SPAdes support for third-party assembly graphs

Natalia Zenkova

Scientific advisor:
Andrey Prjibelski
Anton Korobeynikov

Center for Algorithmic Biotechnology, SPbU
Global problem

- **SPAdes** outputs contigs and genome assembly graph (FASTG/GFA)
- **SPAdes** do not have the ability to accept such graphs as input

**Goal:**

- Implement support for third-party assembly graphs (C++)
Assembly formats

- **FASTG** is a format for faithfully representing genome assemblies in the face of allelic polymorphism and assembly uncertainty.

- **The Graphical Fragment Assembly (GFA)** is a tab-delimited text format for describing a set of sequences and their overlap.
  - **Header** — identifies the type of the line;
  - **Segment** — a continuous sequence or subsequence;
  - **Link** — an overlap between two segment;
  - **Containment** — an overlap between two segments where one is contained in the other;
  - **Path** — an ordered list of oriented segments, where each consecutive pair of oriented segments are supported by a link record.
Stages in SPAdes

- read_conversion;
- construction;
- simplification;
- load_graph;
- hybrid_aligning;
- late_pair_info_count;
- distance_estimation;
- repeat_resolving.
• Studied next assembly formats: FASTG and GFA
• Launched SPAdes
• Understood SPAdes structure
• Launched SPAdes without spades.py
• Launched SPAdes from the specified stage
• Constructed the new stage: load_graph
• Implement support for third-party assembly graphs

The implementation of this functionality will significantly expand the scope of using SPAdes.
Another sequence graph

- SPAdes accepts de Bruijn graph as input

**Question:**

- What if another type of sequence graph is provided?
Algorithm

**Input:** Graph with overlaps = 0, k-mer size \( k \)

**Output:** De Bruijn graph

**Algorithm:**

1. Iterate through vertices
2. Continue outgoing edges and complementary edges
3. Repeat until all vertices are processed

**Example:**
• There are equal edges in input graph
• There is an edge shorter than $k$
• Does it exist?
Git repository:
https://github.com/NataZen/SPAdes
Thank you for attention!
Any questions?