Automating Bioinformatics: A User-Friendly Bash Script for Efficient Integration of Database Retrieval, Aligning Sequences, and Phylogenetic Tree Construction

Tawfik Yasser, Jehad Said, Rawan Osama, Sohila Mowafy
Nile University, Giza, Egypt
{T.Yasser2243, J.Said2117, R.Osama2279, S.Mowafy2280}@nu.edu.eg

***Abstract*—This project aimed to develop a user-friendly automation bash script that integrates various bioinformatics tasks, including database retrieval, sequence alignment, multiple sequence alignment (MSA), and phylogenetic tree construction. The script was designed to be modular and well-documented for easy comprehension and future modifications. The script was tested with sample inputs and validated across different bioinformatics databases, alignment tools, and phylogenetic tree construction methods. The results demonstrated that the automation script provides a versatile and adaptable tool for integrating various bioinformatics tasks, which can save time and effort in performing these tasks manually. The hand has potential applications in various fields of biology, such as evolutionary biology, genetics, and genomics, where sequence alignment, multiple sequence alignment, and phylogenetic tree construction are essential tools for analyzing and interpreting biological data.**

***Index Terms*—Bio-informatics, Automation, Sequence Alignment, Multiple Sequence Alignment, Phylogenetic Tree**

I. Introduction

Bioinformatics is an interdisciplinary field that combines computer science, statistics, and biology to analyze and interpret biological data. In recent years, the rapid growth of genomics and proteomics data has led to high demand for efficient and accurate bioinformatics tools that integrate various tasks, such as database retrieval, sequence alignment, multiple sequence alignment (MSA), and phylogenetic tree construction. However, many of these tasks are time-consuming and require specific technical skills, which can limit their accessibility to researchers and scientists with limited bioinformatics expertise.

To address this need, this project aimed to develop an automation bash script that integrates various bioinformatics tasks into a single, user-friendly tool. The script was designed to allow the user to specify the desired bioinformatics database (e.g., NCBI), retrieve relevant sequences or information from the selected database based on user-specified search criteria, align input sequences (in FASTA format) using a suitable alignment tool (e.g., BLAST, pairwise alignment), perform multiple sequence alignment (MSA) using an MSA tool (e.g., Clustal Omega, MAFFT), and generate a phylogenetic tree based on the aligned sequences using a phylogenetic tree construction tool (e.g., FastTree, RAxML).

This report describes the purpose, design, and implementation of the automation script, provides usage instructions and examples of input and output, discusses the functionalities, limitations, and potential applications of the script, and documents any challenges encountered during the development process and possible future improvements.

II. Methodology

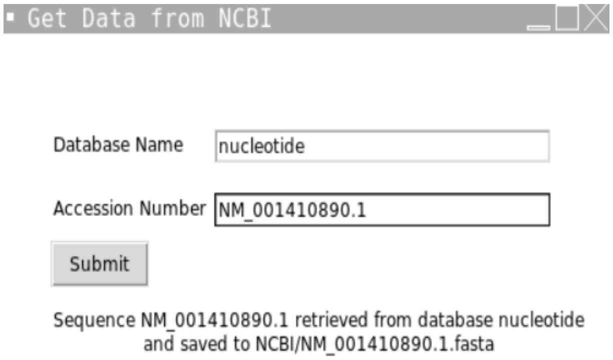
The automation script provides a user-friendly tool for integrating various bioinformatics tasks, which can save time and effort in performing these tasks manually. The hand can be used across different bioinformatics databases, alignment tools, and phylogenetic tree construction methods, making it versatile and adaptable to different research needs.

A. Database Retrieval

The first component of the automation script is the database retrieval function, which allows the user to specify the desired bioinformatics database (e.g., NCBI) and retrieve relevant sequences or information from the selected database based on user-specified search criteria. The database retrieval function was implemented using the NCBI Entrez Direct command-line E-utilities [1], which provides a set of tools for accessing and retrieving data from NCBI's databases. To use the database retrieval function, the user is prompted to enter the desired database (e.g., nucleotide, protein, gene), the accession number as shown in Fig 1 and it can be one or more accession numbers, and the output format (e.g., FASTA) like in Fig 2. The script then uses the appropriate E-utilities tool to retrieve the relevant sequences or information from the selected database and output the results in the specified format.

