Logo

Description automatically generated

Automated Script Toolbox

BMD-301 Project

Spring 2022

**Submitted by:**

Nadeen Nassif 19105325

18102297 Zahraa Omar Elzeiry

19104941 Salma Ayman

19106199 Rawan Ramadan

19200088 Ahmed Hesham

**Under the Supervision of** Prof. Mohamed Elhadidi

Table of Contents

[Abstract: 2](#_Toc107906401)

[Keywords: 2](#_Toc107906402)

[Introduction: 3](#_Toc107906403)

[Methodology: 4](#_Toc107906404)

[Data Retrieval: 4](#_Toc107906405)

[FASTAQ to FASTA: 5](#_Toc107906406)

[Pairwise Alignment (Local & Global) 6](#_Toc107906407)

[Blast: 6](#_Toc107906408)

[MSA: 7](#_Toc107906409)

[Results 8](#_Toc107906410)

[Converting from FASTQ to FASTA 8](#_Toc107906411)

[Pair-wise Alignment (Local &Global) 9](#_Toc107906412)

[Multiple Sequence Alignment (MSA) 10](#_Toc107906413)

[4. Validation 10](#_Toc107906414)

[Phylogenetic Tree 12](#_Toc107906415)

[The phylogenetic Tree 13](#_Toc107906416)

[Blast 14](#_Toc107906417)

[Finally: 17](#_Toc107906418)

[Conclusion 18](#_Toc107906419)

[References 19](#_Toc107906420)

# **Abstract:**

The need of storing, analyzing, and meaningfully interpreting biological data is growing as it is produced at an incredible rate. Bioinformatics is now the method of choice in forensic sciences because of the expanding network of biological information databases, including those for the human genome, transcriptomics, and proteomics. Pipeline frameworks are being used in high-throughput bioinformatics analysis to handle sequence and information. Several innovative pipeline frameworks, including visualization, version tracking, and summary reporting, have been created in recent years to address bioinformatics and reproducible research. The most fundamental kind of pipeline framework can be thought of as scripts written in the Unix shell or another scripting language. Flexible pipelines can be created using variables and conditional logic using scripting. A rare genetic congenital eye condition called primary congenital glaucoma affects newborn children. Most primary congenital glaucoma patients have random occurrences and no recognized family history. By applying methods of bioinformatics to the gene that influences primary congenital glaucoma we found that there are 5 genes contributing to this disease. By doing homology to get the variant, we get from the genome of the n the wild type gene toward the mutated gene to get the gaps, insertion, and deletion. Then try to get the mutated genes by this homology. In this, the paper we will apply methods of the automata to these steps and to know the mutated genes.

### **Keywords:**

Bioinformatics, Automata, Phylogenic tree, Multiple sequence alignment, CYP1B.

# **Introduction:**

We present an automaton, a system that integrates data administration, standardized viral assembly, and a graphical user interface without any gaps. By using the validation process, we can determine whether a file is in the fastQ format by looking at the headers, and the automata will report on the file's validity or invalidity if the sequence contains numerous gaps or not. We can access the information that already exists and compare it to other data to see where there are parallels and discrepancies through an automated script to retrieve the data. To apply these methods, we enter the NCBI tool and searched for the gene we have CYP1B on homo sapiens to create a customized database. Besides, we choose the 5 samples According to (Ali M et al., 2009) “Mutations in Cytochrome P4501B1 (CYP1B1 [MIM \*601771]) at the GLC3A locus account for up to 50% of cases11,12 and more than 100 CYP1B1 mutations have been reported in the Human Gene Mutation Database. No responsible gene has yet been identified at the GLC3B and GLC3C loci” We use for the technique of MSA complex sophisticated to do homology for the sequence of gene CYP1B with all other species to see the similarity, and from the output detect the region variant between species to see if this protein like another species or not. After this get the phylogenetic tree by knowing the nearest branches. A set of visual conditions known as glaucoma can harm the optic nerve and lead to permanent blindness. A rare kind of glaucoma called primary congenital glaucoma (PCG) is characterized by ocular abnormalities that impair the outflow of the aqueous fluid. Children under the age of 3 are the target for PCG. Primary congenital glaucoma (PCG) damage the optic nerve 2,3 afterward.

