Metagenomic Assembly

Tutorial #5

Kleanthis Karavangelas
Karanveer Singh
Dan Drzewicki

Metagenomic Samples

- Snapshots of complex ecosystems at work
- Consist of hundreds of known and unknown species
- Frequently, DNA sequences of these species are the only clue to their evolution

DNA Sequencing

- The process of determining the order of nucleotides in DNA
- Shotgun metagenomic sequencing
  - Sequence random DNA strands from thousands of organisms in parallel
  - Size limit: 100 to 1,000 base pairs
  - Longer sequences subdivided into smaller fragments, called “reads”

Metagenomic Assembly

- Process of joining metagenomic reads into a single sequence
- Merges DNA fragments into the original DNA strands
- Assembly Software
  - SPADES
  - IDBA UD
  - Velvet

Book Analogy

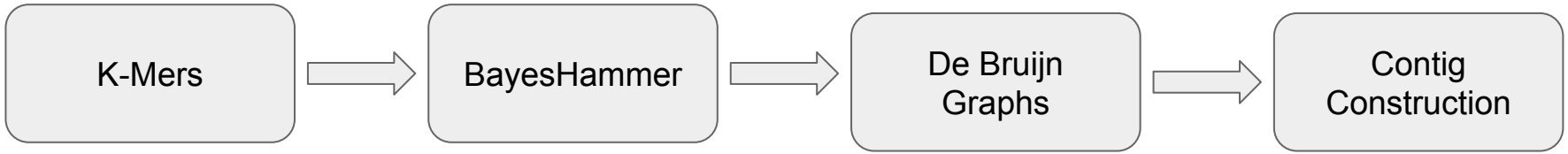
- ➔ Take many copies of a book
- ➔ Pass each of them through a shredder with a different cutter
- ➔ Piece book by looking at shredded pieces only

Original book might have repeated paragraphs, and shredding might result in typos. Excerpts might also be added, making some shreds unrecognizable.

SPADES

- Genome assembly algorithm and software package
- Works with many sequencing technologies including Illumina
- Compatible with both paired ends and single reads
- Ideal for single cell and multi-cell bacterial datasets
  - Doesn’t work as well for mammalian data
- Pipelined process centered around de Bruijn graphs

SPADES PIPELINE



The reads are broken into k-mers that overlap as much as possible.

BayesHammer is an error correction tool. Uses Bayesian clustering to identify erroneous reads and reduce the set of k-mers.

Assemble the De Bruijn graph and iteratively simplify by removing bubbles.

Backtrace the assembly graph to assemble a contiguous sequence.

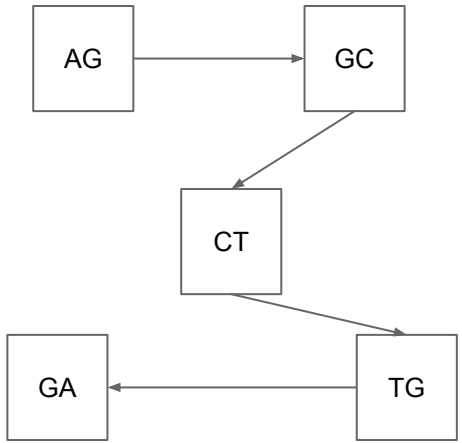
De Bruijn graphs

- Data structure for the shortest common superstring problem
- Directed graph
- Each node represents a unique pattern that exists in some origin sequence given an alphabet
  - Genomic alphabet: [A, C, T, G]
- Each edge represents a sequential overlap
- Forms the basis of many assembly algorithms
  - SPADES
  - Velvet

De Bruijn graphs - Example

Sequence: AGCTGA
3-mers: AGC, GCT, CTG, TGA

(K-1)mers: AG, GC GC, CT CT, TG TG, GA



- One edge per k-mer
- One node for each distinct (k-1)mer
- An Eulerian walk through the graph will return the original sequence
  - Eulerian walk - walk each edge exactly once

De Bruijn graphs

- This was a very simple example
- In practice this is much more difficult
  - K-mers are constructed from random reads
  - Massive amount of nodes and edges
    - Assemblers use larger k-mers
      - Larger k-mers = Greater # of possible nodes
  - These reads may have a lot of overlap
  - Reads are not in order
  - The reads can have mistakes or gaps
  - Nodes will have multiple incoming and outgoing edges

Assembly Statistics

- N50
- GC %
- Largest contig
- Average contig size