SCOLiA

Subtyping and Comparison of LIsteria Assemblies

Sonali Gupta
Allison Nicole Rozanski
Aparna Maddala
Ahish Melkote Sujay
Yuhua Zhang
Yiqiong Xiao

Introduction

Problem Statement

We would like to create a webserver that:

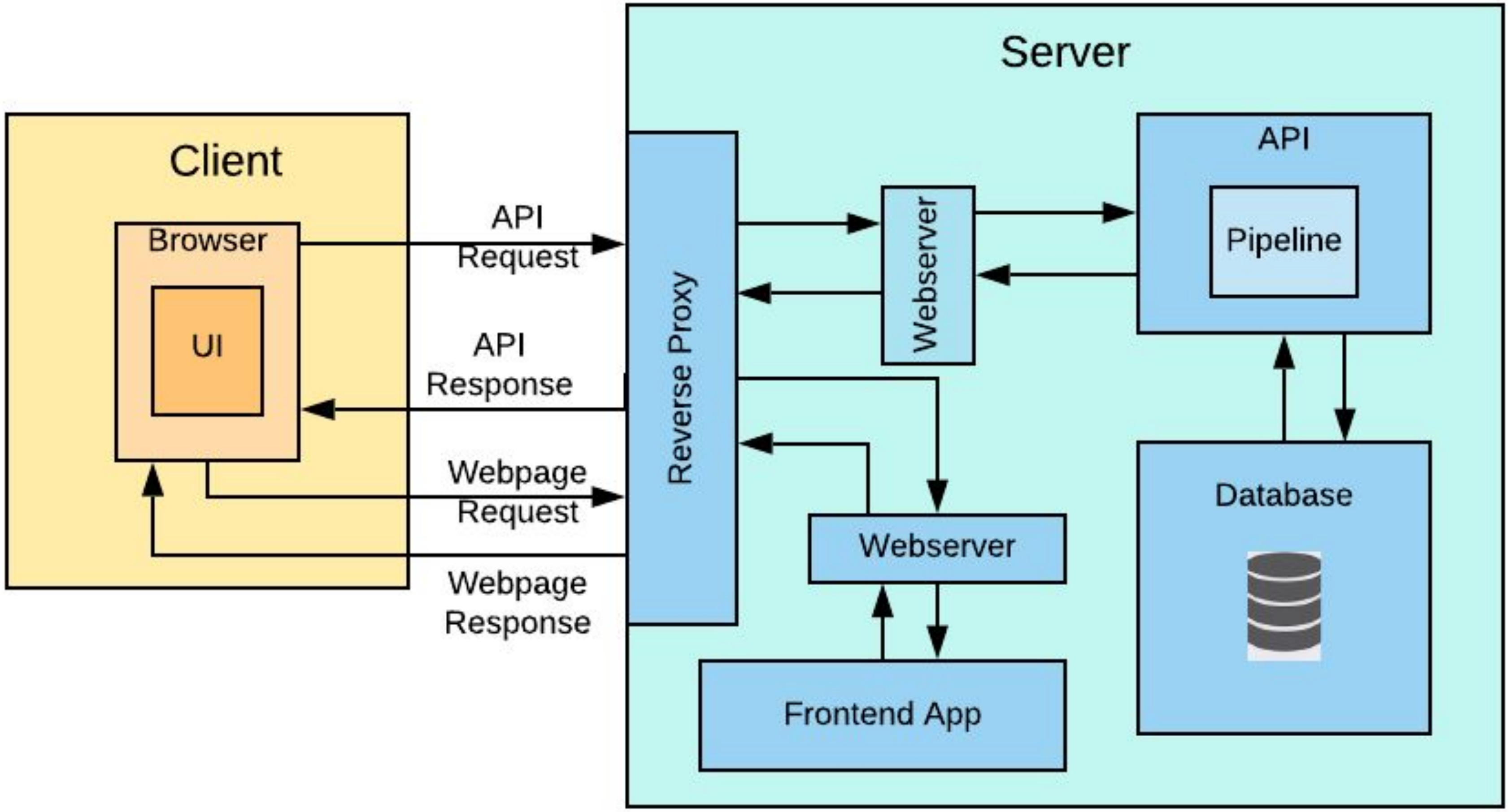
- Makes our work as bioinformaticians accessible to a wider audience
- Is easy-to-use, requires little bioinformatics knowledge to get the analysis results
- Is visually informative, easy on eyes