# Methodology:

This automated script toolbox carries out several important tasks that are the most used in the bioinformatics field, those tasks are:

1. Perform local and global pairwise
2. Perform multiple sequence alignment (MSA)
3. Generating a phylogenetic tree
4. Transforming files from FASTQ to FASTA
5. Validate the FASTq
6. alignment, and phylogenetic tree (user can choose his preferred algorithm)
7. Perform basic statistical operations on a file by calculating how many A, C, G, T, CG content, Introns, exons, etc]

### **Data Retrieval:**

1. Customized Database:
2. Open nucleotide NCBI
3. Searching for the selected gene (CYP1B1)
4. The results were (1396) gene
5. Selecting the Homo-Sapiens genes only to filter the result to be just (330) gene
6. Finally, we downloaded it from (sent to) selecting (FASTA format file)
7. As a result, we have a database fasta file formed of the CYP1B1 genes in all the Homo-Sapiens.
8. Five Samples of Query:
9. From Ensemble, we searched for the (CYP1B1) gene in humans
10. Then downloaded it in Fasta format. Now, we have the first sample of the query.
11. To select the rest of the samples, we searched for the orthologous genes. There were 204 orthologous genes.
12. Focusing on the placental mammals, we choose the Arabian camel, beluga whale, and American mink.
13. Then we choose chimpanzees from the primates.

### **FASTAQ to FASTA:**

1. The user enters the name of the file that needed to be converted
2. Then, enter the name of the output file
3. Applying the command ['sed -n \'1~4s/^@/>/p;2~4p\' {}.fastq > {}.fasta'.format(inFile,outFile)] to the first two lines (Header & Sequence) and replaces every (@) at the beginning of the sequence with the greater than sign (>), so it could be a FASTA formatted file.
4. Finally, the fasta formatted file will be executed successfully.

### **Pairwise Alignment (Local & Global)**

The automation script needs those libraries to be imported to apply the pairwise alignment:

* from Bio.pairwise2 import format\_alignment
* from Bio import pairwise2
* from Bio.Seq import Seq

1. First, it asks the user whether he wants the alignment to be local, or global.
2. If local alignment, the user should enter the first sequence and then the second sequence.
3. Using this command [alignment=pairwise2.align.localxx(seq1, seq2)]
4. If the user chooses global, the command will be [alignment=pairwise2.align.localxx(seq1, seq2)]
5. Finally, using a for loop to loop on the whole sequence, using

for alignment in alignments:

print(format\_alignment(\*alignment))

### **Blast:**

1. First, the program needs to install the blast tool if it was not installed before, by using this command [sudo apt-get install ncbi-blast+], so it asks the user if he needs to install it or not.
2. After that, to install the database, the program needs to retrieve it from NCBI. Using the command [wget https://....] so the user could drop the link of the selected database.
3. After that, the program needs the query that will be aligned against the database, so we also install it using the command [wget https://....]
4. The last two steps result in zipped files. To unzip those files, we use the command [gunzip \*gz] to unzip all files.
5. Then use the command [makeblastdb -in All\_Homo\_CYP1B1.fasta -dbtype nucl] to index the database for easier alignment and because it is a huge file.
6. Finally, run the blast command [ blastn -query Human\_CYP1B1.fna -db All\_Homo\_CYP1B1.fasta -out Blast\_Results.txt]

### **MSA**

1. First, the user should enter how many elements he wants to make a multiple sequence alignment to.
2. Then, the list will be opened (lst = []), so the user could enter the accession number of the selected genes. Using for loop to append the files into the list
3. Using these accession numbers, we will apply ["efetch -db nucleotide -format fasta -id {} > all.fasta"] to retrieve the data from NCBI and concatenate it in one file into the list by applying [','.join(lst)]
4. After that, the user will enter the output format needed
5. Finally, using the command ["muscle -in all.fasta -out alignment.{}"], Multiple sequence alignment will be applied by the MUSCLE tool. Then, the output file will be ready to read or used in other applications.

# **Results**

### **Converting from FASTQ to FASTA**

Firstly, an automated project allows the user to enter the name of the FastaQ file then convert it to the Fasta file and rename it.

