Metagenome Assembly: Explanation, Challenges and Future Trends

S. Momken[a], K. Kavousi[b]*, A. Banaei-Moghaddam[b], D. Moazzami[a]
[a] Department of Algorithms and Computation, University of Tehran, Tehran, Postal code 1417466191, Iran
[b] Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Postal code 1417614411, Iran
*kkavousi@ut.ac.ir

Abstract: For many years, retrieving genomic sequence of undiscovered species was a complicated task, mainly because sequencing of the DNA as a whole was not possible. Therefore, original sequences of the genome had to be assembled de novo from huge number of overlapping small reads from different copies of the genome. When working on a microbiome sample, many of the including species can't be cultured in the laboratory and genomic fragments of all species must be read and assembled later without knowing the origin of each fragment. The set of all these reads is called a metagenome. The above-mentioned circumstances make metagenome assembly even harder than genome assembly. The initial assembly of both genomic and metagenomic data is based on graph algorithms, specially those using De Bruijn graphs. In this paper, we will introduce different stages of metagenome assembly, the algorithms and time complexity of each stage and the influence of each technique on the final result of the assembly. Various challenges are encountered in this process, such as detection and correction of sequencing errors, grouping reads of each genome, finding shared reads and repeated regions, resolving differences between strains and time/memory complexity of the algorithms to examine feasibility of running them on big data. We will list significant metagenome assembly tools and as an example will briefly introduce metaSPAdes [1] (an extended version of SPAdes assembler [2] for metagenomic data). Finally, we will mention new trends and promising approaches in sequencing and assembly of both genomes and metagenomes which can alleviate current difficulties and have revolutionary improvements in length and accuracy of assembled sequences.

Keywords: Metagenome; Assembly; Graph; Machine Learning; Algorithms.

References

[1] S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, "MetaSPAdes: A new versatile metagenomic assembler," Genome Res, 27; 5 (2017) 824-834.
[2] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. L. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, P. A. Pevzner, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," J. Comput. Biol., 19; 5 (2012) 455-477.