Lecture 4: Long-Read Assembly
Student Handout & In-Class Exercises

Course: BINF301 — Computational Biology
Instructor: Tom Michoel
Date: 26/01/2026
Created with Copilot

1 Overview

This handout summarizes the key concepts from **Lecture 4: Long-read Assembly**. Topics covered include:

- Why De Bruijn graphs are not ideal for long-read data
- Overlap-based long-read assembly strategies
- Overview of tools: **Canu, Flye, HiCanu, HiFiAsm**
- Haplotype phasing with long reads
- Scaffolding using Hi-C contact maps
- Assembly polishing and contamination detection

Solution. Lecture 4 emphasizes why DBG-based assembly depends on saturated k-mer coverage, which long-read datasets typically lack due to lower read count. Therefore, modern assemblers rely on overlap graphs.

2 Why Long Reads Break the De Bruijn Graph Assumption

De Bruijn graph assembly assumes most genomic k-mers appear multiple times in the read set. Long-read datasets (PacBio, Nanopore) have:

- **Lower read count** for same coverage
- **High noise** (for earlier long-read technologies)

Thus, k-mer coverage becomes non-uniform and DBG-based assembly becomes unreliable.

Solution. Long reads solve repeats via read length, so the assembly strategy should preserve long-range information, not break reads into small k-mers.

3 Noisy Long-read Data and Assembly Approaches

Early long reads had error rates of 10–20%, making overlap detection difficult.

Common strategies:

- **Hybrid correction:** use accurate short reads to correct long reads
- **Hierarchical correction:** repeated overlap–correct cycles (e.g., Canu)
- **Direct overlapping:** apply approximate matching (e.g., minimizers, MinHash)

Solution. Hybrid correction was popular during early Nanopore/PacBio development, but is less needed with modern HiFi reads.

4 Canu

Canu is a long-read assembler derived from the Celera assembler.

It operates in three phases:

1. **Correction** – detect overlaps, estimate corrected length, output corrected reads
2. **Trimming** – identify unsupported regions and remove them
3. **Assembly** – build the final overlap graph and output contigs

Canu uses the MHAP (MinHash Alignment Process) algorithm for fast overlap detection:

- Decompose reads into k-mers
- Hash k-mers and select *min-mers*
- Fraction of shared min-mers approximates sequence similarity

Solution. MHAP dramatically reduces the cost of all-vs-all comparisons, which is the primary bottleneck in OLC assemblers.

5 Flye

Flye uses a **repeat graph** rather than a classical overlap graph.

Key ideas:

- Build **disjointigs**: arbitrary merges of overlapping fragments
- Use disjointigs to form a draft assembly graph
- Distinguish:
  - **Bridged repeats** – some read spans the repeat
  - **Unbridged repeats** – resolved through subtle sequence differences
- Output final contigs after graph simplification

Solution. Flye's repeat graphs are robust for genomes with complex repeat structures, especially microbial and eukaryotic genomes.

6 HiFi Reads and Assemblers (HiCanu, HiFiAsm)

6.1 HiCanu

Optimized for high-accuracy PacBio HiFi reads.

Key features:

- Homopolymer compression before overlap detection
- Overlap-based trimming
- Error correction using read pileups

Solution. HiCanu produces extremely clean contigs because HiFi reads resolve many error modes during overlap correction.

6.2 HiFiAsm

A haplotype-aware assembler that:

- Performs all-vs-all overlaps
- Identifies **informative SNP positions**
- Groups reads into haplotypes using consistency checks
- Builds a string graph where haplotypes appear as "bubbles"

Solution. HiFiAsm is the current state-of-the-art for diploid HiFi genome assembly due to its haplotype resolution.

7 Haplotype Phasing

Diploid and polyploid organisms have heterozygous positions that create bubbles in assembly graphs.

Phasing strategies include:

- Read-based phasing (HiFiAsm)
- Trio-binning using parental data
- Hi-C based chromosome-scale phasing

Solution. Long reads often contain enough SNPs to directly assign them to haplotypes, making phasing far more accurate than short-read methods.

8 Scaffolding Using Hi-C

Hi-C provides chromosome-scale contact information:

- Contigs with strong Hi-C link density likely belong to the same chromosome
- Tools such as SALSA2 and YaHS construct chromosome-scale scaffolds
- Hi-C maps also reveal misassemblies (disruptions in the diagonal contact pattern)

Solution. A good Hi-C contact map shows a strong diagonal; off-diagonal blocks often indicate structural errors.

9 Polishing and Decontamination

After assembly:

- **Polishing:** tools such as `Pilon` improve base accuracy using mappings
- **Decontamination:** BlobTools, NCBI FCS detect foreign sequences
- **Organellar assembly:** tools such as MitoHiFi and OATK target mitochondrial/chloroplast genomes

Solution. Decontamination is vital for metagenome-derived samples and field-collected specimens.

10 In-Class Exercises

Exercise 1: MinHash Overlaps

A long read is decomposed into the following 6 k-mers:

$$\{AATCG, ATCGT, TCGTA, CGTAC, GTACG, TACGA\}.$$

The min-hash function selects the lexicographically smallest k-mer as the signature.

1. Compute the min-mer for the read.
2. Determine whether two reads overlap if they share the same min-mer.

Solution. The smallest k-mer is `AATCG`. Two reads sharing this min-mer are likely to overlap, but additional min-mers or sketches are typically needed for robust detection.

Exercise 2: Repeat Graph Reasoning

You observe two repeat copies in a Flye graph: one bridged, one not.

1. Explain how Flye resolves the bridged repeat.
2. Explain how Flye resolves the unbridged repeat.

Solution. Bridged repeats are resolved by direct read traversal. Unbridged repeats are resolved using small sequence differences and consistency between disjointigs.

Exercise 3: Hi-C Misassembly Detection

Given a Hi-C contact heatmap with a disrupted diagonal:

1. Interpret the meaning of an abrupt diagonal break.
2. Suggest a repair strategy.

Solution. A diagonal break indicates a misjoin between contigs. Tools like SALSA2 or YaHS can break and reassemble the problematic junction based on Hi-C support.

Exercise 4: Haplotype Bubbles

You see a bubble in a HiFiAsm graph representing two possible paths.

1. Identify what genomic feature this represents.
2. Describe a criterion to decide which path belongs to haplotype A vs B.

Solution. This represents heterozygosity. Reads supporting each variant (SNPs, small indels) determine the haplotype assignment.

Exercise 5: Homopolymer Compression

Explain why homopolymer compression improves overlap detection for HiFi reads.

Solution. Most HiFi read errors occur in homopolymer length estimation. Compressing runs (e.g., AAAA $\rightarrow$ A) removes this error source before alignment.