A screenshot of a computer

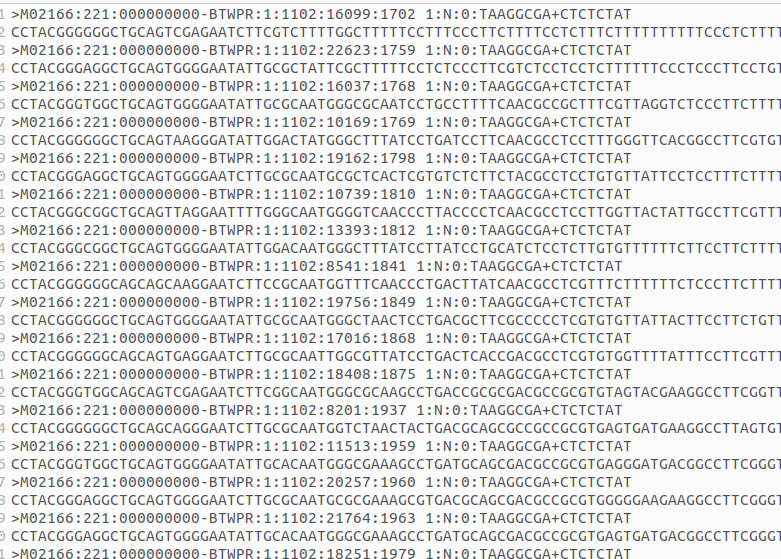
Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated



**The FastaQ file**



**Fasta file**

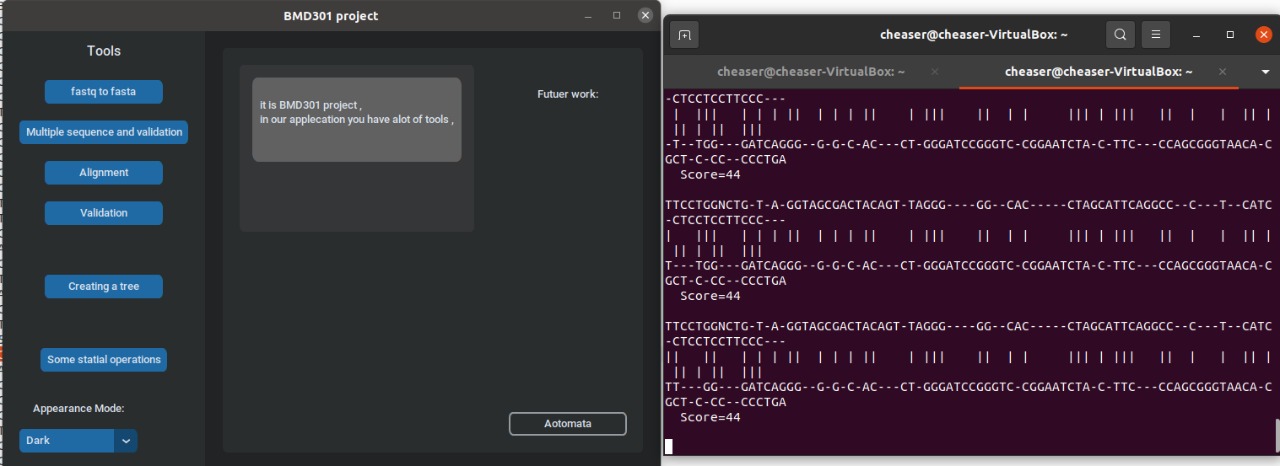
### **Pair-wise Alignment (Local &Global)**

Secondly, Local alignment can align any 2 sequences accurately and can identify more variants

