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 and index of K-mers.
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

Errors Handling

Additional Info

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...