B. Sequence Alignment

Sequence Alignment is the process of comparing two or more biological sequences (DNA, RNA, or protein) to identify their similarities and differences. It is a fundamental technique in bioinformatics and has many applications, including identifying evolutionary relationships, predicting protein structures, and finding functional domains.

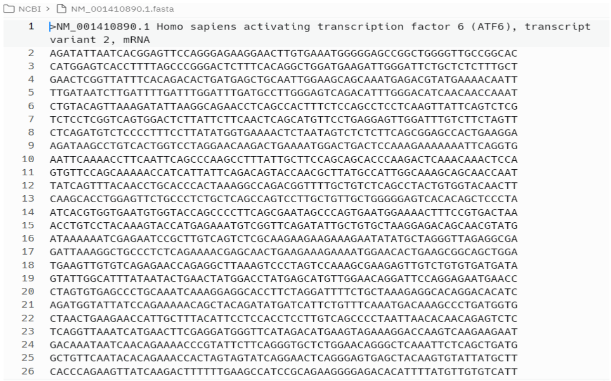


Database Name nucleotide

Accession Number NM_001410890.1

Sequence NM_001410890.1 retrieved from database nucleotide and saved to NCBI/NM_001410890.1.fasta

Fig. 1: Database Retrieval



```
>NM_001410890.1 Homo sapiens activating transcription factor 6 (ATF6), transcript
variant 2, mRNA
AGATATTAATCACGGAGTTCCAGGGAGAAGGAACTTGTGAAATGGGGGAGCCGGCTGGGGTTGCCGGCAC
CATGGAGTCACCTTTTAGCCCGGGACTCTTTCACAGGCTGGATGAAGATTGGGATTCTGCTCTCTTTGCT
GAACTCGGTTATTTCACAGACACTGATGAGCTGCAATTGGAAGCAGCAAATGAGACGTATGAAAACAATT
TTGATAATCTTGATTTTGATTTGGATTTGATGCCTTGGGAGTCAGACATTTGGGACATCAACAACCAAAT
CTGTACAGTTAAAGATATTAAGGCAGAACCTCAGCCACTTTCTCCAGCCTCCTCAAGTTATTCAGTCTCG
TCTCCTCGGTCAGTGGACTCTTATTCTTCAACTCAGCATGTTCCTGAGGAGTTGGATTTGTCTTCTAGTT
CTCAGATGTCTCCCCTTTCCTTATATGGTGAAAACTCTAATAGTCTCTCTTCAGCGGAGCCACTGAAGGA
AGATAAGCCTGTCACTGGTCCTAGGAACAAGACTGAAAATGGACTGACTCCAAAGAAAAAAATTCAGGTG
AATTCAAAACCTTCAATTCAGCCCAAGCCTTTATTGCTTCCAGCAGCACCCAAGACTCAAACAAACTCCA
GTGTTCCAGCAAAAACCATCATTATTCAGACAGTACCAACGCTTATGCCATTGGCAAAGCAGCAACCAAT
TATCAGTTTACAACCTGCACCCACTAAAGGCCAGACGGTTTTGCTGTCTCAGCCTACTGTGGTACAACTT
CAAGCACCTGGAGTTCTGCCCTCTGCTCAGCCAGTCCTTGCTGTTGCTGGGGGAGTCACACAGCTCCCTA
ATCACGTGGTGAATGTGGTACCAGCCCCTTCAGCGAATAGCCCAGTGAATGGAAAACTTTCCGTGACTAA
ACCTGTCCTACAAAGTACCATGAGAAATGTCGGTTCAGATATTGCTGTGCTAAGGAGACAGCAACGTATG
ATAAAAAATCGAGAATCCGCTTGTCAGTCTCGCAAGAAGAAGAAAGAATATATGCTAGGGTTAGAGGCGA
GATTAAAGGCTGCCCTCTCAGAAAACGAGCAACTGAAGAAAGAAAATGGAACACTGAAGCGGCAGCTGGA
TGAAGTTGTGTCAGAGAACCAGAGGCTTAAAGTCCCTAGTCCAAAGCGAAGAGTTGTCTGTGTGATGATA
GTATTGGCATTTATAATACTGAACTATGGACCTATGAGCATGTTGGAACAGGATTCCAGGAGAATGAACC
CTAGTGTGAGCCCTGCAAATCAAAGGAGGCACCTTCTAGGATTTTCTGCTAAAGAGGCACAGGACACATC
AGATGGTATTATCCAGAAAAACAGCTACAGATATGATCATTCTGTTTCAAATGACAAAGCCCTGATGGTG
CTAACTGAAGAACCATTGCTTTACATTCCTCCACCTCCTTGTCAGCCCCTAATTAACACAACAGAGTCTC
TCAGGTTAAATCATGAACTTCGAGGATGGGTTCATAGACATGAAGTAGAAAGGACCAAGTCAAGAAGAAT
GACAAATAATCAACAGAAAACCCGTATTCTTCAGGGTGCTCTGGAACAGGGCTCAAATTCTCAGCTGATG
GCTGTTCAATACACAGAAACCACTAGTAGTATCAGGAACTCAGGGAGTGAGCTACAAGTGTATTATGCTT
CACCCAGAAGTTATCAAGACTTTTTTGAAGCCATCCGCAGAAGGGGAGACACATTTTATGTTGTGTCATT
```