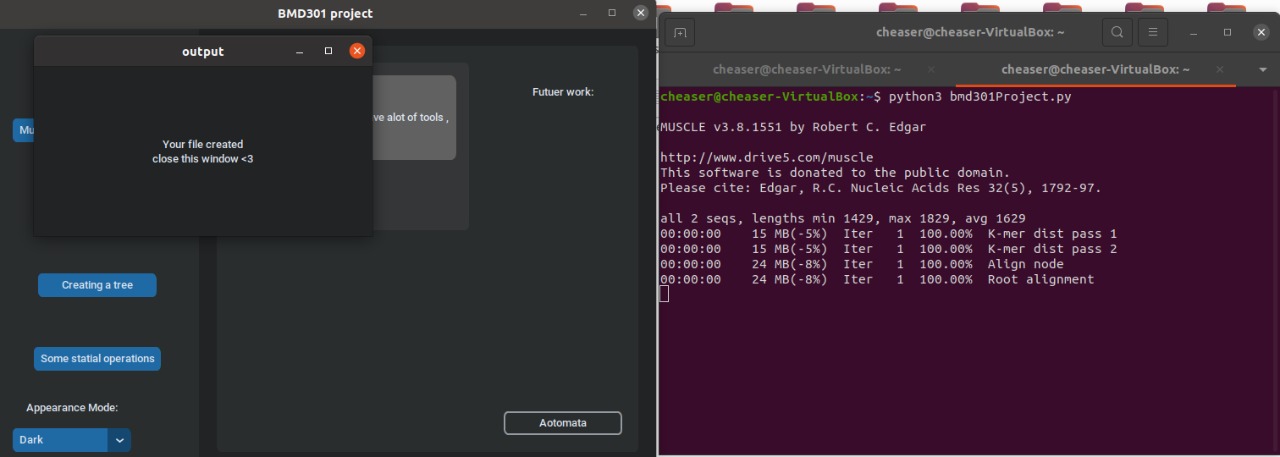
Text

Description automatically generated

Global alignment is a pairwise alignment that can align every nucleotide in every sequence. It is used when the sequences in the query are similar and equal size.



### **Multiple Sequence Alignment (MSA)**

Thirdly, MSA and validation: In the terminal, we can see the process of forming the MSA file

### 4. Validation

**Validate of FASTQ file**

The following details should be included in the result report:

1. Number of readings in the Sequence FASTq file.
2. The minimum, maximum, and typical read length.
3. A list of errors that include the read of sequence included in

* Header must begin with a '@'.
* Sequence
* ‘+’ sign
* Quality Score

Mention the error's line number, for example:

Graphical user interface, text, application

Description automatically generated

**Validate FASTQC**

We enable users to run a variety of quality control tests on raw sequence data produced in FASTQ format by high throughput sequencing platforms like Illumina According to the output result user will be able to identify the quality of sequence data.

Graphical user interface, text

Description automatically generated

### **Phylogenetic Tree**

The phylogenetic tree is used to know which species are having the same traits

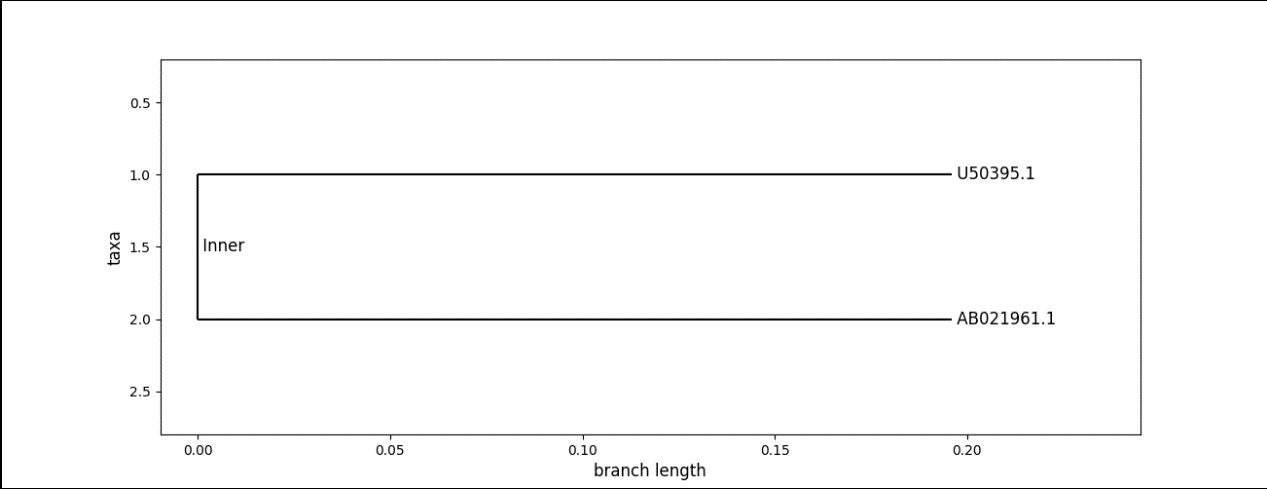
phylogenetic tree: the user enters the file name which the extension is .clw

