# A-Bruijn Graph Approach to de novo genome assembly

Mikhail Dvorkin, Alexander Kulikov, Max Alekseyev

July 5, 2011

A common approach to assembling a genome from short reads is construction of the de Bruijn graph on k-mers from the reads and finding a traversal of edges in this graph. We propose a new approach that allows to decrease the graph size without losing the information from the input data.

We earmark only a small fraction of all k-mers such that each read contains at least two earmarked k-mers. Earmarked k-mers are then represented by vertices so that each read is represented by a line graph on such vertices. Further gluing vertices corresponding to the same k-mers results in an A-Bruijn graph, where each read is present as a path. After simplifications, non-branching paths in the A-Bruijn graphs reveals the genome contigs as sequences of reads; the contig content can be then found by consensus over the corresponding reads.

The A-Bruijn graph has a number of advantages as compared to the traditional de Bruijn graph: it has less vertices and thus reduces memory usage (that is particularly important for large genomes); it as well captures the repeat structure of the genome being assembled; since it is based only on a small fraction of all k-mers, it is less sensitive to the errors in the reads.

Many algorithms (graph simplification, mate pairs analysis etc.) that work on de Bruijn graphs can be adapted to work on A-Bruijn graphs.