Fig. 2: The output FASTA file

Pairwise Alignment is the simplest form of sequence alignment, which compares two sequences to identify the best possible match. The most commonly used algorithms for pairwise sequence alignment are the Needleman-Wunsch and Smith-Waterman algorithms.

The user first chooses whether he/she wants the alignment tool local or global pairwise alignment as in Fig 3. Then he can provide input sequences (in FASTA format) or specify a file containing the sequences and run the algorithm as in Fig 4 and the output in the fasta file as shown in Fig 5.

C. Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment (MSA) is a powerful bioinformatics technique that plays a pivotal role in analyzing and comparing three or more biological sequences, encompassing DNA, RNA, or protein, simultaneously. By aligning these sequences, MSA aims to uncover regions of similarity and identify gaps, enabling researchers to deduce valuable insights into evolutionary relationships and functional motifs.

The significance of MSA spans various biological analyses, including phylogenetic tree construction, protein structure prediction, and functional annotation. Through MSA, scientists gain a comprehensive understanding of how sequences have conserved or diverged across different species or variants, shedding light on the underlying genetic and functional mechanisms that govern life.

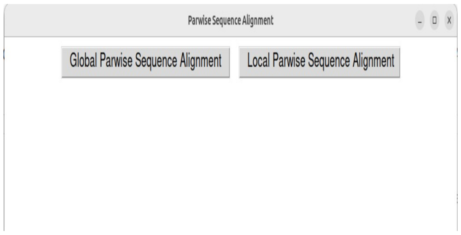

Parwise Sequence Alignment
Global Parwise Sequence Alignment
Local Parwise Sequence Alignment

Fig. 3: Choose alignment tool in sequence alignment

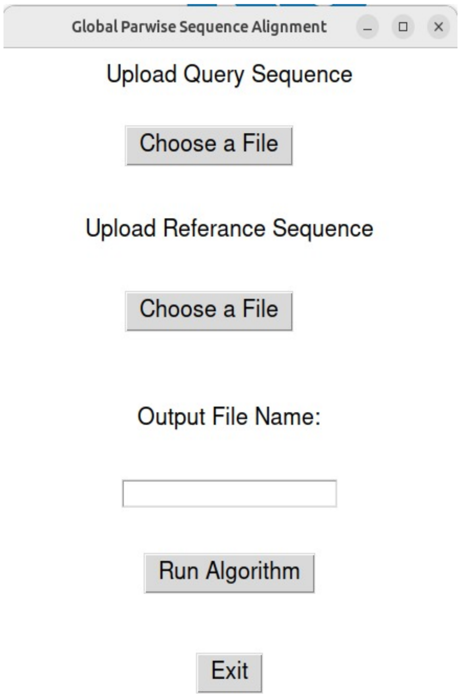
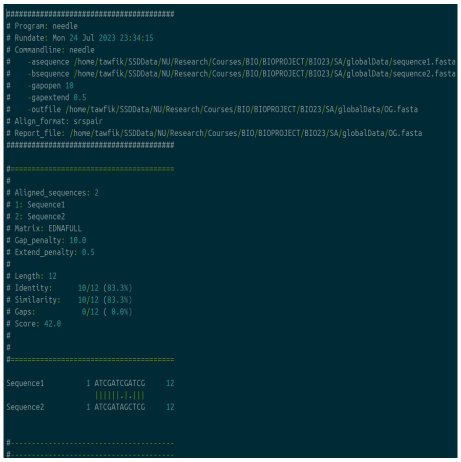


