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

Anton Alexandrov, Sergey Kazakov, Sergey Melnikov, Alexey Sergushichev[1], Anatoly Shalyto, Fedor Tsarev

St. Petersburg National Research University of Information Technologies, Mechanics and Optics
Genome Assembly Algorithms Laboratory
197101, Kronverksky pr. 49, St. Petersburg, Russia

In this paper we present a method for de novo genome assembly that splits the process into three stages: quasicontigs assembly; contigs assembly from quasicontigs; contigs postprocessing with microassembly. We have carried out an experiment of assembling the *E. Coli* genome from an Illumina Genome Analyzer 160-fold coverage paired-end reads library SRR001665 with insert sizes of about 200 bp and got 247 contigs with an N50 size of 53720 and covering 98% of the reference genome.

The first stage uses a de Bruijn graph built from all the input data. For each pair of reads a path connecting reads' beginning *k*-mers is searched for, assuming that reads are directed inwards. For this we are searching for all paths connecting these *k*-mers with lengths bounded from up and down by *a priori* limits of insert sizes. This is done by a pair of simultaneous breadth-first searches starting from the *k*-mers. If all paths found have the same length and are similar to each other then we have a sequence likely to be in the genome. We call these sequences quasicontigs as they are far from being contigs but are greater than raw reads.

For the second stage the previously assembled quasicontigs are used. In the beginning short ones are thrown out to get to a reasonable size of an input data, e.g. 10-fold coverage can be kept. Then contigs are assembled with the algorithm based on the overlap-layout-consensus approach.

The third stage is similar to OLC and scaffolding. We are trying to order the contigs and fill the gaps between them. At first all of the paired-end reads are aligned to the contigs using Bowtie (reads in a pair are aligned independently). Then if both reads in a pair are aligned but 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. After that all pairs of reads with at least one read aligned to one of these contigs are used to build a relatively small (thus, microassembly) de Bruijn graph.

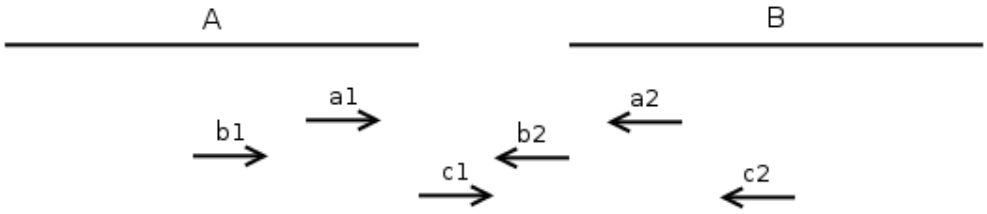

Figure. Contigs A and B are bridged, reads a1 and a2 are bridging, pairs (b1, b2) and (c1, c2) can be used for microassembly.

As graph is small and "local" we are likely to find a path connecting reads in a bridging pair using the same technique as in the first stage of the whole algorithm (quasicontigs assembly). This path gives us a distance between contigs and a filling sequence. After the distance is determined (it's accurate, not like in scaffolding) we have a layouting tasks similar to the one of the second stage.

On the *E. coli* dataset after the first stage we had about 10 million quasicontigs with a total size of two Gbp. Then this data was truncated to 175 Mbp. After the second phase there were 525 contigs with an N50 size of 17804 and a maximum size of 73908. After the third phase there were 247 contigs with an N50 size of 53720 and a maximum size of 167319.

[1] Corresponding author. Email: alsergbox@gmail.com