COSC480

Project: Sequence analyzer

Diana Castano

Introduction

The Sequence Analyzer project is designed to analyze DNA and protein sequences from FASTA and FNA files.

DNA, short for deoxyribonucleic acid, is like a blueprint for life. It contains the instructions needed to build and maintain living organisms from bacteria and plants to humans. Think of it as a set of instructions that tell the cells what to do and how to do it. DNA is made up of smaller units called nucleotides, which are like the letters of an alphabet. These nucleotides form a long, twisted ladder-like structure called a double helix. The 4 nucleotides are A, T, G, C. These 4 letters combine in different orders and different numbers giving light to the different shapes of life.

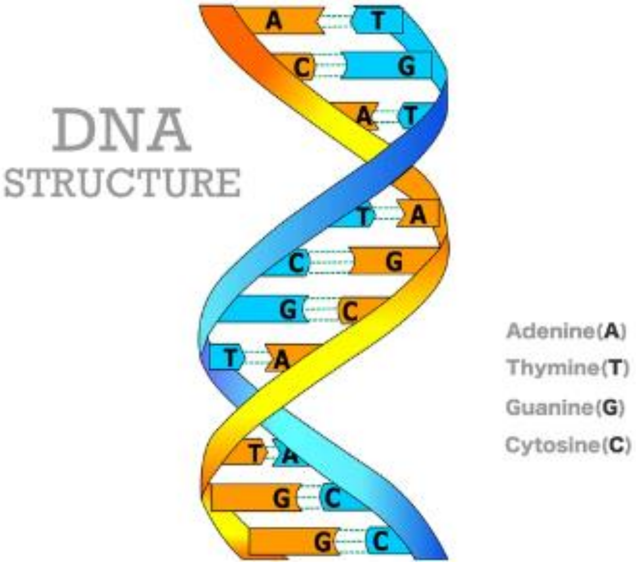

Fig1. Represents a DNA fragment. The DNA is a long string composed of 4 letters A,T,G,C which repeat along the sequence and perform multiple combinations (Shutterstock, n.d.).

DNA is translated into amino acids which are the building blocks of the proteins that carry out the instructions written in the DNA. Proteins perform a wide range of tasks in our bodies, from building and repairing tissues to helping our immune system fight off infections.

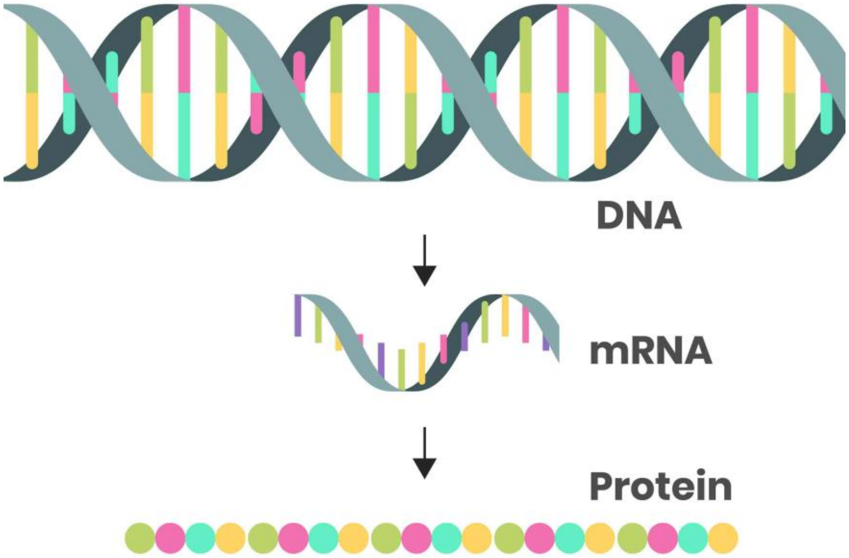

Fig 2. According with the order and the combinations of letters in the DNA. It will be translated into different Amino acids which are the building blocks of proteins (The Conversation, 2021).

If we were to draw an analogy between DNA and proteins, often considered the "program of life," and a Python program, we might liken DNA to the foundational elements of Python such as if statements, loops, and variables, etc. These foundational elements can then be translated into functions, akin to proteins, which govern various functionalities within the program.

The program is designed for analyzing DNA and protein sequences using a graphical user interface (GUI). Users can select sequence files, analyze them, and view the results in a user-friendly manner. This tool is particularly useful for researchers and biologists who need to process and visualize sequence data.

How to use the program

1. Launching the Program:
   - Ensure you have the necessary libraries installed in python (tkinter, matplotlib, numpy, biopython). “Biopython is a collection of freely available Python modules for computational molecular biology” (Biopython contributors, 2024).
   - Run the program file. The GUI window will open.
2. Selecting Files:
   - Click the "Select Files" button.
   - A file dialog will appear. Select the DNA (e.g., .fna) or protein sequence files (e.g., .fasta) you want to analyze. You can select multiple files at once.
3. Analyzing Sequences and viewing the results:
   - After selecting the files, click the "Analyze Sequences" button.
   - The program will process the files and display the results in the text area and plot.
   - For DNA sequences, the program will count and display nucleotide occurrences and plot a bar chart.
   - For protein sequences, the program will calculate and display amino acid composition and plot the results.

