TITLE: NePDeCA - New Product Detection for Comparative Analysis

AUTHORS: OLIVEIRA, M. S.[1]; SÁ, P. H. C. G.[2]; ALVES, J. T. C.[3]; M, B.[1]; F, H.[1]; VERAS, A. A. O.[1]

INSTITUTION: [1]FEDERAL UNIVERSITY OF PARÁ (RUA ITAIPU, 36, CEP 68464-000, TUCURUÍ, PARÁ, BRAZIL); [2]FEDERAL RURAL UNIVERSITY OF AMAZONIA (RODOVIA PA 140, CEP 68680-000, TOMÉ-AÇU, PARÁ, BRAZIL); [3]PARÁ STATE UNIVERSITY (AV. HILÉIA, S/N, CEP 68502-100, MARABÁ, PARÁ, BRAZIL).

ABSTRACT:
The advances of second-generation sequencing platforms provided an exponential growth in the storage of biological information in public databases such as the NCBI (National Center for Biotechnology Information). This has boosted the genomic knowledge of several organisms and enabled numerous other analyzes to be performed. Among them we can cite the comparative analysis. The main task of the comparative analysis is the use of homology to identify orthologous and paralogous genes. It can also be used in the structural annotation process. However, most of the computational tools used to perform comparative analysis are started by extensive and complex command lines and use complete genomes as input for analysis which leaves out an expressive number of draft genomes already deposited in public databases. In order to solve these problems, we developed the NePDeCA tool, capable of performing comparative analysis automatically using complete and drafts genomes. It also allows identification of products missing from the original genomic sequence due to some errors, such as errors from the assembly genome process. The new products identified in this process are added to the original genomic sequence and then the comparative analysis is done. The methodology used to perform analysis in the NePDeCA tool is based on a pipeline tasks that are: Mapping reads (Bowtie2); *De novo* Assembly (SPADes); annotation process (Web RAST platform); local search to similarity (local BLAST) and comparative analysis (PGAP). The pipeline was divided into two parts, the first part focuses on identify new products and update the input file, so, when complete, the file is sent to the annotation process. The second part make a comparative analysis. To validate the tool, was used seventeen organisms: two draft genomes and fifteen complete genomes. All organisms are available to download in SRA database. To check the NePDeCA accuracy we perform two analysis: one using PGAP by command line and another with the NePDeCA tool. The alfa value observed was 0.6545 and 0.6613 respectively. The results highlight that the new gene products identification impacts the alfa value, without altering the number of organisms present in the analysis. The NePDeCA tool has a intuitive graphical interface and enables users to resume the tasks in case of error or interruption.

KEYWORDS: Bioinformatics, Comparative analysis, NGS, Products Genes, Software