TITLE: A TOOL FOR CLOSING GAPS IN GENOME ASSEMBLY BASED ON THE GROUPING OF REPETITIVE REGIONS

AUTHORS: NEGRÃO, D. M.; PANTOJA, G,B; SILVA, A. L. ; RAMOS, R.T.

INSTITUTION: [1]UNIVERSIDADE FEDERAL DO PARÁ - (R. Augusto Corrêa, 1 - Guamá, Belém - PA, 66075-110) - BRASIL.

ABSTRACT:

The assembly of the genome refers to the process of obtaining a large number of short sequences of DNA from sequencing and through assembly techniques to try to reassemble the representation of the original chromosomes from which the DNA originated. The assembly techniques are divided into two assembly categories by reference, where information from a known sequence (reference) is used for the reconstruction of the sequence of interest (consensus sequence) and *de novo* which consists of comparing the sequencing readings between each other, being the result of the overlap of the alignments used to construct a consensus sequence. There are many problems that hinder the assembly of genomes and consequently lead to Gaps, such as very small fragments of sequences, regions with low coverage, repetitive regions among others. Because of these and other problems, the assembly of genomes becomes a very complex task, and consequently, many genomes are not completely represented in their assemblies, as is demonstrated in the GOLD database. To try to lessen one of these assembly-related problems, we have developed a tool that uses machine learning techniques to group repetitive regions of the genome and thereby calculates an ideal k-mer to be used in the assembly for each group, thereby generating contigs that are used to close GAPs and thus represent regions that were not previously represented in the default assembly. The tool was developed in the Python programming language and uses the spades software to do assembly again, the grouping of the repetitive regions is done with the K-means algorithm, and the closing of gaps is executed by Gapblaster software. In addition, a graphic web interface was developed in HTML and JavaScript to make the tool user-friendly in a way that makes it easy to use by the scientific community. As a result, the tool demonstrated improvement in assembly data such as an increase in N50 and completeness of the genome.

Keywords: Genome assembly, Machine Learning, Gaps Closure, Clustering.

Development Agency: CNPq