Graphical user interface, application

Description automatically generated

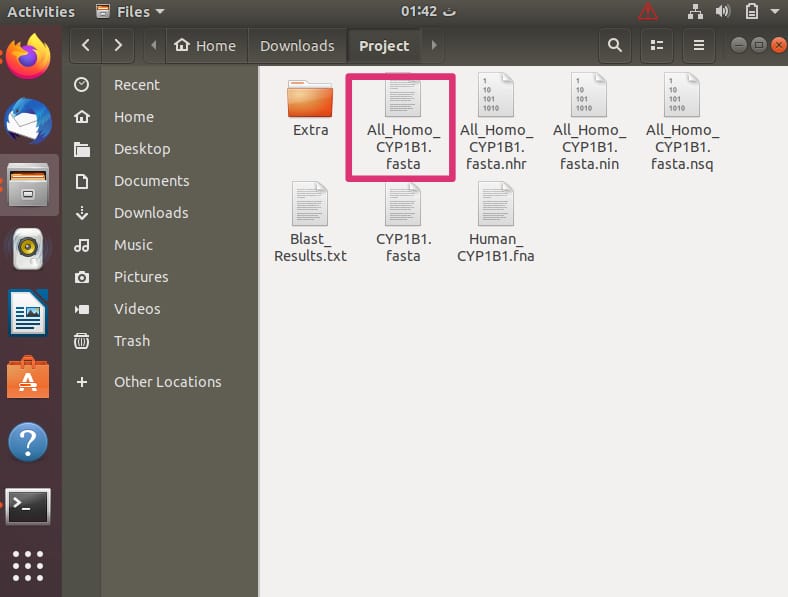
The output files are

Graphical user interface, application, Teams

Description automatically generated

### **Blast**

1. The database of the gene (CYP1B1) in homosapiens = 330 gene

