Automated Script

Bioinformatics BMD301 / Dr. Mohamed El Hadidi / 30.1.2023

Team:

Mayar Tarek	202002151
Malak Essam	202000935
Mohamed ElSayed	19105616
Nada Adel	202000554
Sarah Mahmoud	202000408
Shahd Fekry	202000908

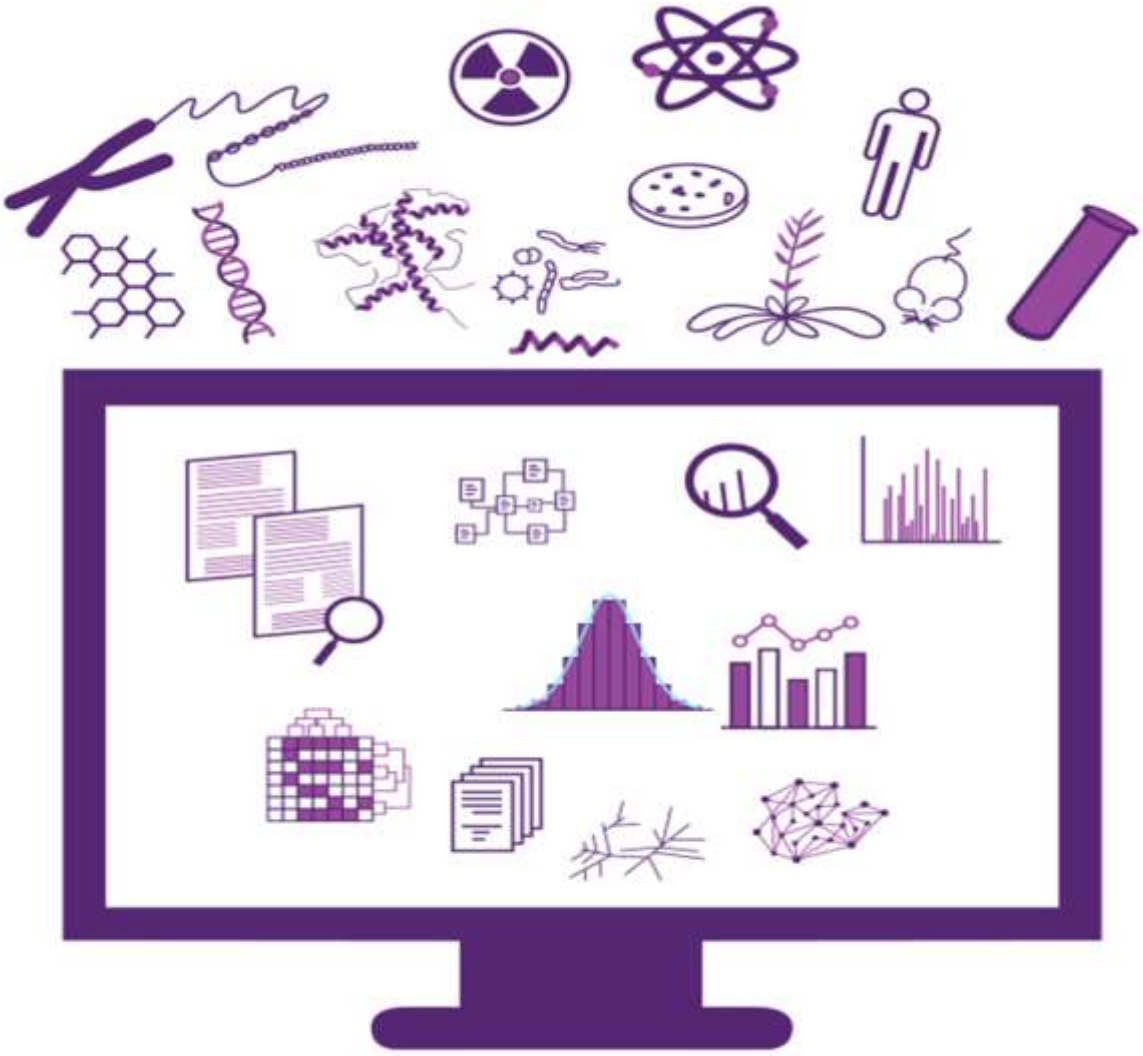


Table of Contents

ABSTRACT

Bioinformatics is the application of computer technology to the understanding and effective use of biological and biomedical data. It is the field concerned with storing, analyzing, and interpreting Big Data collected by life-science studies or acquired in a clinical setting. All bioinformaticians use Linux instead of Windows because bioinformatics options are limited on windows. When compared to other operating systems such as Windows and MacOS, Linux has a more robust command-line interface. We can perform both simple and complex tasks through its terminal. We can create files, remove files, move files, and do other basic operations. Even though the Linux terminal can run dozens of command lines, it is not automated, and you must do each step in a specific process independently. As a result, we created a fully automated command line application with a graphical user interface. By simply clicking the name of our application, the user can do a variety of operations automatically. For example, converting FASTQ to FASTA, counting K-Mers, pairwise or multiple sequence alignment, and so on. Lastly, we created a user-friendly graphical user interface to help the user perform tasks easily.

