UDC 004.021

Combining De Bruijn Graphs, Overlap Graphs and Microassembly for *De Novo* Genome Assembly

A. A. Sergushichev, A. V. Alexandrov, S. V. Kazakov, F. N. Tsarev, A. A. Shalyto

Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics, Russia, 197101, St. Petersburg, Kronverkskiy pr., 49, alserg@rain.ifmo.ru, alexanrdov@rain.ifmo.ru, svkazakov@rain.ifmo.ru, tsarev@rain.ifmo.ru, shalyto@mail.ifmo.ru

In this paper we present a method for *de novo* genome assembly that splits the process into three stages: quasicontig assembly; contig assembly from quasicontigs; contig postprocessing with microassembly. The first stage uses de Bruijn graph, the second one uses overlap graph. We have carried out experiments of assembling the *E. Coli* genome (size ≈ 4.5 Mbp) and *Maylandia zebra* genome (size ≈ 1 Gbp). Advantage of proposed method is a low memory consumption.

Key words: genome assembly, contigs, de Bruijn graph, overlap graph, microassembly.

References

1. *Illumina, Inc.* Available at: http://www.illumina.com/ (Accessed 18, May, 2012).

2. Böckenhauer H.-J., Bongrratz D. *Algorithmic Aspects of Bioinformatics*. Springer, 2007, 396 p.

3. Pevzner P. A. 1-Tuple DNA sequencing: computer analysis. *J. Biomol. Struct. Dyn.* 1989. vol. 7, pp. 63–73.

4. Zerbino D. R., Birney E. Velvet : Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 2008, vol. 18, pp. 821–829.

5. Butler J., MacCallum I., Kleber M., Shlyakhter I. A., Belmonte M. K., Lander E. S., Nusbaum C., Jaffe D. B. ALLPATHS : De novo assembly of wholegenome shotgun microreads, *Genome Research*, 2008, vol. 18, pp. 810–820.

6. Simpson J. T., Wong K., Jackman S. D., Schein J. E., Jones S. J., Birol I. ABySS : A parallel assembler for short read sequence data. *Genome Research*, 2009, vol. 19, pp. 1117–1123.

7. Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., Li Y., Li S., Shan G., Kristiansen K., Li S., Yang H., Wang J., Wanget J. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 2010, vol. 20, pp. 265–272.

8. Pevzner P. A., Tang H., Waterman M. S. EULER : An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci.*, 2001, no. 98, pp. 9748–9753.

9. Aleksandrov A. V., Kazakov S. V., Melnikov S. V., Sergushichev A. A., Tsarev F. N., Shalyto A. A. Errors Correction Method in the Readings Set of Nucleotide Sequence. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2011, no. 5, pp. 81–84 (in Russian).

10. Okanohara D., Sadakane K. Practical entropy-compressed rank/select dictionary. *Comput. Research Reposit.*, 2006. Available at: http://arxiv.org/abs/cs/0610001 (Accessed 18, May, 2012).

11. Chikhi R., Rizk G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *Algorithms in Bioinformatics*, 2012, pp. 236–248.

12. Gusfield D. *Algorithms on String, Trees and Sequences. Computer Science and Computational Biology*. Cambridge Univ. Press, 1997, 554 p. (Rus. ed.: Gusfild D. *Stroki, derev'ia i posledovatel'nosti v algoritmakh. Informatika i vychislitel'naia biologiia*. St. Petersburg, Nevskii dialekt Publ., 2003, 656 p.).

13. *The Assemblathon*. Available at:
http://www.assemblathon.org (Accessed 18, May, 2012).