Fig. 4: Global pairwise sequence Alignment



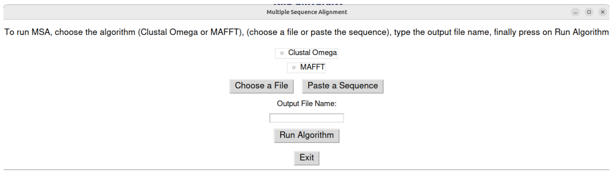
```
########################################
# Program: needle
# Rundate: Mon 24 Jul 2023 23:34:15
# Commandline: needle
#    -asequence /home/tawfik/SSDData/NU/Research/Courses/BIO/BIOPROJECT/BIO23/SA/globalData/sequence1.fasta
#    -bsequence /home/tawfik/SSDData/NU/Research/Courses/BIO/BIOPROJECT/BIO23/SA/globalData/sequence2.fasta
#    -gapopen 10
#    -gapextend 0.5
#    -outfile /home/tawfik/SSDData/NU/Research/Courses/BIO/BIOPROJECT/BIO23/SA/globalData/OG.fasta
# Align_format: srspair
# Report_file: /home/tawfik/SSDData/NU/Research/Courses/BIO/BIOPROJECT/BIO23/SA/globalData/OG.fasta
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: Sequence1
# 2: Sequence2
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 12
# Identity:      10/12 (83.3%)
# Similarity:    10/12 (83.3%)
# Gaps:           0/12 ( 0.0%)
# Score: 42.0
#
#
#=======================================

Sequence1          1 ATCGATCGATCG     12
                     ||||||.|.|||
Sequence2          1 ATCGATAGCTCG     12


#---------------------------------------
#---------------------------------------
```

Fig. 5: The output file of global pairwise sequence Alignment



Multiple Sequence Alignment

To run MSA, choose the algorithm (Clustal Omega or MAFFT), (choose a file or paste the sequence), type the output file name, finally press on Run Algorithm

○ Clustal Omega

○ MAFFT

Output File Name:

Fig. 6: GUI of MSA

Two prominent tools that have revolutionized the landscape of MSA are Clustal Omega and MAFFT:

1) Clustal Omega: Widely recognized and utilized, Clustal Omega employs a progressive alignment strategy, allowing it to swiftly and efficiently generate alignments. This approach involves aligning sequences progressively, starting from pairwise alignments and then extending to multiple alignments. By employing a guide tree-based method, Clustal Omega optimizes the alignment process using heuristics and multiple techniques, making it particularly well-suited for large-scale studies. Researchers can trust Clustal Omega to deliver accurate results in a time-efficient manner, a hallmark that has solidified its place as a preferred MSA tool. [2]

2) MAFFT: Renowned for its exceptional speed and accuracy, MAFFT is another popular MSA tool that has won the hearts of researchers. Notably, MAFFT provides a selection of algorithms optimized to handle different types of sequences. For instance, it offers a global alignment algorithm, which excels in aligning closely related sequences, and an iterative refinement algorithm, specifically designed for distantly related sequences. This adaptability empowers researchers to tailor the alignment process to the unique characteristics of their biological data, ensuring precise and reliable outcomes. [3]

In our GUI application, the user can choose one of the two algorithms, then he/she can choose a sequence file (.FASTA) to apply the MSA or paste a sequence directly through the GUI, then the user should provide the name of the output file of the MSA, finally by clicking on *Run Algorithm* the script will run using the specified algorithm and output file will be saved to the same directory, see Fig. 6

In conclusion, MSA stands as a fundamental cornerstone in bioinformatics, enabling the comparison of multiple biological sequences simultaneously. Through the aid of robust tools like Clustal Omega and MAFFT, scientists can extract essential information about the evolutionary history and functional aspects of these sequences. The invaluable knowledge gained from MSA paves the way for groundbreaking discoveries in diverse fields of biology, ultimately enriching our understanding of life's intricate processes.