INTRODUCTION

Bioinformatics is essential for management of data in modern biology and medicine. The bioinformatics toolbox includes computer software programs such as BLAST and Ensemble, which depend on the availability of the internet. Analysis of genome sequence data, particularly the analysis of the human genome project, is one of the main achievements of bioinformatics to date. Prospects in bioinformatics include its future contribution to functional understanding of the human genome, leading to enhanced discovery of drug targets and individualized therapy. The main tools of a bioinformatician are computer software programs and the internet. A fundamental activity is sequence analysis of DNA and proteins using various programs and databases available on the world wide web. Anyone, from clinicians to molecular biologists, with access to the internet and relevant websites can now freely discover the composition of biological molecules such as nucleic acids and proteins by using basic bioinformatic tools. This does not imply that handling and analysis of raw genomic data can easily be carried out by all. Bioinformatics is an evolving discipline, and expert bioinformaticians now use complex software programs for retrieving, sorting out, analyzing, predicting, and storing DNA and protein sequence data. The growth of bioinformatics has been a global venture, creating computer networks that have allowed easy access to biological data and enabled the development of software programs for effortless analysis. Multiple international projects aimed at providing gene and protein databases are available freely to the whole scientific community via the internet. The escalating amount of data from the genome projects has necessitated computer databases that feature rapid assimilation, usable formats, and algorithm software programs for efficient management of biological data. These databases include both “public” repositories of gene data as well as those developed by private companies.

The easiest way to identify databases is by searching for bioinformatic tools and databases in any one of the commonly used search engines. One of the simplest and better-known search tools is called BLAST (basic local alignment search tool, at www.ncbi.nlm.nih.gov/BLAST /). This algorithm software can search databases for genes with similar nucleotide structure (fig (fig3)3) And allows comparison of an unknown DNA or amino acid sequence with hundreds or thousands of sequences from human or other organisms until a match is found. Databases of known sequences are thus used to identify similar sequences, which may be homologues of the query sequence.

METHODOLOGY

Data used

In the **pairwise alignment**, for the nucleotides alignment: we downloaded single sequence FASTA files from NCBI as a DNA query for organisms (Mus musculus, Cricetulus griseus, and Bos primigenius), as shown in Fig. 1, and we downloaded the reference genome sequence as a DNA database (Fig. 2).

As for the protein alignment, we used the cow organism as a protein query and the human as a reference database.

Figure 1

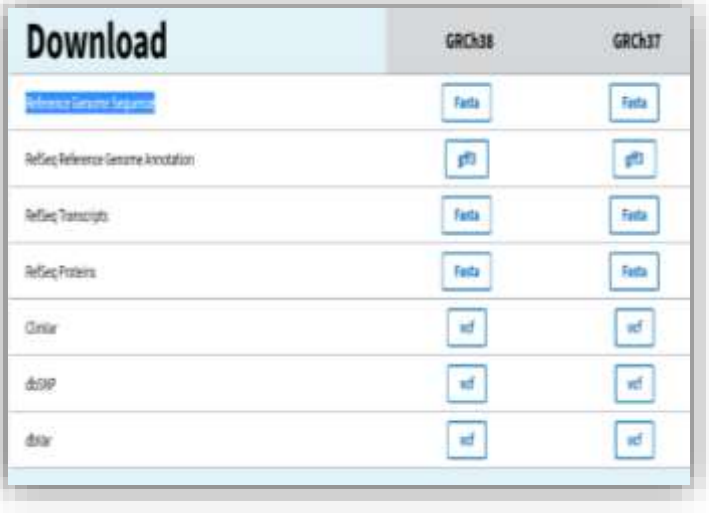


Figure 2

In the **multiple sequence alignment**, we downloaded four FASTA files from NCBI using the accession numbers of organisms (Mus musculus, Cricetulus griseus, and Bos primigenius, Homo sapiens). We added them all to a single file called multiple.fasta in order to have a multiple-sequenced FASTA file using the line of code shown in Figure 3 on the Ubuntu terminal on the Linux operating system:

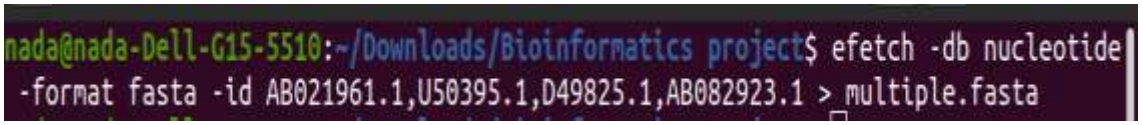


Figure 3

For the conversion of a FASTQ file to a FASTA file, we downloaded a FASTQ file from NCBI.

```
1 @ERR000001.1 IL2_62_3_1_346_881/1
2 GAACTAAGTGAACTGAAACATCTAAGTAACTTAAGG
3 +
4 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
5 @ERR000001.2 IL2_62_3_1_583_614/1
6 GATCCTACTATTACAATAATGCATTACAATATTACT
7 +
8 IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Figure 4

Bash Script

1. Pairwise Alignment

Pairwise Sequencing Alignment is used to find similarities between two biological sequences that may reveal functional, structural, and/or evolutionary links between protein or nucleic acid (Embl-Ebi, 2022).

1.1. Local Alignment

Local alignment called Smith-Waterman algorithm find one, or more, alignments describing the most similar region(s) within the sequences to be aligned (Embl-Ebi, 2022).They are can align protein and nucleotide sequences.

1.1.1. blastn

BLASTn (Nucleotide BLAST) compares one or more nucleotide query sequences to a subject nucleotide sequence or a nucleotide sequence database (Library Guides at UC Berkeley, 2022).

```
blastn -query $Qu -db $DB -out $Qu.vs.$DB.blastn_results.txt
```

Figure 5

1.1.2. blastp

BLASTp (Protein BLAST) compares one or more protein query sequences to a subject protein sequence or a protein sequence database (Library Guides at UC Berkeley, 2022).

