

Bioinformatics and Computational Biology Lab

Course code: CSE 430

Project Title: Liver Cancer

Project Documentation

**Submitted To: -**

Jobair Ahammed Tusher

Lecturer

Department of CSE, UITS

**Submitted By: -**

Name: Al Shahriar Refath

ID: 2114951020

SUBMISSION DATE: 06-01-2025

**Introduction :**

Liver cancer is one of the most common cancer diseases in the world and has a complicated genetic and environmental background. If genetic mutations associated with liver cancer are identified, then early diagnosis and cure is easily possible. This project uses molecular biology and computational biology methods to assay DNA sequences and search for a signal of hepatocarcinogenesis. Through this project, the value of sequence alignment for cancer genomics can be seen by using computational approaches.

**Libraries:**

The following Python libraries are used in the project:

**Bio:** Includes a set of modules performing the global alignment based on the Needleman-Wunsch algorithm and the local alignment based on the Smith-Waterman algorithm required for DNA sequences comparison.

**difflib:** It can compute and expresses the degree of genetic variation in the form of sequence similarity percentages.

**os:** Clarification It does file I/O which includes loading the reference genome.

**Problem Statement:**

Identification of liver cancer associated genetic variations requires precise computational methods. This project addresses the following challenges:

* In what manner can computation techniques identify mutations associated with liver cancer
* Looking at the case of DNA sequence variations, what is the significance of the global and local sequence alignments
* This paper explains how similarity thresholds can actually decide important genetic variations from a reference genome

**Methodology:**

The project incorporates a highly-defined biological analysis of genetic data linked to liver cancer. The methodology consists of the following steps:

**Reference Genome Loading:**

* There is also a FASTA file associated with liver cancer that is called liver.fna.
* The script opens the file to read, we throw the heading and store the DNA sequence into a string which will be analysed.

**User Input DNA Sequence:**

* This script asks the user to enter a DNA sequence that wants to be analyzed.
* The input sequence is validated to ensure it contains only valid DNA bases: A, T, G, and C.

**Sequence Similarity Calculation:**

For the percentage similarity of the reference and user DNA sequences, the difflib.SequenceMatcher introduces the sequence matcher percentage.

A threshold of 90% similarity is used to classify the sequence:

**Infected:** If dissimilarity is below threshold stating that there is significant variation.

**Not Infected:** However, if similarity which is computed into a percentage reaches or exceeds the defined threshold, HOX can confidently say that the sought match in the sample space of genome scale is very similar to the reference genome.

**Sequence Alignment:**

**Global Alignment:**

* Cohesion Analysis done using the Needleman-Wunsch algorithm taken from (pairwise2.align.globalxx).
* Selects the number of matching values that defines the best overall match with the given user sequences.
* Gives an alignment score in order to judge the quality of the matching.

**Local Alignment:**

* It was performed using the Smith-Waterman algorithm under the (pairwise2.align.localxx).
* More tailored to identification of regions within the sequences that are the most conserved.
* Recovers an alignment score and generates graphical representation of the conserved sequences.

**Result Interpretation:**

The script displays:

* Consequently, similarity percentage has something to do with the infection status.
* Global alignment almost results, including the best alignment and its score.
* The local alignment outcome with the options of conserved regions, and the score of alignment.

**Error Handling:**

If the reference genome file, namely liver.fna is not found, the script outputs this message and stops the script.

Any DNA sequences containing characters other than ‘A’, ‘C’, ‘G’ and ‘T’ are eliminated and the user is asked to input a correct sequence.

**Output Visualization:**

The alignments and scores are then displayed in a format that can be understood by the human interest, and this make the result of the analysis easy to understand.

Thus, this methodology makes it possible to study DNA sequences for genetic mutations linked to liver cancer comprehensively.

**Result & Analysis:**

The script performs the following analysis:

**Reference Genome Loading:** Reads a reference genome for liver cancer from a file..

**User Sequence Input:** Prompts the user to input DNA sequence that will be used in a comparison.

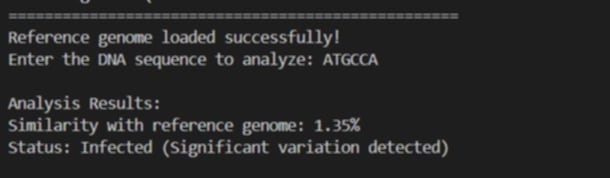
**Sequence Validation:** Input data should contain only valid bases of a DNA sequence, that is, A, T, G or C.

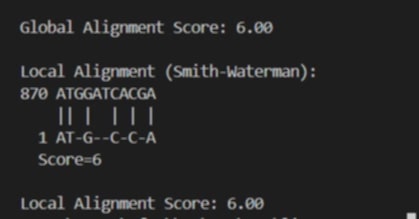
Comparison and Alignment:

**Similarity Calculation:** Calculates the percent representativeness of the user sequence to the reference one.

**Global Alignment:** Employing the Needleman–Wunsch algorithm it determines the best overall alignment and its score.

**Local Alignment:** Uses SW-align for sequence alignment to conserved region and the score obtained.





**Conclusion:**

This project shows how bioinformatics can be used in examining genetic mutations linked to liver cancer. In a manner, the project brings out aspects such features of similarity calculation as well as sequence alignments that provide plausible genetic mutation that is related to liver cancer. The strategies of using the broad and narrow alignment tools when analyzing genomic features demonstrated the scientific potential of the approach, and will allow for deeper investigation of the genetic markers in the context of cancer research in the future.