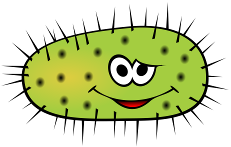
Project in Algorithms in Molecular Biology

Leena Salmela

May 4th, 2015

The genome assembly problem



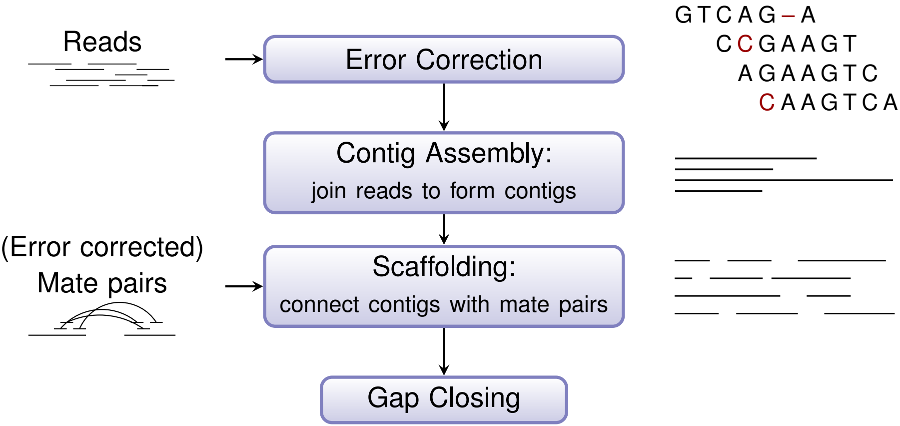
$\Longrightarrow$

AATTCTAGAGGAAATTACAAT
AAGTAAAGTATGATTTAGC
ATTAGCGAAAACCCTCAATT
AGGAAATTACAATAAAGTAAA
TACAATAAAGTAAAGTATGA
CGAAAACCCTCAATTCTAG
AATTACAATAAAGTAAAGTATG

$\Downarrow$

```
ATTAGCGAAAACCCTCAATT           TACAATAAAGTAAAGTATGA
                AATTCTAGAGGAAATTACAAT AAGTAAAGTATGATTTAGC
                        AGGAAATTACAATAAAGTAAA
     CGAAAACCCTCAATTCTAG    AATTACAATAAAGTAAAGTATG
---------------------------------------------------------
ATTAGCGAAAACCCTCAATTCTAGAGGAAATTACAATAAAGTAAAGTATGATTTAGC
```

Genome assembly pipeline



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 errrors

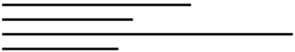
```
GTCAGAA–GTCGTGGTAACCCTTGATACCACCGGTTCA
--------------------------------------
GTCAGAA–GTCGTGGTAACCCTTGATA
  CAGAAAGTCGTGGTAACCCTTGATACC
  CAGAA–GTCGTGGTAACCCTTGATACC
        GTCGTGGTAACC–TTGATACCACCGG
            TGGTAACCCTTGATACCACCGGTTC
             GGTACCCCTTGATACCACCGGTTCA
```

Contig assembly

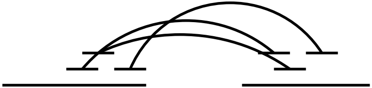
- **Input**: Corrected reads
- **Output**: Longer contiguous sequences (=contigs) reconstructed from the reads
- **Approaches**:
  - Overlap-Layout-Consensus
  - Eulerian path

Scaffolding problem

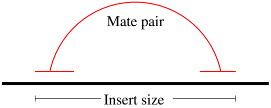
- Input:
  - Set of contigs (contiguous sequences)



  - Set of mate pairs and their insert size

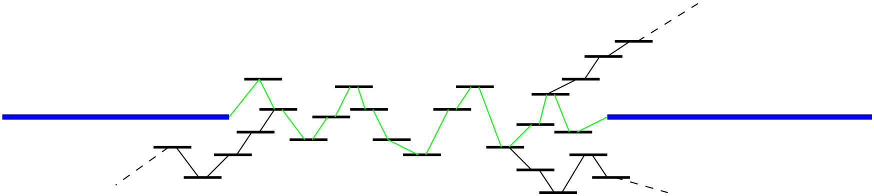


- Find a linear ordering of the contigs such that the number of mate pairs whose pairwise distance equals the insert size is maximized.



Gap closing

- Input: Scaffolds (=linearly ordered contigs) and reads
- Output: Scaffolds where gaps between contigs have been filled



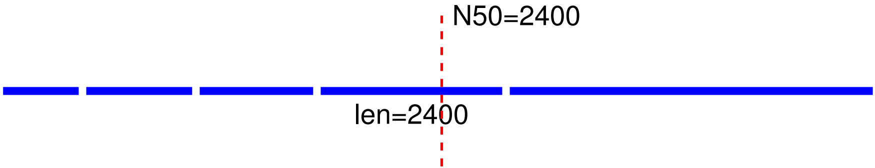
Validation: How good is the assembly?

- 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

$\Longrightarrow$ 50% of the sequence is in contigs longer than or equal to N50.



Project

Error correction

- *k*-mer spectrum methods
  - Quake:
    - Kelley et al.: Quake: quality-aware detection and correction of sequencing errors. Genome Biol 11:R116, 2010.
- Multiple alignments based methods
  - Coral:
    - Salmela and Schröder: Correcting errors in short reads by multiple alignments. Bioinformatics 27(11):1455–1461, 2011.
- For an overview of methods see:
  - Yang et al.: A survey of error-correction methods for next-generation sequencing. Briefings in Bioinformatics 14(1):56–66, 2012.

Contig assembly

Most of these assemblers are pipelines performing several phases.

- De Bruijn graph based methods
  - Velvet
    - Zerbino and Birney: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18:821-829, 2008
  - SOAPdenovo
    - Luo et al.: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1:18, 2012.
  - IDBA-UD
    - Peng et al.: IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28(11):1420–1428, 2012.
- Overlap-layout consensus
  - SGA
    - Simpson and Durbin: Efficient construction of an assembly string graph using the FM-index. Bioinformatics 26(12):i367-i373, 2010.

Scaffolding

- SSPACE (a greedy method)
  - Boetzer et al.: Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27(4):578–579, 2011.
- BESST
  - Sahlin et al.: BESST - Efficient scaffolding of large fragmented assemblies, BMC Bioinformatics 15:281, 2014.
- SCARPA
  - Donmex and Brudno: SCARPA: scaffolding reads with practical algorithms, Bioinformatics 29(4):428–434, 2013.

Gap closing

- Gap2Seq
  - Salmela et al.: Gap filling as exact path length problem. In Proc. RECOMB 2015, LNBI 9029, pp. 281–292, 2015.
- GapFiller
  - Nadalin et al.:GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics 13:S8, 2012.

Useful libraries

- SeqAn (http://www.seqan.de/)
- GATB (http://gatb.inria.fr/)