Project in Algorithms in Molecular Biology

Leena Salmela

March 13th, 2017
The genome assembly problem

\[
\begin{align*}
\text{ATTCTAGAGGAAATTACAAT} \\
\text{AAGTAAAGTATGATTTAGC} \\
\text{ATTAGCGAAAAACCCTCAATT} \\
\text{AGGAAATTACAATAAAGTAAA} \\
\text{TACAATAAAAGTAAAGTATGA} \\
\text{CGAAAAACCCTCAATTCTAG} \\
\text{AATTACAATAAAGTAAAGTATG}
\end{align*}
\]

\[
\downarrow
\]

\[
\begin{align*}
\text{ATTAGCGAAAAACCCTCAATT} & \quad \text{TACAATAAAAGTAAAGTATGA} \\
\text{AATTCTAGAGGAAATTACAAT} & \quad \text{AAGTAAAGTATGATTTAGC} \\
\text{AGGAAATTACAATAAAGTAAA} & \quad \text{TACAATAAAAGTAAAGTATGA} \\
\text{CGAAAAACCCTCAATTCTAG} & \quad \text{AATTACAATAAAGTAAAGTATG} \\
\text{ATTAGCGAAAAACCCTCAATTCTAG} & \quad \text{AATTACAATAAAGTAAAGTATGATTTAGC}
\end{align*}
\]
Genome assembly pipeline

- **Reads**
  - Error Correction
  - Contig Assembly:
    - join reads to form contigs
  - Scaffolding:
    - connect contigs with mate pairs
  - Gap Closing

- **Mate pairs**

- **(Error corrected)**

- **GTCA G-A**
  - **CGAGAAGT**
  - **AGAAGTC**
  - **GAAGTCA**

- **Leena Salmela**

- **Project in Algorithms in Molecular Biology**

- **March 13th, 2017**
Error correction

- Sequencing machines make reading errors
- Depending on technology, these can be mismatches, insertions, and/or deletions
- Genome assembly without sequencing errors would be simpler
- Exploit redundancy in sequencing to correct the errors
Contig assembly

- **Input**: Corrected reads
- **Output**: Longer contiguous sequences (=contigs) reconstructed from the reads
- **Approaches**:
  - Overlap-Layout-Consensus
  - Eulerian path
Scaffolding problem

- Input:
  - Set of contigs (contiguous sequences)
  - Set of mate pairs and their insert size

- Find a linear ordering of the contigs such that the number of mate pairs whose pairwise distance equals the insert size is maximized.
Gap closing

- Input: Scaffolds (=linearly ordered contigs) and reads
- Output: Scaffolds where gaps between contigs have been filled
Validation: How good is the assembly?

- How fragmented is the assembly?
- How well does the assembly reflect the used data?
- How complete is the assembly?
- Are there misassemblies?
N50: A measure for the length of contigs

- Order the contigs from shortest to longest
- Find the midpoint in terms of total sequence length
- The length of the contig in that point gives the N50 statistic of the set

\[ \text{N50} = 2400 \]

\[ \text{len} = 2400 \]

\[ \Rightarrow 50\% \text{ of the sequence is in contigs longer than or equal to N50.} \]
Project
Error correction

- *k*-mer spectrum methods
  - Quake:

- Multiple alignments based methods
  - Coral:

- For an overview of methods see:
Contig assembly

Most of these assemblers are pipelines performing several phases.

- De Bruijn graph based methods
  - Velvet
  - SOAPdenovo
  - IDBA-UD
  - SPAdes
- Overlap-layout consensus
  - SGA
Scaffolding

- SSPACE (a greedy method)

- BESST

- SCARPA
Gap closing

- Gap2Seq

- GapFiller
Useful libraries

- SeqAn (http://www.seqan.de/)
- GATB (http://gatb.inria.fr/)
- SDSL (https://github.com/simongog/sdsl-lite)