NanoOpt- Nanopore Protocol Optimization Toolkit

Component Specification

1. Software Components

a. Probe Design Optimizer

- **Function:** Optimizes probe sequences for PCR amplification and nanopore sequencing, ensuring specificity and avoiding dimerization.
- **Inputs:** Preprocessed sequence data, probe design constraints.
- **Outputs:** Optimized probe sequences and reports on dimerization risks.

b. Sequence Validation Engine

- **Function:** Analyzes and validates nanopore sequencing data from raw electrical signal to final sequence accuracy assessment.
- **Components:**
  - Quality Control Module: Assesses the quality of the raw signal data.
  - Basecalling Module: Converts raw signals into nucleotide sequences.
  - Read Trimming Module: Removes adapters and low-quality sequences from the reads.
  - Quality Assessment Module: Evaluates the quality of basecalled reads.
  - Read Mapping/Alignment Module: Aligns reads to a reference genome or assembles them de novo.
  - Variant Calling Module: Identifies variations from the reference sequence.
  - Data Analysis Module: Compares the final, cleaned reads against known reference sequences to validate accuracy.
- **Inputs:** Raw electrical signal data from nanopore sequencing, known reference sequences.
- **Outputs:** Accuracy metrics, mismatch identification, and a detailed validation report.

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. Preliminary Plan

1. Develop and test **the Probe Design Optimizer** with algorithms capable of producing optimized probes and identifying dimerization.
2. Set up the automated pipeline for **Quality Control** with the selected tools.
3. Integrate **basecalling** pipeline, ensuring compatibility with raw signal data.
4. Develop or integrate a read **trimming** module, verifying its efficacy in removing undesired sequences.
5. Implement the **post-trimming quality assessment** to validate read integrity after trimming.
6. Incorporate the **read mapping/alignment** pipeline, testing alignment accuracy against known reference genomes.
7. If necessary, include a **variant calling** step and verify its accuracy in identifying known variants.
8. Finalize the data analysis module to **compare the sequences against reference data**.
9. Document the pipeline process and provide training for lab members.
10. Launch the sequence validation engine for broader use within the research team, collecting feedback for continuous improvement.