```
blastp -query $Qu.small.fasta -db $DB -out $Qu.vs.$DB.blastp_results.txt
```

Figure 6

1.1.3. tblastx

Tblastx compares a protein query against a nucleotide sequence database to map a protein to genomic DNA (Library Guides at UC Berkeley, 2022).

```
tblastx -query $Qu.small.fasta -db $DB -out $Qu.vs.$DB.tblastx_results.txt
```

Figure 7

1.1.4. blastx

BLASTx (translated nucleotide sequence searched against protein sequences): compares a nucleotide query sequence to a protein sequence database. Blastx is especially beneficial when the query sequence's reading frame is unknown or contains faults that might result in frame shifts or other coding issues (Library Guides at UC Berkeley, 2022).

```
blastx -query  $Qu.small.fasta -db $DB -out $Qu.vs.$DB.blastx_results.txt
```

Figure 8

1.1.5. tblastn

Tblastn is a search tool for finding homologous protein coding regions in unannotated nucleotide sequences. It compares a protein query sequence against six-frame translations of a database of nucleotide sequences (Library Guides at UC Berkeley, 2022).

```
tblastn -query $Qu.small.fasta -db $DB -out $Qu.vs.$DB.tblastn_results.txt
```

Figure 9

1.2. Global Alignment

The Needleman-Wunsch algorithm is widely used for optimal global alignment, assigning a score to every possible alignment. We used Needle (EMBOSS) on Linux to create an optimal global alignment of two sequences using the algorithm (Embl-Ebi, 2022).

```
nada@nada-Dell-G15-5510:~/Downloads/Bioinformatics project$ needle
Needleman-Wunsch global alignment of two sequences
Input sequence: seq2.fasta
Second sequence(s): seq3.fasta
Gap opening penalty [10.0]: 3
Gap extension penalty [0.5]: 3
Output alignment [u50395.needle]: global
```

Figure 10

2. Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is the alignment of three or more similar length biological sequences (protein or nucleic acid). The result can be used to infer homology and the evolutionary connections between the sequences investigated (Embl-Ebi, 2022).

2.1. Muscle

Accurate MSA tool, especially good with proteins. Suitable for medium alignments (Embl-Ebi, 2022). Three different formats for the muscle alignment were used that are HTML, MSF, and clustalw. The

phylogenetic tree is an evolutionary tree of different organisms are identified and organized in a hierarchical structure in which closely related species are physically placed near each other.

2.1.1. MSF:

```
muscle -in $input -out $output.msf -msf
```

```
muscle -in $input -out $output.msf -msf -tree1 $output_tree.phy
plottree $output_tree.phy
```

Figure 11

2.1.2. HTML:

```
muscle -in $input -out $output.html -html
```

```
muscle -in $input -out $output.html -html -tree1 $output_tree.phy
plottree $output_tree.phy
```

Figure 13

2.1.3. CLUSTALW:

```
muscle -in $input -out $output.clw -clw
```

```
muscle -in $input -out $output.clw -clw -tree1 $output_tree.phy
plottree $output_tree.phy
```

Figure 14

2.2. Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments (Embl-Ebi, 2022).

```
kalign -i $input -f clu -o $output.clu
```

Figure 15

3. FASTQ to FASTA

FASTQ	FASTA
Text-based format that stores both sequence and associated sequence quality values.	Text-based format that stores only DNA and protein sequences.
Uses four lines description. • A line starting with @, containing the sequence ID. • One or more lines that contain the sequence. • A new line starting with the character +, and being either empty or repeating the sequence ID. • One or more lines that contain the quality scores.	Uses one line description. • The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column.

Using the line of code as shown in following figure 16 to convert the FASTQ file into a FASTA file to change from 4 lines to 2 lines and get rid of the quality scores and the (+) sign that indicates the split between the sequence and the scores to only a single lined sequence.

Figure 16

4. K-mers

K-mers are k-length substrings of a biological sequence that are used in bioinformatics. K-mers, which are made up of nucleotides (such as A, T, G, and C) and are primarily used in the context of computational genomics and sequence analysis. For example, the sequence AGAT would have four monomers (A, G, A, and T), three 2-mers (AG, GA, AT), two 3-mers (AGA and GAT) and one 4-mer (AGAT).

```
import pandas as pd
import numpy as np
import screed
from collections import Counter
def build_kmers(sequence, ksize):
    kmers = []
    n_kmers = len(sequence) - ksize + 1

    for i in range(n_kmers):
        kmer = sequence[i:i + ksize]
        kmers.append(kmer)

    return kmers

def read_kmers_fron_file(filename, ksize):
    all_kmers = []
    for record in screed.open(filename):
        sequence = record.sequence

        kmers = build_kmers(sequence, ksize)
        all_kmers += kmers
        y= len(all_kmers)
        print("sum of",ksize,"kmers = ",y)

    return Counter(all_kmers).most_common()
inputFileName = input("Enter name of input file: ")
n=int(input("Enter number of kmers : "))
for i in range (1,n+1):


  df = pd.DataFrame(read_kmers_fron_file(inputFileName, i))
  df.index = np.arange(1, len(df) + 1)
  print(df,"\n")
```

Figure 17

We implemented the K-mers in Python and called them in the bash script with the following Linux command:

```
python3 kmers.py
```

RESULTS

This is the first page of our automated script graphical user interface and represents the toolbox's main menu, from which the user can select the desired operation as shown in (fig.18).

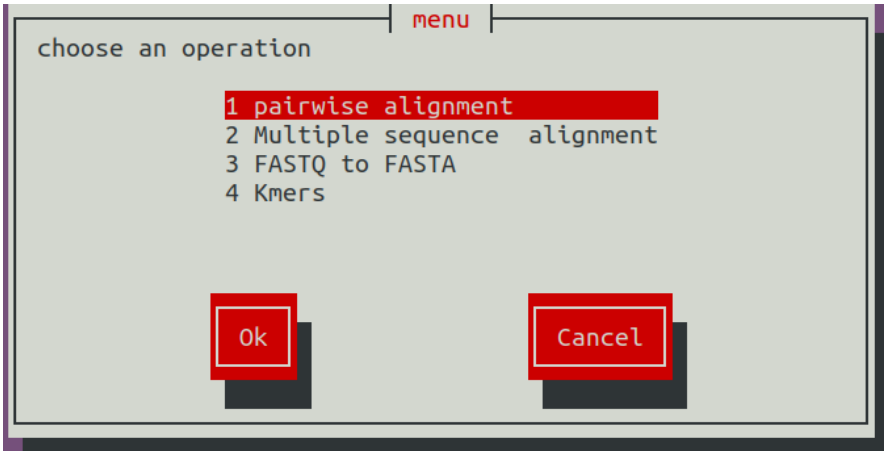


Figure 18

First operation: pairwise alignment, the user can select whether the alignment is global or local as shown in (fig.19).

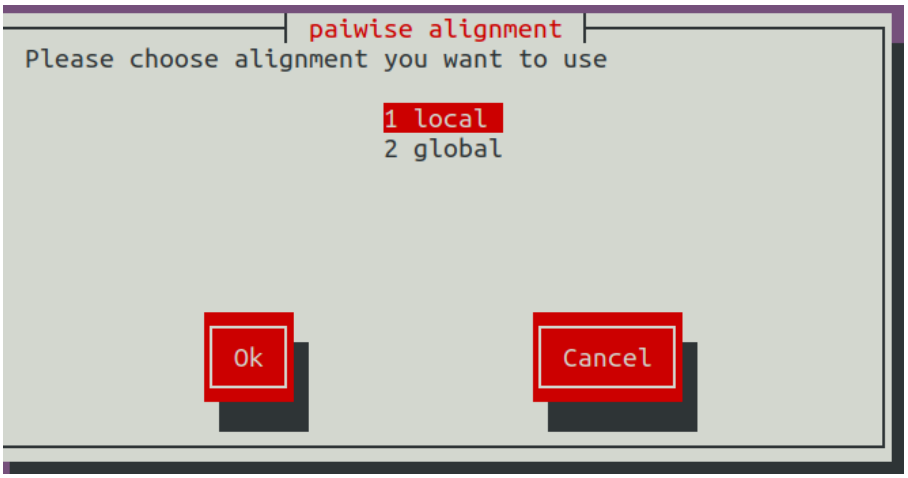


Figure 19

If the user clicks on **local alignment**, user will be able to choose from five different types of BLAST methods for alignment as shown in (fig.20).

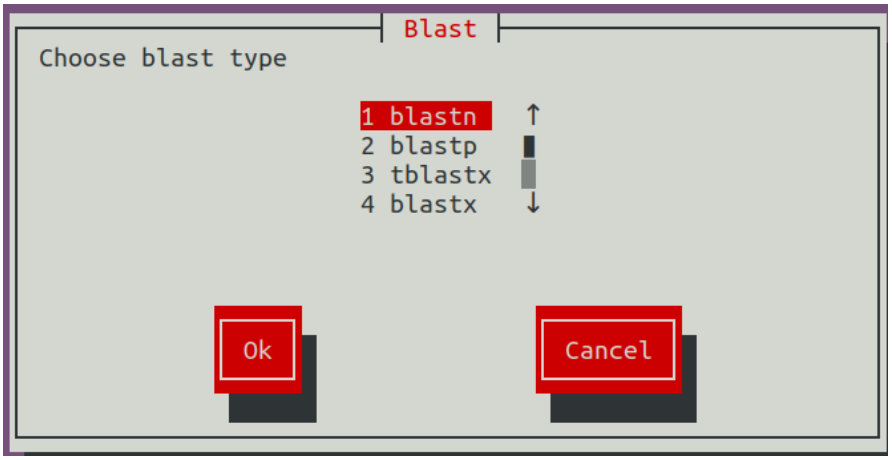
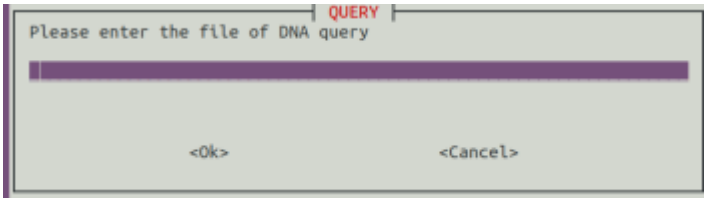
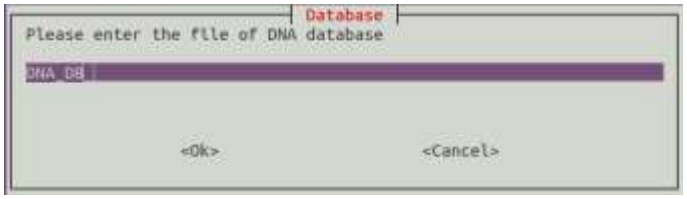


Figure 20

In **Blastn**, the user will be asked for the query file and database file as input.





Output file:

The output of the blast file consists of the alignment, the bit score, the E value, and the gaps. The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size.

bit-score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity.

Gaps are spaces introduced into an alignment to compensate for insertions and deletions in one sequence relative to another.

Figure 21

In **blastp**, the user will be asked for the protein query file and protein database file as input.

Output file:

```
BLASTP 2.9.0+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.


