Expandable de Novo Genome Assembler for Short-Read Sequence Data

Nikolay Vyahhi Sergey Nurk Anton Bankevich Max Alekseyev Pavel Pevzner

Algorithmic Biology Lab, Academic University of the Russian Academy of Sciences, St. Petersburg, Russia

http://bioinf.spbau.ru/en/

Abstract

De novo genome sequence assembly is the essential step to reveal genomic sequences of different species world-wide. Currently there exists various genome assemblers for short-read NGS data, such as Velvet, SOAPdenovo, ALLPATH, ABySS and others.
We present new open-source de Bruijn graph-based assembler currently in development on C++, which uses novel algorithmic ideas such as context-free graph approach and also have agile and expandable software architecture. It requires affordable amount of memory and computations while giving high quality results. It provides solid basis for single-cell and mammalian assemblers in the near future.

In the Poster

1. De Bruijn graph construction, index and callbacks.
2. Error correction and corruption based on topology of the de Bruijn graph (e.g. tip clipping, bulge removal).
3. Efficient data structures for storing and handling of the additional information in the de Bruijn graph (e.g. coverage or distances between pairs of edges).

Graph Construction

- Vetexes: K-mers. Edges: $(K+1)$-mers.
- All simple pathes (i.e. without branchings) are condensed.
- Hash-index $(K+1)$-mer $\rightarrow (edge, offset)$.
- 2 bits per nucleotide in sequences. 0 bits overhead for sequences with compile-time lengths (i.e. K).
- Every edge/vertex knows it's reverse-complementary edge/vertex.

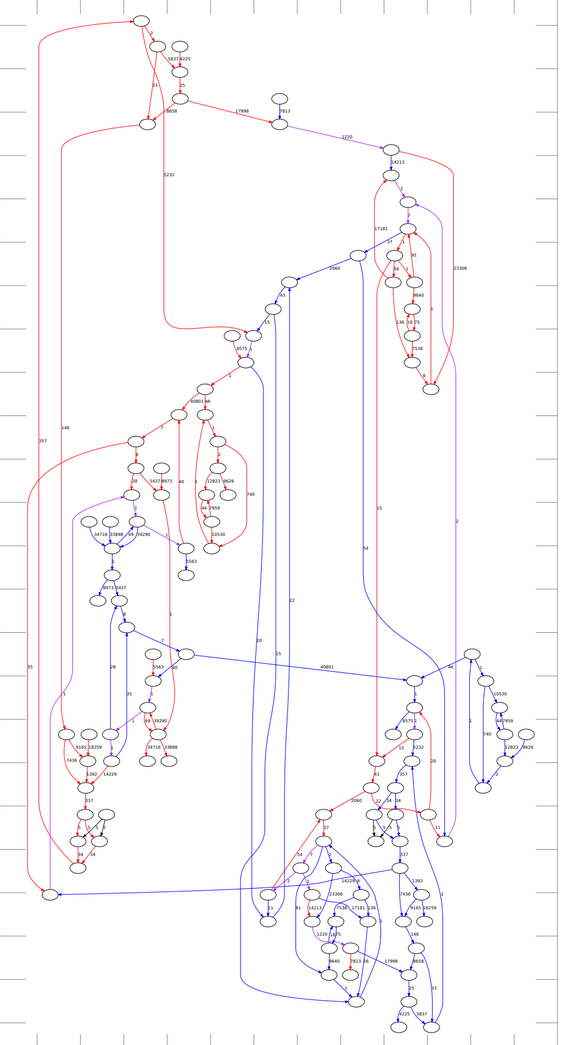
Graph Operations

- Addition/deletion of edge/vertex.
- Concatenation of a path.
- Split of an edge in a given position.
- Projection of one edge (with all additional info) to another edge.

Callback technique: each time a modifying operation is performed corresponding event is triggered and all objects that listen to graph events are notified about what exactly happened to the graph. It allows to implement any additional graph information as an external structure and hide its maintaining logic from the graph and from processing algorithms.

Errors Handling

- **Tip clipping**.
- **Bulge removal**.
- **Low-coverage edges removal**.



Additional Info

- Coverage.
- Distances between edges.

Practical results

E.Coli MG1655-K12.
12M Illumina paired-end reads (length = 100, gap = 20).
Correction with Quake (Hammer for future).
...
PROFIT!!!

Conclusions

Something was done.

Future

0. Repeat resolving.
1. Single Cell (Bacterial) Assembler.
2. Mammalian Genomes Assembler (see the poster of *Mikhail Dvorkin* and *Alexander Kulikov*: Earmark graph approach to de novo genome assembly).
3. Transcriptome Assembler, Cancer Genomes Assembler and other customizations...