

Bioinformatics and Computational Biology Lab

Course code: CSE 430

Project Title: Lung Cancer

Project Documentation

**Submitted To: -**

Jobair Ahammed Tusher

Lecturer

Department of CSE, UITS

**Submitted By: -**

Name: Jannatul Eshrat Sinthia

ID: 2114951011

SUBMISSION DATE: 06-01-2025

**Introduction :**

Cancer of the lungs is regarded among the leading causes of death due to tumors in the entire globe. The follow-up here is that early identification, together with early and accurate diagnostic measurements, play a huge role in improving the overall survival rates. It is possible to conduct sequence analyzes in bioinformatics and it may assist in identifying genetics associated with lung cancer. This project mainly focuses on the determination of the shifts of human lung cancer-associated polyomavirus through DNA sequences alignment.

**Libraries:**

The following libraries and tools are utilized in this project:

**os:**To process the file paths and omitted the checking of the input files.

**difflib:**Contains procedures to calculate alignments and similarities between two sequences using the SequenceMatcher.

**Biopython:**Among them, the pairwise2 module is for Global and Local alignments with Needleman-Wunsch as well as Smith-Waterman algorithms.

**Problem Statement:**

The aim of this project is to design an efficient bioinformatics pipeline that aligns the sequences uploaded by a user to the reference genome of lung cancer-associated polyomavirus. The objectives are:

* To determine the level of similarity between the users DNA sequence and the reference genome.
* To decide whether the difference is remarkably significant in order to determine whether one is infected or whether there has been a mutation.
* For using BLAST to get the value for identity comparison, sequence alignment score and to carry out both global and local alignment.

**Methodology:**

The methodology for the project "Lung Cancer" is outlined in the following steps:

**Data Preparation:**

Reference Genome:

* A file with sequence data for the lung cancer–associated polyomavirus is read in from the file cancer\_lung.fna and used as the reference sequence.
* First, the sequence is cleaned, by eliminating any header lines, and all sequence lines are merged into a single string.
* User DNA Sequence:
* A DNA sequence given by the user is entered for analysis.
* The sequence is checked to make sure they do not include any other character than A,T,G and C; return true if the check passes.

**Similarity Analysis**

Sequence Similarity:

* The difflib.SequenceMatcher library is used to find out the percentage match between the given DNA sequence and the reference genome list.
* To segregate normal control from potentially infected or mutated, the match percentage is set at the 90 percent.

**Sequence Alignment**

Global Alignment:

* Specifically, we calculate the global alignment of the two sequences using the Needleman-Wunsch algorithm through Biopython package, pairwise2.align.globalxx.
* This technique is taken to arrange sequences right from end to end to measure overall similarity.

Local Alignment:

* Biopyton’s pairwise2.align.localxx Smith-Waterman aligning computes the local alignments that identify domain-typical regions of maximum similarity.
* It is also particularly useful if the pattern of variation or mutation is localised.

**Decision Making**

Infection Status:

* Depending on the similarity ratio, the tool infers whether the user sequence contains fairly high degree of genetic divergence (possibly due to infection or mutation).
* If the similarity is less than 90%, the sequence is more marked as “Infected.”

Alignment Scoring:

* The alignment scores from both the global as well as the local alignment methods are utilized for more detailed study of the quality of match in the sequences.

**Output Results**

The tool provides:

* Similarity percentage.
* Infection status to include the type of status later in the study distinct between individuals being infected and those not infected at all.
* The tools producing specific alignment report containing alignment visualization and both global and local alignment scores.

**Result & Analysis:**

The analysis is performed through the following steps:

**Input Validation:** The user's DNA sequence is validated to ensure it contains only valid nucleotides (A, T, G, C).

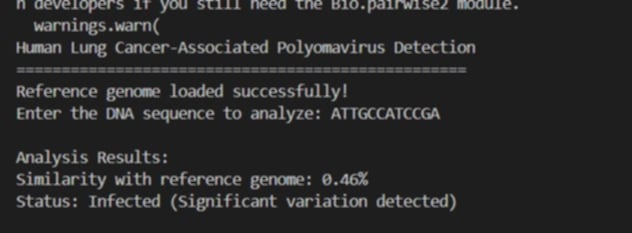
**Reference Genome Loading:** The reference genome is loaded from a file, ignoring any header information.

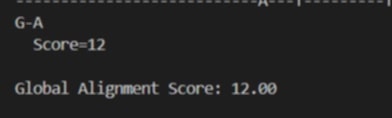
**Similarity Calculation:** The similarity percentage between the reference and user-provided sequences is calculated using the difflib SequenceMatcher.

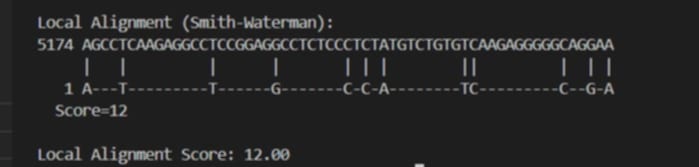
**Alignment Analysis:**

* Including the comparison of the whole strings with the Needleman-Wunsch algorithm.ength of both sequences.
* By applying the Smith-Waterman algorithm, and thereby to detect exact matches and work out the best bit score for comparison of these sequences. within the sequences.

**Infection Status Determination:** Similarity threshold used is set at 90% to separate the user as one that is most likely infected from the unaffected one.

****

****

****

**Findings:**

* The degree of identity of the sequences compared to the reference genome is used to define to the extent there are significant genetic differences.
* Match quality information in the global and local alignments scores used in the evaluation give better understanding.

**Conclusion:**

This project can be used to show how bioinformatics can be used to find suspected lung cancer related genomic difference. Using both, the sequence similarity check and the alignment, the tool can be effectively used to analyze DNA sequences. It raises awareness of the practical value of computational applications in progressive cancer investigation and diagnostic techniques.