Homework #4: Genome Assembly Comparison and Analysis

Course: Bioinformatics Computing (CS 494/594)
Due: October 17

Overview

In this assignment, you will explore *genome assembly comparisons* using MUMmer. You will compare (1) two *Bacillus anthracis* strains and (2) a sample assembly provided by the instructor against a reference genome. Your goals are to identify sequence-level differences, summarize assembly statistics, and interpret biological or technical implications to the best of your ability.

Part A. Comparing Two Anthrax Strains (20 points)

1. **Background:**
   We will reuse the two *B. anthracis* strains from Homework #3:
   - **Ames ancestor (NC_007530):** virulent strain
   - **Ames (NC_003997):** non-virulent laboratory strain
2. **Set up:**
   - Visit the MUMmer website.
   - Read about **nucmer**, **dnadiff**, and **show-snps** to understand alignment and SNP/indel detection workflows.
3. **Genome alignment:**
   - Install MUMmer (any version ≥ 2.x or 4.x)
   - Run **nucmer** using *Ames ancestor* as the reference and *Ames* as the query.
   - Submit the resulting `.delta` file. *(5 points)*
4. **Variant summary:**
   - Use **show-snps -C** to summarize SNPs and indels.
   - Submit your `.snps` file. *(5 points)*
5. **Interpretation:**
   - Write a short paragraph (~150–200 words) summarizing differences between strains:
     - How many SNPs/indels are there?
     - Any evidence of structural variation?
     - What biological differences might you hypothesize? *(10 points)*

Part B. Comparing a Test Assembly to a Reference (15 points)

1. **Setup:**
   - Download the test assembly (FASTA format) from the course website.
   - Download the corresponding reference genome.
2. **Assembly evaluation:**
   - Use **nucmer** to align the provided assembly to the reference.

    - Submit the resulting `.delta` file. *(5 points)*

3. **Assembly quality summary:**
    - Generate statistics using:
        - **show-coords** (alignment summary)
        - **show-snps** or **dnadiff** (optional, for variant assessment)
    - Include in your write-up:
        - Number of contigs
        - Total assembly size
        - N50 and average contig size
        - General comparison metrics (coverage, gaps, errors if relevant)
    - Evaluate: Does this look like a *good* assembly? Why or why not? *(10 points)*
4. EC for more curious students: +2 bonus points for mummerplot visualization of the draft assembly versus the reference

---

Deliverables Checklist

- `*.delta` files (2 total)
- `*.snps` file (from Part A)
- Written report (1–2 pages, combining both parts)
- Assembly summary table (recommended)