Introduction

Design Objectives

- Users should be able to
    - Go through the entire pipeline from genome assembly to comparative genomics
    - Only execute individual steps
    - Easily execute the remainder of the pipeline from any intermediate step

Architecture Design



MVC Framework - MODEL

MySQL

- Corresponds to all the data-related logic that the user works with
- Constitutes the computation, execution and visualisations
- Completely abstracted from the user

MVC Framework - View

Javascript, CSS, HTML

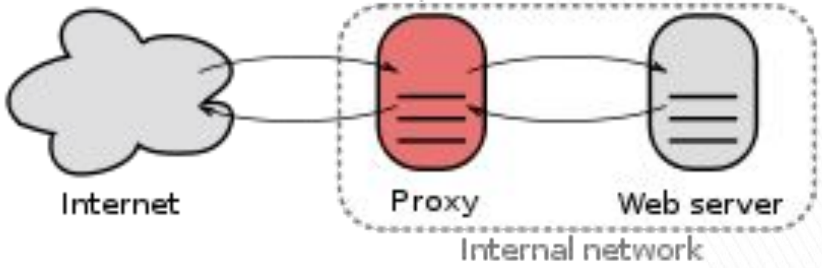
- Used for all the UI logic of the application
- Separates user from backend processing

MVC Framework - Controller

Python(Flask)

- Acts as an interface between Model and View components
- Responsible for validation of inputs from view and outputs from model before sending data to either of them
- Responsible for invocation of specific responses based on the requests received

Other Components



- Web server:

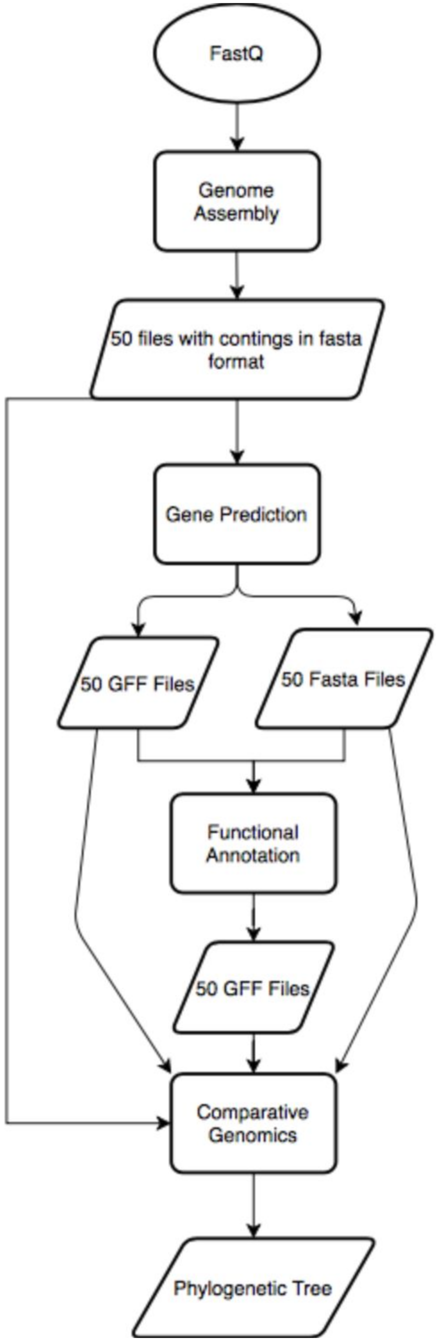
  Software that understands URLs (web addresses) and HTTP (the protocol your browser uses to view webpages).

  It can be accessed through the domain names of websites it stores, and delivers their content to the end-user's device.

- Reverse Proxy:

  Takes requests from the Internet and forwards them to servers in an internal network. Those making requests to the proxy may not be aware of the internal network.

