NanoAlign- Nanopore Protocol Alignment Toolkit

Functional Specification

Background

Nanopore sequencing is predominantly employed for long-read sequencing tasks. However, its application for short-read sequencing, particularly for tuberculosis (TB) sequences ranging from 40 to 100 bases, remains under-explored. Optimizing the process requires specialized probe designs for PCR amplification and rigorous validation against known reference sequences. There is an unmet need for a computational tool to facilitate these optimization steps.

User Profile

- Lab Members in a Bioengineering Research Lab
- Domain Knowledge: Expertise in bioengineering, molecular biology, and DNA sequencing. Focused on global health solutions such as point-of-care diagnosis for HIV and TB.
- Computing Skills: Basic knowledge in nanopore sequencing and bioinformatics. Require straightforward and user-friendly outputs.

Use Cases

1. Primer Design

- Objective: Design primers for PCR amplification from given DNA sequences.
- Interactions: Users input DNA sequences, and the system outputs optimized primer sequences.

2. Sequence Alignment

- **Objective:** Align nanopore-sequenced fragments against known reference sequences and set up necessary software environments. The output file can be used in Geneious Prime to check detailed alignment information
- **Interactions:** Users provide a directory containing a sample sheet and subfolders with FASTQ files. The system aligns sequences and manages software dependencies like Java, Docker, and Nextflow.

Component Specification

1. Software Components

a. Primer Design Module (primer_design.py):

- **Function:** Generates optimized primer sequences from input DNA sequences in FASTA.
- **Input:** DNA sequences in FASTA format.
- **Output:** List of primer sequences.

b. Sequence Alignment Module (main_Parsed.py):

- **Function:** Aligns sequencing data to a reference genome and sets up environmental dependencies.
- **Input:** Directory containing a sample sheet and subfolders with FASTQ files.
- **Output:** Alignment results in bam files.

2. Interactions to Accomplish Use Cases

Primer Design:

c. Read DNA sequences from a FASTA file.
d. Generate and output primer sequences.

Sequence Alignment:

a. Install and set up Java, Docker, and Nextflow.
b. Read sequencing data and align it to a reference genome using Nextflow workflows.

3. Project Plan

1. 11.13.2023 – 11.19.2023:
    a. Develop and test **the primer design** with algorithms capable of producing optimized primer and identifying dimerization.
2. 11.21.2023 – 11.29.2023
    a. Develop and test **alignment** accuracy against known reference genomes.
    b. Finalize the data analysis module to **compare the sequences against reference data**.
3. 11.30.2023 – 12.09.2023
    a. Document the pipeline process and provide training for lab members.
    b. Package the pipeline

4. Additional Notes:

The “main_Parsed.py” script's functionality is dependent on external tools (Java, Docker, Nextflow), and thus, it is crucial to ensure these are installed and configured correctly.