BIOEN 537 – Computational System Biology
Hanwen Gu *Assignment: First draft of functional Spec* *Due: Nov 8th 2023*

NanoOpt- Nanopore Protocol Optimization Toolkit

Component Specification

1. Software Components

a. Data Manager:

- **Function:** Manages and processes Fasta & Excel sequence data.
- **Input:** FASTA file containing target DNA/RNA sequences, reference sequences, and barcode sequences.
- **Output:** preprocessed data for use in subsequent components (FASTA, FASTQ, BAM); formatted Excel files for output data like optimized probe sequences and validation results, assign reads to the correct FASTA File.

b. Probe Design Optimizer

- **Function:** Generates optimized probe sequences for PCR amplification and nanopore sequencing, while minimizing dimerization risks.
- **Inputs:** Preprocessed sequence data and probe design constraints from the Data Manager.
- **Outputs:** Optimized probe sequences and reports on dimerization risks.

c. Quality Control:

- **Function:** Assesses the quality of raw nanopore sequencing data.
- **Input:** Raw electrical signal data from nanopore sequencing.
- **Output:** Quality metrics for the raw data and flagged data for further processing or exclusion.

d. **Basecalling (including read trimming and/or Quality Assessment):**

   a. **Function**: Converts raw electrical signals from nanopore sequencing into trimmed nucleotide sequences
   b. **Input:** High-quality raw signal data, as assessed by the Quality Control module.
   c. **Output:** Basecalled nucleotide sequences ready for further analysis.

e. Sequence Validation Engine

- **Function:** Analyzes and validates nanopore sequencing data from raw electrical signal to final sequence accuracy assessment.
- **Components:**
    - **Read Mapping/Alignment**: Aligns reads to a reference genome.
    - **Variant Calling**: Identifies variations from the reference sequence.
    - **Data Analysis:** Compares the final, cleaned reads against known reference sequences to validate accuracy.
- **Inputs:** Raw electrical signal data, sample sheet with reference sequences and barcode sequences.

- **Outputs:** Accuracy metrics, mismatch identification

2. Interactions to Accomplish Use Cases

For Use Case 1: Universal Probe Design

- The Data Manager receives user-inputted target DNA/RNA sequences and constraints in Excel format. It validates and preprocesses this data.
- The preprocessed data is passed to the Probe Design Optimizer, which generates optimized probe sequences and identifies potential dimerization issues.
- The Probe Design Optimizer outputs these optimized probe sequences and any dimerization flags back to the user through the Data Manager, which may present this information in an Excel file or through a simplified graphical interface.

For Use Case 2: Sequence Validation

- The Quality Control Module processes raw electrical signal data using automated pipelines with tools like *NanoPlot* or *PycoQC*.
- High-quality signal data is fed into the Basecalling Module where software like *Guppy* translates it into nucleotide sequences.
- The Read Trimming Module uses tools like *Porechop* to trim adapters and low-quality bases from the sequences.
- Post-trimming, the Quality Assessment Module reassesses the read quality to ensure high fidelity for the following steps.
- The Read Mapping/Alignment Module aligns the high-quality reads to a reference genome, using tools like *Minimap2*.
- If variant analysis is needed, the Variant Calling Module identifies discrepancies using software like *Medaka*.
- Finally, the Data Analysis Module compiles the results, comparing the processed sequences against known references to generate accuracy metrics.

3. Project Plan

1. 11.13.2023 – 11.19.2023:
    a. Develop and test **the Probe Design Optimizer** with algorithms capable of producing optimized probes and identifying dimerization.
    b. Set up the automated pipeline for **Quality Control** with the selected tools.
    c. Integrate basecalling, read trimming, and quality assessment modules.
2. 11.21.2023 – 11.29.2023
    a. Test **read mapping/alignment** accuracy against known reference genomes.
    b. If necessary, include a **variant calling** step and verify its accuracy in identifying known variants.

   c. 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. Launch the sequence validation engine and collect feedback.