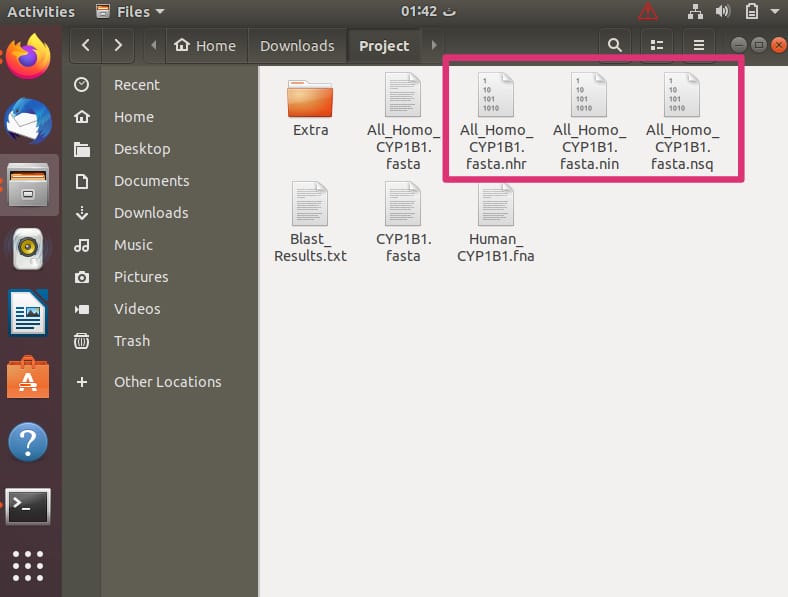


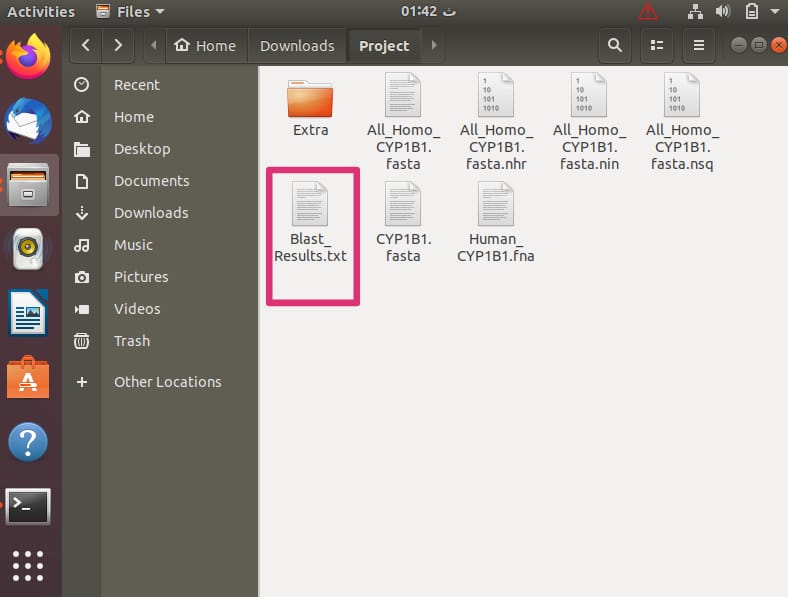
1. The query = CYP1B1 gene in human

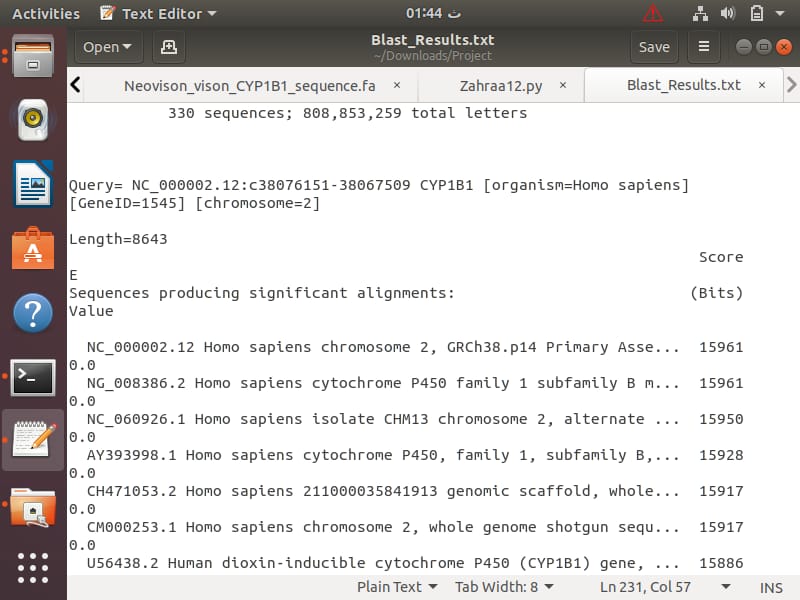
**Graphical user interface, application

Description automatically generated**

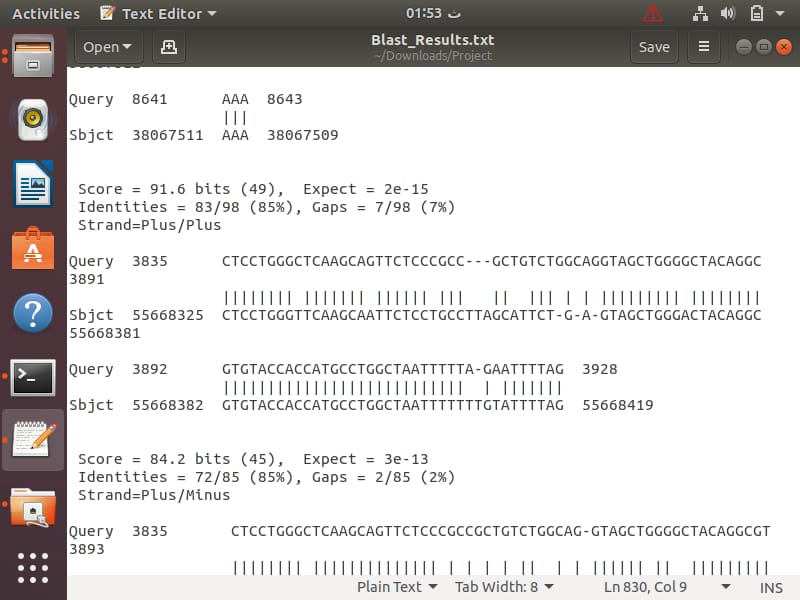
1. Those 3 files are the results of indexing the database



