov/genbank/fastaformat/>