Identifying Alternative Splicing Events in RNAseq data using De Bruijn Graphs and Bloom Filters

Ricardo Medeiros da Costa Junior, André Yoshiaki Kashiwabara

Universidade Tecnológica Federal do Paraná - Campus Cornélio Procópio, Departamento de Computação, Programa de Pós-Graduação em Bioinformática - PPGBIOINFO

Alternative splicing (AS) is a post-transcriptional mechanism in which multiple functional transcripts might be produced from a single gene. In particular, a gene encoding the protein may produce different proteins through pre-mRNA AS events. In this process, some exons may be included or excluded from the final messenger RNA (mRNA). In consequence, AS mRNA translated protein contains differences in their amino acid sequences and often in their biological functions. The AS process allows the human genome directly synthesize many proteins that could be expected from the 20,000 protein-coding genes. Recent studies have linked abnormally spliced mRNAs with cancerous cells.

In 2012, it was proposed an algorithm for the identification and quantification of polymorphisms of data from RNA-seq when the reference genome is not available without assembling of full transcripts. Although this algorithm identify both approximate tandem repeats, SNPs (single nucleotide polymorphism) and AS, it is only focused on quantifying AS. Due this method, it was possible to realize that annotation of AS events have been underestimated, which 56% of AS identified in the tested dataset were not present in the current notes. However, the algorithm has some limitations. Like most new assemblers based on DBG, (De Bruijn Graphs) the construction of graph requires a very high cost of memory and must be run on a cluster.

In 2016 an article was published which proposes an improvement to an assembler based on DBG. It was removed MPI (message-passing system) and it was implemented Bloom Filter, that is a probabilistic data structure, in the construction of DBG. It was possible run it in a personal computer rather than a cluster. Bloom filter is a probabilistic data structure created by Burton Howard Bloom in 1970, which is used to test whether an element is a member of a set. False positive combinations are possible, but false negative are not, because of that Bloom filter is considered 100% recall rate. That is, it returns 100% of the relevant results.

As the construction of the DBG of this assembler is very similar to that algorithm that indentify and quantify AS, the purpose of this work is the implementation of the Bloom Filter in the identification and quantification AS algorithm, reducing the cost of memory for the creation of DBG, allowing that runs efficiently on a personal computer.