4. The result file of the blast alignment



5



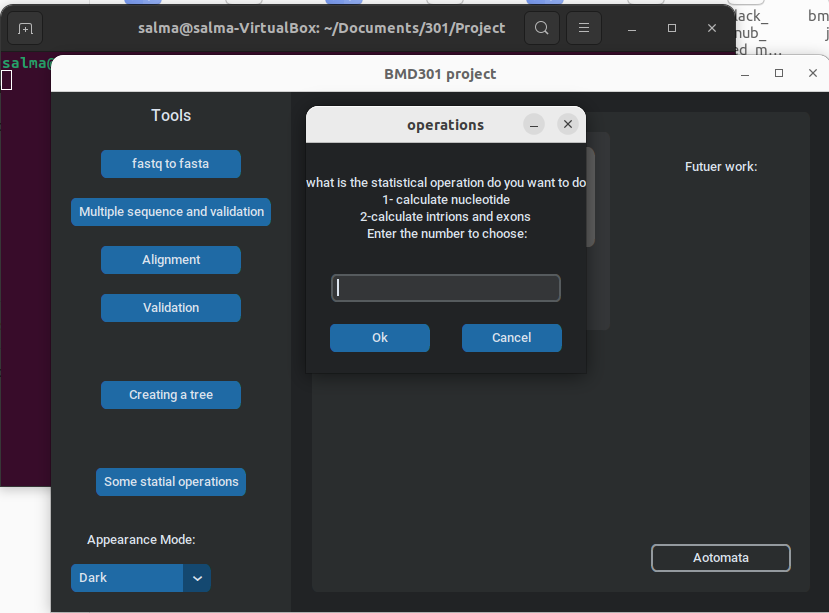
6

### **Finally:**

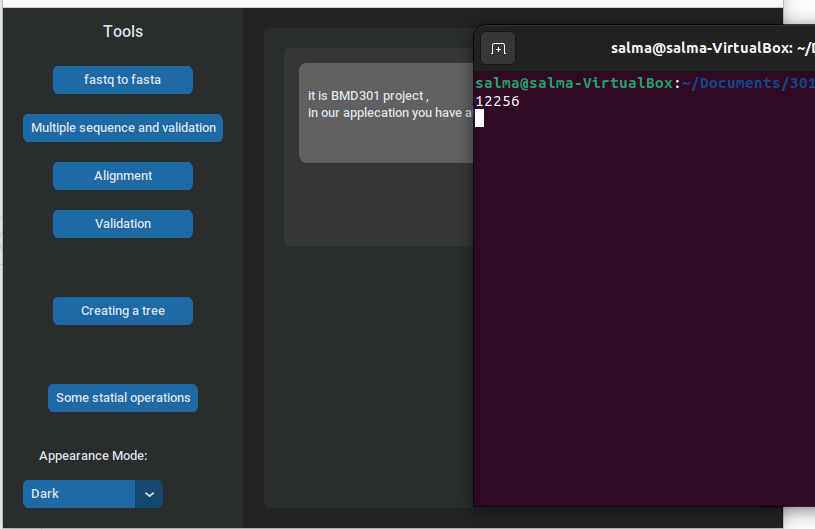
The user enters the nucleotide that they want to calculate and the file name of the output appears in the terminal. The user will know the number of exons and introns to know the coding area in the sequence