Reference for composition-based statistics: Alejandro A. Schaffer,
L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri
I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001),
"Improving the accuracy of PSI-BLAST protein database searches with
composition-based statistics and other refinements", Nucleic Acids
Res. 29:2994-3005.


Database: P_OB.faa
           12,006 sequences; 8,355,456 total letters


Query= XP_005201558.1 dynamin-like 120 kDa protein, mitochondrial isoform
X3 [Bos taurus]

Length=400

                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

NP_570844.1 dynamin-like 120 kDa protein, mitochondrial isoform 2...  788     0.0
NP_570846.1 dynamin-like 120 kDa protein, mitochondrial isoform 4...  689     0.0
NP_001005336.1 dynamin-1 isoform 2 [Homo sapiens]                     62.8    2e-10
XP_047273373.1 dynamin-3 isoform X17 [Homo sapiens]                   56.2    1e-08
XP_047273466.1 arf-GAP with Rho-GAP domain, ANK repeat and PH dom...  29.6    2.5
XP_047301616.1 titin isoform X2 [Homo sapiens]                        29.3    3.5
NP_001393916.1 lamin isoform G [Homo sapiens]                         28.5    5.7
NP_001076437.1 TOM1-like protein 2 isoform 3 [Homo sapiens]           28.5    6.2
XP_005256519.1 TOM1-like protein 2 isoform X1 [Homo sapiens]          28.5    6.3
XP_047274740.1 mitogen-activated protein kinase kinase kinase 5 i...  28.5    6.3
NP_982271.3 synaptojanin-1 isoform b [Homo sapiens]                   28.5    7.3
XP_047297806.1 synaptojanin-1 isoform X20 [Homo sapiens]              28.1    7.8
NP_001153778.1 synaptojanin-1 isoform d [Homo sapiens]                28.1    8.1
XP_047296991.1 synaptojanin-1 isoform X4 [Homo sapiens]               28.1    8.1
NP_001340531.1 lysosomal cobalamin transporter ABCD4 isoform k [H...  27.7    9.1
NP_001340532.1 lysosomal cobalamin transporter ABCD4 isoform k [H...  27.7    9.1
NP_001340538.1 lysosomal cobalamin transporter ABCD4 isoform l [H...  27.7    9.4
NP_001340535.1 lysosomal cobalamin transporter ABCD4 isoform l [H...  27.7    9.4


>NP_570844.1 dynamin-like 120 kDa protein, mitochondrial isoform
2 [Homo sapiens]
Length=924

 Score = 788 bits (1827),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 347/400 (87%), Positives = 354/400 (89%), Gaps = 36/400 (9%)

Query  1    MWRVRQAAVACEVCQSLVKHSSGTKGSLPLQKLHLVSRSIYHTHHPTLKLQRPQLRMSLQ  60
            MWR+R+AAVACEVCQSLVKHSSG KGSLPLQKLHLVSRSIYH+HHPTLKLQRPQLR S Q
Sbjct  1    MWRLRRAAVACEVCQSLVKHSSGIKGSLPLQKLHLVSRSIYHSHHPTLKLQRPQLRTSFQ  60

Query  61   QFSSLTSLPLRKLKLSPVKYGYQPHRNFWPARLAARLLKLRYLILGSAVGGGYTAKKTFD  120
            QFSSLT+LPLRKLK SP+KYGYQP RNFWPARLA RLLKLRYLILGSAVGGGYTAKKTFD
Sbjct  61   QFSSLTNLPLRKLKFSPIKYGYQPRRNFWPARLATRLLKLRYLILGSAVGGGYTAKKTFD  120

Query  121  QWKDMIPDLSDYKWIVPDIVWEIDEYIDLEKIRKALPNSEDLAKLAPDFDKIAESLSLLK  180
            QWKDMIPDLS+YKWIVPDIVWEIDEYID
Sbjct  121  QWKDMIPDLSEYKWIVPDIVWEIDEYIDF----------------------------  149
```

If the user clicks on **global alignment**, the user will be asked for the two input FASTA files, gap opening and extension, to perform the Needleman algorithm. (fig.22).

Output file result:

```
########################################
# Program: needle
# Rundate: Mon 30 Jan 2023 16:07:50
# Commandline: needle
#    -asequence seq2.fasta
#    -bsequence seq3.fasta
#    -gapopen 3
#    -gapextend 3
#    -outfile global
# Align_format: srspair
# Report_file: global
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: U50395.1
# 2: AB082923.1
# Matrix: EDNAFULL
# Gap_penalty: 3.0
# Extend_penalty: 3.0
#
# Length: 2569
# Identity:    1468/2569 (57.1%)
# Similarity:  1468/2569 (57.1%)
# Gaps:         858/2569 (33.4%)
# Score: 5504.0
#
#
#=======================================

U50395.1          1 TTGGGATCAGGGGGCACCTGGGATCCGGGTCCGGAATC-TACTTCCCAGC     49
                                                       | |.|||.||| |
AB082923.1        1 -----------------------------------CGTGCTTTCCA-C     12

U50395.1         50 G--GGTAACACGCTCCCCCCTGAAGACCTGAAGCCTGGCTGACTTCTTGG     97
                    |  |||-|||||||  -|||||  || -||  |||-|-|||-|||| -||
AB082923.1       13 GACGGTGACACGCT--TCCCTG--GA-TTG--GCCAGACTGCCTTC-CGG     54