The development process

The development process followed several key stages:

1. Requirements Gathering:
   - Determined the need for a user-friendly tool for analyzing DNA and protein sequences.
   - Identified core functionalities: file selection, sequence reading, nucleotide/amino acid analysis, and result visualization.

2. Design:

- Planned the sequence reading and analysis functions using Biopython for sequence manipulation and Matplotlib for visualization.

3. Environment set up:

- Installed the libraries matplotlib for plotting, numpy and biopython for sequence processing.
- Gathered the files: the genomes and protein files were obtained from Bacterial And Viral Bioinformatics Resource Center (Bacterial and Viral Bioinformatics Resource Center, n.d.) and The National Center for Biotechnology Information (National Center for Biotechnology Information, n.d.) respectively.

4. Implementation of DNA and Protein Analysis functions in the following order:

DNA Sequence Analysis Functions:

- Reading Sequences from File**:** The function reads sequences from a given file, removing unwanted lines/identifiers, and returns a string of concatenated sequences.
- Nucleotide Counting: The function counts the occurrences of each nucleotide (A, T, G, C, N) in a DNA sequence and returns a dictionary with the count of each nucleotide.
- Plotting Nucleotide Counts: The function plots the count of each nucleotide in a DNA sequence.

Protein sequence analysis functions:

- Reading Protein Sequences from File: The function reads protein sequences from a FASTA file and returns a list of protein sequences.
- Calculating Sequence Lengths: The function calculates the lengths of protein sequences and returns a list of lengths corresponding to each sequence.
- Analyzing Amino Acid Composition: The function analyzes the amino acid composition of protein sequences and returns a list of dictionaries with the count of each amino acid for each sequence.

- Plotting sequence lengths: The function is designed to visualize the distribution of sequence lengths. This is useful in bioinformatics to understand the range and frequency of different sequence lengths within a dataset.
- Analyzing Protein Sequence: The function analyzes the given protein sequence and returns a dictionary containing the count of each amino acid.
- Plotting Amino Acid Composition: The function plots the amino acid composition using colors.

Main Functions: two mains were created.

- Main for DNA Sequence Analysis: The main function reads multiple sequence files, counts nucleotides in each, prints the results and plots the nucleotide counts.
- Main for Protein Sequence Analysis: The main_protein function performs analysis of protein sequences. It calculates sequence lengths, analyzes amino acid composition, and plots the results.

5. Testing:

   - Each function was being tested and modified in every step.
   - Conducted tests with various DNA and protein sequence files to ensure accurate analysis.

6. Refactoring and Graphic User Interface integration:

   - Interface Design: A tkinter GUI was designed with buttons for file selection and sequence analysis, and a text area for displaying results.
   - Refactoring and Integration: The sequence analysis functions, and the main functions were integrated into the GUI, enabling user interaction.

7. Final testing:

   - Tested the GUI and overall functions for usability and responsiveness.

Things that went well

- **Functionality:** The core functionality of reading sequences, analyzing data, and displaying results worked as expected.

- **User Interface:** The GUI successfully compiled and integrated the functions as expected.
- **Visualization:** The use of matplotlib for plotting nucleotide counts and amino acid compositions added a valuable visual component to the analysis.

Challenges faced

- **File Handling:** Ensuring the program correctly reads and processes different formats of FASTA files required careful handling and a lot of trial.
- **Color Coding:** Assigning distinct and meaningful colors to amino acids for visualization involved some trial and error until decided to include the color=colors argument.
- **Biopython Library:** The use of this library was challenging because it implied learning to implement new things.
- **Implementation of Blast algorithm:** It was intended to implement blast algorithm for comparing sequences, however it was challenging to get it running in the given time frame.

Future developments/ Work that can be done

- **Advance Analysis:** Implementation of more advanced sequence analysis methods, such as motif finding and 3D protein structure prediction.
- **Implementation of blast algorithm:** for performing multiple sequence mapping, retrieval and alignments.
- **Improve GUI:** Enhancing the GUI with more features, such as saving analysis results to a file in different formats and more interactive plots.

References

- Shutterstock. (n.d.). Dna helix molecule blue yellow colors royalty-free images, stock photos & pictures. Retrieved June 6, 2024, from https://www.shutterstock.com/search/dna-helix-molecule-blue-yellow-colors
- The Conversation. (2021). *file-20210408-15-1pshkcj.jpg (2262×1773)* [Image]. The Conversation. Retrieved June 6, 2024, from file-20210408-15-1pshkcj.jpg (2262×1773) (theconversation.com)
- Bacterial and Viral Bioinformatics Resource Center. (n.d.). Retrieved May7, 2024, from https://www.bv-brc.org/
- National Center for Biotechnology Information. (n.d.). Retrieved June 8, 2024, from https://www.ncbi.nlm.nih.gov/
- Biopython contributors. (2024). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Retrieved from https://biopython.org/