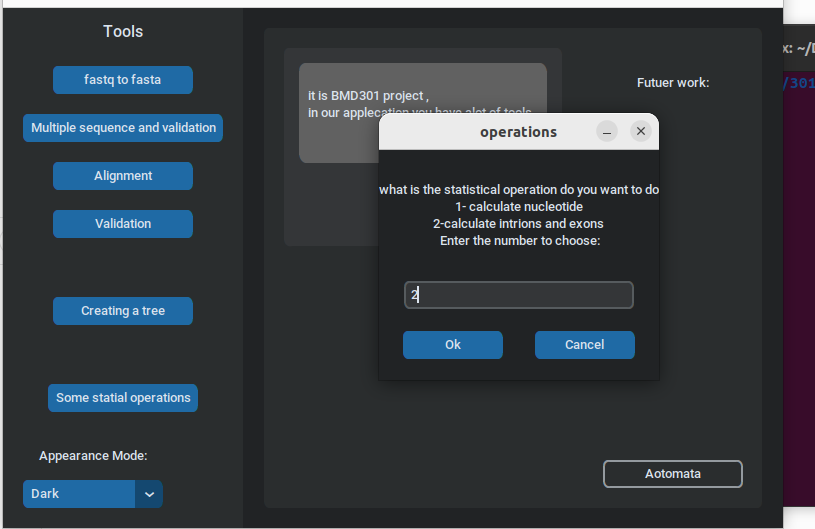
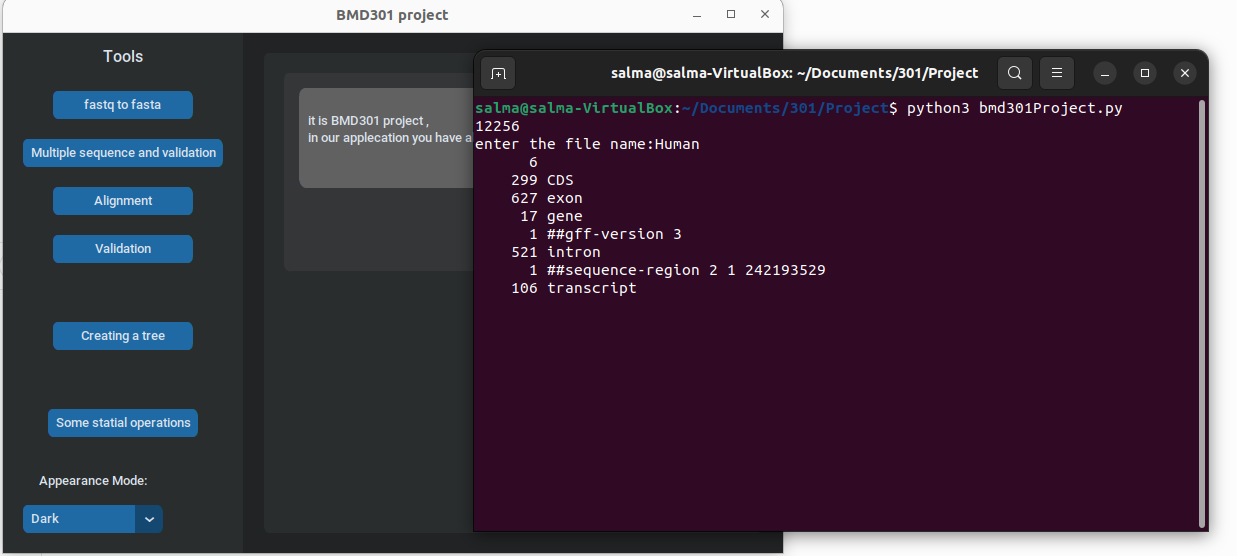
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated





# **Conclusion**

Mainly most tools are not user-friendly, and the results-driven focus of science means that many important non-functional software requirements are often overlooked. We established an automated and reproducible integrative Automated Script by python and bash script that will help the user focus more on revolutionizing and improving the result. by eliminating the requirement for any manual Task. Our software operates form a user-friendly graphic interface that allows software management and installation by users. Users of Bio2Byend have two options on the GUI. The first option is Manual, where the user can select the process, he or she wishes to execute on a file they have entered, such as pairwise, validation Fastq, or simple statistical operations.

The user of the second, automatic option need only provide the file name and choose the desired outcome. To sum up, we used the Bio2Byend tool to examine the gene that affects primary congenital glaucoma and discovered that the CYP1B1 gene mutation causes disease. We then selected 5 species (Arabian Camel, Human, Beluga Whale, Chimpanzee, American Mink). By performing alignment, we can determine the gaps, insertions, and deletions from the human genome's wild type gene toward the mutant gene.

# **References**