Combining de Bruijn graph, overlaps graph and microassembly for de novo genome assembly

Anton Alexandrov, Sergey Kazakov, Sergey Melnikov, Alexey Sergushichev, Anatoly Shalyto, Fedor Tsarev
St. Petersburg National Research University of Information Technologies, Mechanics and Optics
Genome Assembly Algorithms Laboratory

Assembler architecture

De Bruijn graph

Quasi-contigs assembly

Errors correction

Overlap graph

Contigs assembly

Microassembly

Errors correction

• Reads truncation
• K-mers frequency analysis
• Split k-mers into buckets according to their prefixes

Quasicontigs assembly

1. Set of paired-end reads

2. De Bruijn graph

3. Unique paths

4. Non-unique paths

5. Unique paths are transformed to quasicontigs

Contigs assembly

• Quasicontigs are given as input to “overlap-layout-consensus” module
• Short quasicontigs are thrown out to get to a reasonable size of an input data, e.g. 10-fold coverage

Microassembly

• All of the paired-end reads are aligned to the contigs with Bowtie (reads in a pair are aligned independently).
• If both reads in a pair are aligned to different contigs such reads are called bridging and the contigs are called bridged (see Figure).
• For every pair of bridged contigs we can infer their order from orientations of alignments of the bridging reads.
• All pairs of reads with at least one read aligned to one of these contigs are used to build a relatively small de Bruijn graph.
• In this graph we search for the path connecting two contigs in the same as quasicontigs are assemblded

Experiments

• Dataset – E. Coli genome 160-fold coverage paired-end reads library SRR016655 with insert sizes of about 200 bp.
• We got about 10 million quasicontigs with a total size of two Gbp.
• This data was truncated to 175 Mbp.
• After contigs assembly there were 525 contigs with an N50 size of 17804 and a maximum size of 73908.
• After microassembly there were 247 contigs with an N50 size of 53720 and a maximum size of 167319. This contigs cover 98% of the reference genome.

Acknowledgements

• This research was supported by the Ministry of Science and Education of Russia in the context of Federal Program “Research and development on priority directions of scientific-technological complex of Russia” and Federal Program “Scientific and pedagogical personnel of innovative Russia”

alserg@rain.ifmo.ru http://genome.ifmo.ru/en/