Pipeline

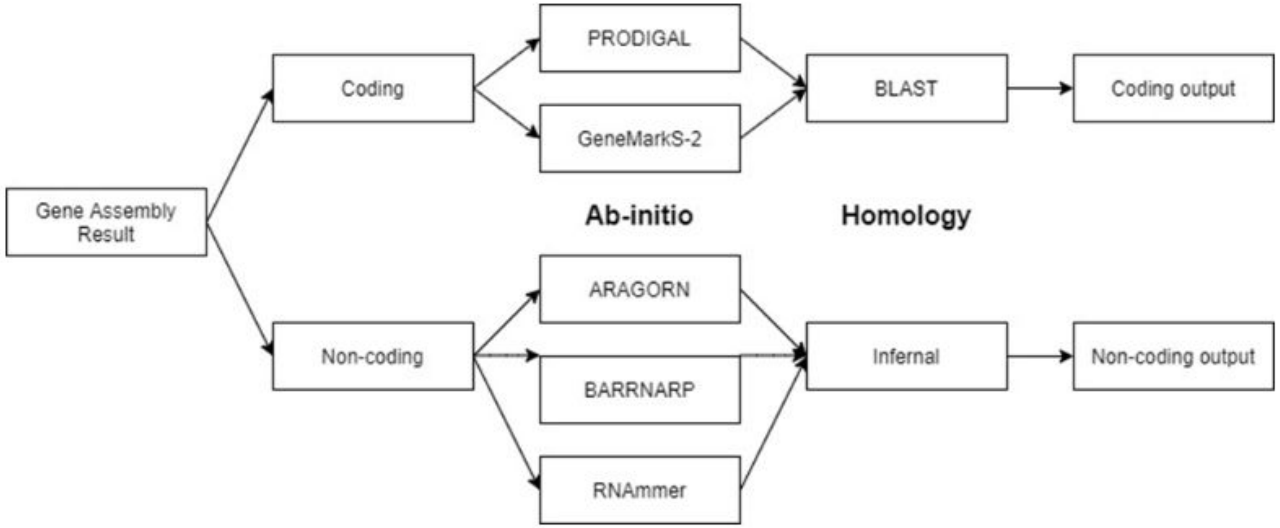

FastQ
Genome Assembly
50 files with contings in fasta format
Gene Prediction
50 GFF Files
50 Fasta Files
Functional Annotation
50 GFF Files
Comparative Genomics
Phylogenetic Tree

Genome Assembly

- Input files
  - Paired-end fastq files for Listeria monocytogenes
- Process
  - Perform quality control and trimming using fastp
  - Assemble genomes and plasmids using SPAdes
- Output files
  - HTML quality control report
    - Generated by MultiQC
  - FASTA contig files
    - Generated by SPAdes and Plasmid SPAdes

Gene Prediction

- Process
  - Coding
    - Run both PRODIGAL and GeneMarkS-2
    - Use BLAST for validation and retrieve the coding output
  - Non-coding
    - Run ARAGORN, BARRNARP, and RNAmmer
    - Use Infernal for validation and retrieve the non-coding output



- Input
  - Assembled genomes and plasmids from the genome assembly pipeline
- Output files
  - Coding: FASTA files, GFF files
  - Non-coding: FASTA files, GFF files

Functional Annotation

Input

- 50 files from gene prediction in .fna format

Processes

1. Cluster using UCLUST
2. eggNOG, CARD ,VFDB, PilerCR, SignalP and HMMTOP
3. merge functional annotations

Output

- 50 files in .gff and summary of annotation results i.e. annotation count for each tool, .gff display and significant antibiotic and resistance genes present

Comparative Genomics

fasta → Fast-ANI → Heatmap

fasta → StringMLST → Tree

fasta
reference.fasta → kSNP → ML/NJ tree

gff → BPGA → Tree

fasta → ARIBA → gff

- May not make all the tools available in the end
- Newick format

Flask

- One of the most widely used Python based web frameworks

- Reasons for selection:
  - Offers simple development
  - Easy to deploy
  - Fine-grained control
  - Provides flexible frameworks and is minimal
  - Familiarity with Flask

Flask

MySQL

- Flask does not support databases natively, gives flexibility on database choice best for application
- Flask-SQLAlchemy provides a flask friendly wrapper to the SQLAlchemy package.
- SQLAlchemy is an Object Relational Mapper (ORM) and supports several database engines including MySQL
- MySQL due to familiarity with server and installation
- Will Utilize MySQL for storing file paths and intermediate outputs

Task Delegation

BACKEND

Genome Assembly Pipeline: Aparna, Yiqiong
Gene Prediction Pipeline: Sonali, Ahish
Functional Annotation Pipeline: Allison Rozanski
Comparative Genomics Pipeline: Yuhua

FRONTEND

Aparna, Sonali, Yiqiong

Next Steps

- Backend development
- Webserver integration, configuration

Demo