U50395.1         98 GT-GCTGCCATGGAGGAGCCACAGTCAGA-CCTCAGCATCGAGCTCCCTC    145
                    || .||||||||||||||||.||||||| ||| |||-|||||-|||||
AB082923.1       55 GTCACTGCCATGGAGGAGCCGCAGTCAGATCCT-AGCGTCGAGCCCCCTC    103
```

Figure 22

Second operation: multiple sequence alignment (MSA), user will be able to choose type of algorithm (muscle or Kalign). As shown in (fig.23).

If the user clicks on Muscle, the user will be able to choose from three different muscle formats (Clw, HTML, and MSF), and then be asked for the MSA file (multiple.fasta).

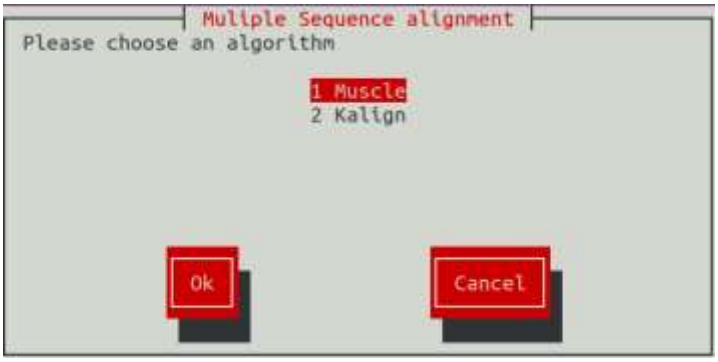


Figure 23

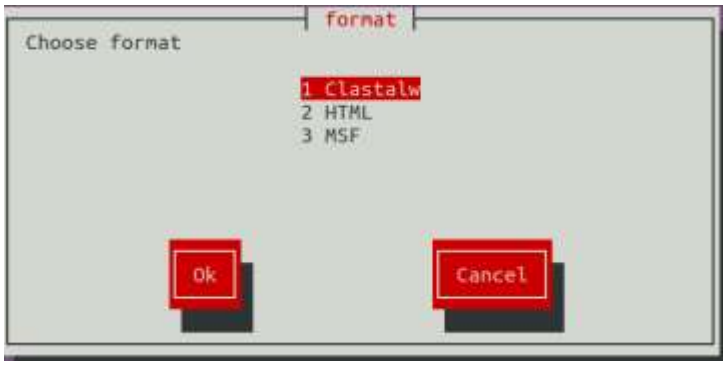


Figure 24

Output File:

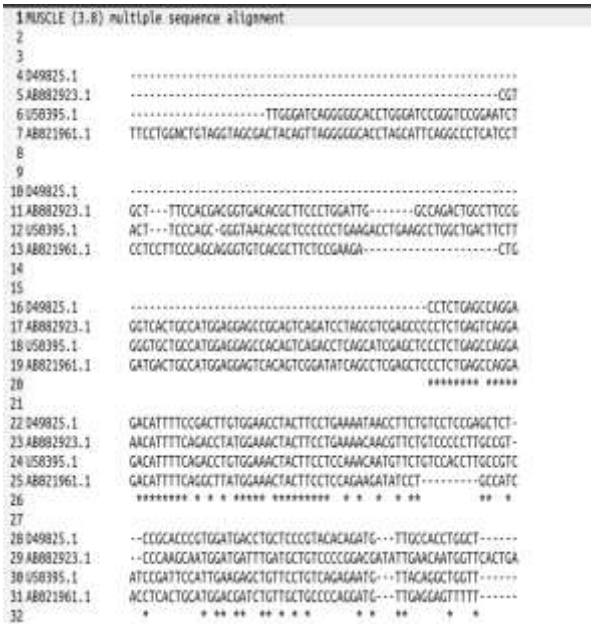


Figure 25 Clustalw format

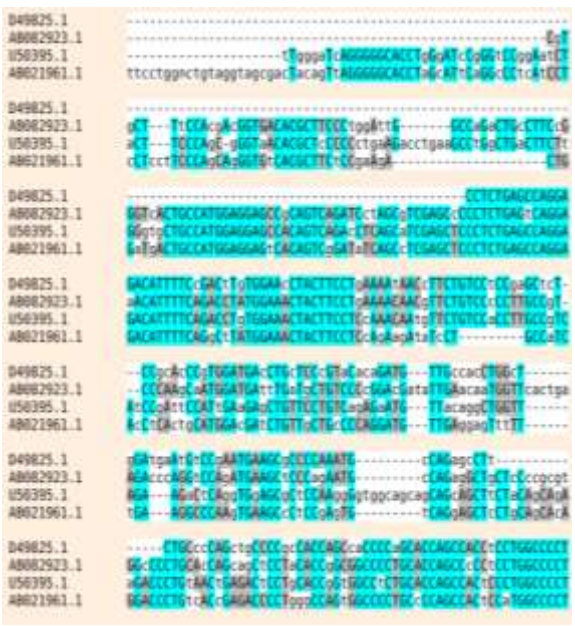


Figure 26 Html format

Output Tree:

```
1
2 U58395.1 Cricetulus griseus wild type tumor supressor P53 (p53) mRNA, complete cds:0.675811
3 ,
4 (
5 AB821961.1 Mus musculus mutant p53 mRNA, complete cds:0.65614
6 ,
7 (
8 D49825.1 Bos primigenius p53 mRNA, partial cds:0.497326
9 ,
10 AB882923.1 Homo sapiens mRNA for P53, complete cds:0.497326
11 ):0.158814
12 ):0.8188781
13
14 ;
```

Figure 31 New wick format

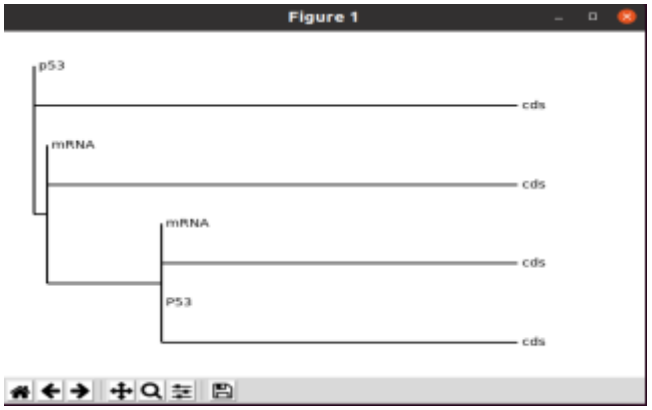


Figure 32 Tree visualization

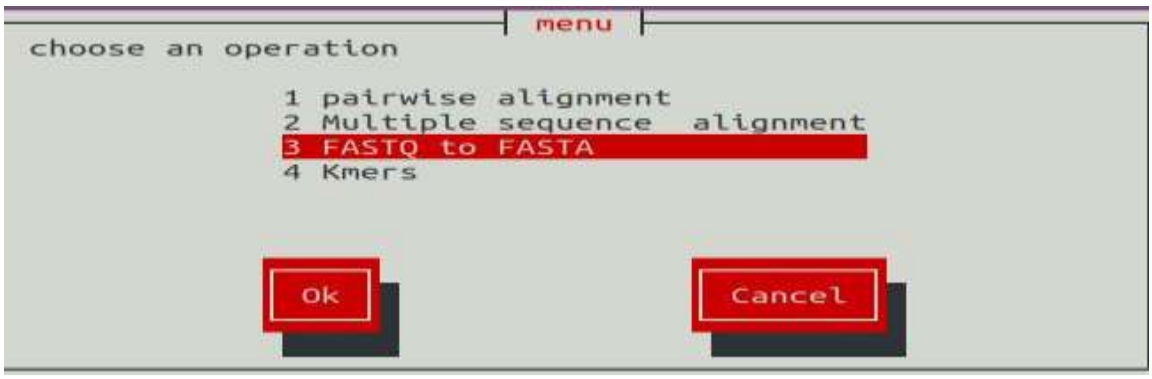
If the user clicks on Kalign, The user will be asked for the MSA file (multiple.fasta).

Output File:

```
1 Kalign (2.0) alignment in ClustalW format
2
3
4 AB821961      TTCCTGGNCTGTAGGTAGCGACTACAGTTAGGGGGCACCTAGCATTCAGCCCCTCATCCT
5 U58395        TT----GGGATCAGGGGGCACCTGGGATCCGGGTCCGGAATCTACTTCCCAGCGGGTAAC
6 D49825        C-----------------------------------------------------------
7 AB882923      C-----------------------------GTG----------CTTTCCACGACGGTGAC
8
9
10 AB821961     CCTCCTTCCCAGCAGGGTGTCACGCTTCTCCGAAGACTGGATGACTGCCATGGAGGAGTC
11 U58395       ACGCTCCCCCCTGAAGACCTGAAGCCTGGCTGACTTCTTGGGTGCTGCCATGGAGGAGCC
12 D49825       ------------------------------------------------------------
13 AB882923     ACGCTTCCCT-----GGATTG--GCCAGACTGCCTTCCGGGTCACTGCCATGGAGGAGCC
14
15
16 AB821961     ACAGTCGGATATCAGCCTCGAGCTCCCTCTGAGCCAGGAGACATTTTCAGGCTTATGGAA
17 U58395       ACAGTCAGACCTCAGCATCGAGCTCCCTCTGAGCCAGGAGACATTTTCAGACCTGTGGAA
18 D49825       ---------------------------CTCTGAGCCAGGAGACATTTTCCGACTTGTGGAA
19 AB882923     GCAGTCAGATCCTAGCGTCGAGCCCCCTCTGAGTCAGGAAACATTTTCAGACCTATGGAA
20
21
22 AB821961     ACTACTTCCTCCAGAAGATATCC---------TGCCATCACCTCACTGCATGGACGATCT
23 U58395       ACTACTTCCTCCAAACAATGTTCTGTCCACCTTGCCGTCATCCGATTCCATTGAAGAGCT
24 D49825       CCTACTTCCTGAAAATAACCTTCTGT---CCTCCGAGCTCTCCGCACCCGTGGATGACCT
25 AB882923     ACTACTTCCTGAAAACAACGTTCTGTCCCCCTTGCCGTCCCAAGCAATGGATGATTTGAT
26
27
28 AB821961     GTTGCTGCCCCAGGATGTTGAGGAGTTTTTT---GAAGGCCCAAG---------TGAAGC
29 U58395       GTTCCTGTCAGAGAATGTTACAGGCTGGTTA---GAAGACTCAGGTGGAGCGCTCCAAGG
30 D49825       GCTCCCGTACACACGATGTTGCCACCTGGCTGGATGAATGTCCGAA---------TGAAGC
31 AB882923     GCTGTCCCCGGACGATATTGAACAATGGTTCACTGAAGACCCAGGTCCAGA---TGAAGC
```

Figure 27 Kalign alignment

Third operation: FastQ conversion, the user will input the FastQ file, and it will be converted to a FASTA file.



Input File:

```
@ERR000001.1 IL2_62_3_1_346_881/1
GAACTAAGTGAACTGAAACATCTAAGTAACTTAAGG
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@ERR000001.2 IL2_62_3_1_583_614/1
GATCCTACTATTACAATAATGCATTACAATATTACT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@ERR000001.3 IL2_62_3_1_389_877/1
GGGAGACAATGCAGAGGTTGAAAGATGTATCTGAAA
+
IIIIIIIIIIIIIIIIIII>IIICIIIII-IIIII
@ERR000001.4 IL2_62_3_1_284_606/1
TTAACGACCGTACCGAAAGTGGACTTAAGTAGTATG
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII8
@ERR000001.5 IL2_62_3_1_480_810/1
GGTTTGCTTCAAGAATAGCTTTGGTTTGTAAAGGTT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@ERR000001.6 IL2_62_3_1_576_286/1
GATTTGTCAATCACTCGTGTTCCTTCCTATGTTTGT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@ERR000001.7 IL2_62_3_1_641_293/1
GGAAATGAAGGAAATGGAATTGCGTATTGTTGAATC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII2IIII
```

Figure 28 Fastq

Output File:

```
>ERR000001.1 IL2_62_3_1_346_881/1
GAACTAAGTGAACTGAAACATCTAAGTAACTTAAGG
>ERR000001.2 IL2_62_3_1_583_614/1
GATCCTACTATTACAATAATGCATTACAATATTACT
>ERR000001.3 IL2_62_3_1_389_877/1
GGGAGACAATGCAGAGGTTGAAAGATGTATCTGAAA
>ERR000001.4 IL2_62_3_1_284_606/1
TTAACGACCGTACCGAAAGTGGACTTAAGTAGTATG
>ERR000001.5 IL2_62_3_1_480_810/1
GGTTTGCTTCAAGAATAGCTTTGGTTTGTAAAGGTT
>ERR000001.6 IL2_62_3_1_576_286/1
GATTTGTCAATCACTCGTGTTCCTTCCTATGTTTGT
>ERR000001.7 IL2_62_3_1_641_293/1
GGAAATGAAGGAAATGGAATTGCGTATTGTTGAATC
>ERR000001.8 IL2_62_3_1_801_750/1
GGGATTTTAAAATTATTATTATATTTAAGAATAAGA
>ERR000001.9 IL2_62_3_1_386_889/1
TTATGTAGTACCTTTGTAATTATAATCATGATGATA
>ERR000001.10 IL2_62_3_1_866_369/1
GTCTTGAGTGAAGTTAAGGCCGAAGGCTTTGACAAA
>ERR000001.11 IL2_62_3_1_677_478/1
GTAACTACATGTAAAATGTAATCCATAACCGTGTAA
```

Figure 29 Fasta file

Fourth operation: Counting Kmers, the user will input the FASTA file and the number of Kmers.

```
nada@nada-Dell-G15-5510:~/Downloads/Bioinformatics project$ bash last.sh
Enter name of input file: seq2.fasta
Enter number of kmers : 2
sum of 1 Kmers =  1829
   0    1
1  C  523
2  G  439
3  A  434
4  T  433

sum of 2 Kmers =  1828
     0    1
1    CC  186
2    CT  158
3    TG  151
4    CA  137
5    GG  129
6    TC  122
7    AA  122
8    GA  117
9    AG  117
10   GC  109
11   AC  106
12   TT  102
13   AT   89
14   GT   83
15   TA   58
16   CG   42
```

Figure 30 Kmers count

CONCLUSION

This report focuses on bioinformatics and its applications in modern biology and medicine. The main tools of a bioinformatician are computer software programs and the internet, with sequence analysis of DNA and proteins being a fundamental activity. Bioinformatics is an evolving discipline and expert bioinformaticians use complex software programs for retrieving, analyzing, and storing biological data. The growth of bioinformatics has been a global venture, creating computer networks and freely available gene and protein databases accessible to the scientific community. The report explains the use of Linux as the preferred operating system for bioinformatics due to its robust command-line interface, and the development of a fully automated command-line application with a graphical user interface for easier use. The report describes the methodology used for data collection and analysis, which includes downloading FASTA files from NCBI for pairwise and multiple sequence alignment, and downloading a FASTQ file for conversion to a FASTA file. The report concludes by highlighting the importance of bioinformatics in the discovery and understanding of biological data, and its potential contribution to the development of new drugs and individualized therapy.

REFERENCES

Brihadiswaran, G. (2020) Bioinformatics 1: K-Mer counting, Medium. The Startup. Available at: https://medium.com/swlh/bioinformatics-1-k-mer-counting-8c1283a07e29#:~:text=What%20are%20k%2Dmers%3F,which%20consists%20of%2011%20nucleotides

Embl-Ebi Pairwise sequence alignment, EBI. Available at: https://www.ebi.ac.uk/Tools/psa/#:~:text=Pairwise%20Sequence%20Alignment%20is%20used,(protein%20or%20nucleic%20acid) .

National Center for Biotechnology Information. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/

(S.webteam), www-core (2022, February 21). What is bioinformatics and how do we use it? @yourgenome · Science website. Retrieved January 30, 2023, from https://www.yourgenome.org/facts/what-is-bioinformatics-and-how-do-we-use-it/

TEAM CONTRIBUTION

Abstract, Introduction – Sarah Mahmoud

Methodology, Results – All Team members

Conclusion – Mohamed ElSayed