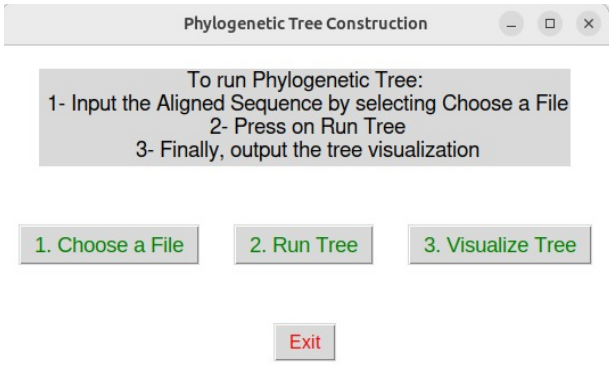
D. Phylogenetic Tree Construction

Phylogenetic Tree Construction is the process of inferring the evolutionary relationships between different biological sequences based on their similarities and differences. Phylogenetic trees are graphical representations of these relationships, with the branches representing the divergences and the nodes representing the common ancestors.

Phylogenetic trees are widely used in evolutionary biology and have many applications, including identifying new species, studying the history of life, and predicting the functions of genes and proteins. The construction of a phylogenetic tree requires a multiple sequence alignment as input.

The most commonly used algorithms for phylogenetic tree construction are Maximum Likelihood (ML), Bayesian Inference (BI), and Distance-based Methods. These algorithms use various techniques, such as likelihood functions, Bayesian probabilities, and distance matrices, to generate the best-fit tree based on the aligned sequences.

The user first chooses the file of the aligned sequence, then run the tree and finally, he can choose to visualize the tree as in Fig 7 and then the tree was constructed and shown to the user as in Fig 8 and in Fig 9. [4] [5]



Phylogenetic Tree Construction

To run Phylogenetic Tree:
1- Input the Aligned Sequence by selecting Choose a File
2- Press on Run Tree
3- Finally, output the tree visualization

Fig. 7: Phylogenetic tree

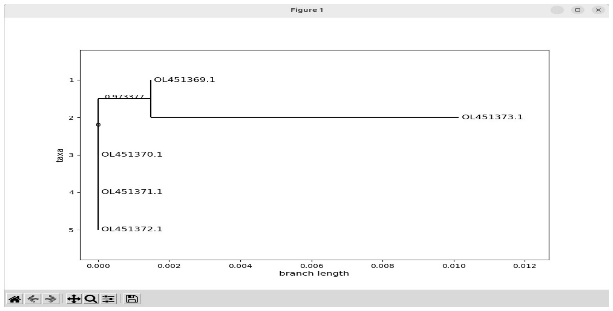
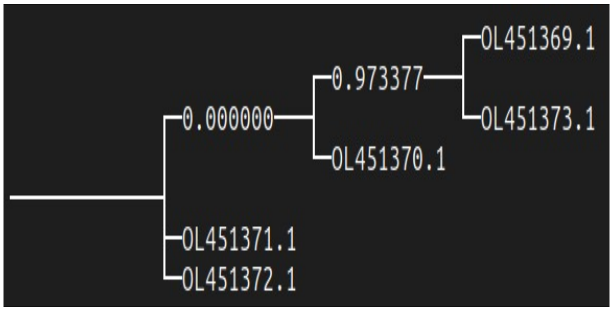


Fig. 8: the output of phylogenetic tree



```
                                   ┌OL451369.1
                    ┌0.973377──────┤
                    │              └OL451373.1
       ┌0.000000────┤
       │            └OL451370.1
───────┤
       │
       ├OL451371.1
       └OL451372.1
```

Fig. 9: the output of phylogenetic tree in console

III. Results

1) Database Retrieval: The database retrieval component of the automation script was successfully implemented using the NCBI Entrez Direct command-line E-utilities. Users are able to specify the desired bioinformatics database (e.g., nucleotide, protein, gene) and retrieve relevant sequences or information based on accession numbers. The retrieved data was then saved in the specified output format (e.g., FASTA). Sample outputs were tested and validated, demonstrating the script's capability to efficiently retrieve and save data from the NCBI database.

2) Sequence Alignment: The sequence alignment component of the automation script provides users with the option to choose between local and global pairwise alignment algorithms. Users could either provide input sequences in FASTA format or specify a file containing the sequences. The script executed the selected alignment algorithm and saved the aligned sequences to an output FASTA file. Sample inputs and outputs were tested, showing the script's ability to perform pairwise alignment accurately and generate aligned sequences ready for further analysis.

3) Multiple Sequence Alignment: The multiple sequence alignment component of the automation script allows users to choose between two popular MSA tools, Clustal Omega and MAFFT. Users could input sequences either through a file or directly via the graphical user interface (GUI). The script applied the selected MSA algorithm to the sequences and saved the aligned sequences to the specified output file. The script's performance was tested using sample sequences, and the resulting alignments were visually inspected to confirm the correctness of the MSA process.

4) Phylogenetic Tree: [6] The phylogenetic tree construction component of the automation script uses aligned sequences as input to generate phylogenetic trees. Users could provide the aligned sequence file, and the script ran a selected tree construction algorithm (e.g., Maximum Likelihood, Bayesian Inference, or Distance-based Methods). The constructed tree was either displayed visually in the GUI or shown in the console, depending on the user's preference. The accuracy and reliability of the constructed trees were evaluated by comparing them with known phylogenetic relationships.

5) Script Usability and Modularity: The developed automation script was designed to be user-friendly and well-documented, providing clear instructions and input options for each component. The modularity of the script allowed easy customization and extension for future modifications and incorporation of additional bioinformatics tasks. Feedback from users who tested the script highlighted its ease of use and potential to save time and effort in performing bioinformatics analyses.

IV. conclusion

In conclusion, the Bio-informatics Automation Script toolbox [7] successfully achieved its objective of developing a user-friendly bash script that integrates various bio-informatics tasks, including database retrieval, sequence alignment, MSA, and phylogenetic tree construction. The script provides a flexible and modular tool that allows users to retrieve relevant sequences or information from a selected bio-informatics database, align sequences using a suitable alignment tool, perform multiple sequence alignment using an MSA tool, and generate a phylogenetic tree based on the aligned sequences using a phylogenetic tree construction tool. The script has been tested with sample inputs, and its functionality has been validated across different bio-informatics databases, alignment tools, and phylogenetic tree construction methods. The script's design and implementation have been well-documented, making it easy to comprehend and modify in the future. Overall, the Bioinformatics Automation Script project provides an excellent example of how scripting can be used to automate complex bioinformatics tasks, saving time and effort for researchers.

V. Future Work

The automation script currently supports database retrieval from NCBI, but it could be extended to support other databases, such as UniProt or Ensembl. Also, it could be extended to support other tree construction methods, such as maximum parsimony or maximum likelihood, and provide options for customizing the tree construction parameters. we also could optimize it to run in parallel on multiple processors or computing nodes, to speed up the processing of large datasets.

References

[1] Kans, J. (2023, June 30). Entrez Direct: e-Utilities on the Unix Command line. Entrez Programming Utilities Help - NCBI Bookshelf.

[2] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). ClustalW: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Research, 22(22), 4673-4680. doi:10.1093/nar/22.22.4673

[3] Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. Nucleic Acids Research, 30(14), 3059-3066. doi:10.1093/nar/gkf436

[4] https://itol.embl.de/

[5] https://itol.embl.de/upload.cgi

[6] https://github.com/TawfikYasser/bio23/tree/main/PhylogeneticTree

[7] https://github.com/TawfikYasser/bio23/tree/main