TITLE: SMAPS – A WEB TOOL TO EXTENDS CONTIGS TO REDUCE GAPS WITH UNMAPPED READS

AUTHORS: POMPEU, L. N.; LOBATO, A. R. F.; SILVA, A.; RAMOS, R. T. J.

INSTITUTION: UNIVERSIDADE FEDERAL DO PARÁ, BELÉM, PA (ESPAÇO INOVAÇÃO, AVENIDA PERIMETRAL, KM 01 – GUAMÁ. CEP: 66075-750)

ABSTRACT:
The advent of the Next Generation Sequencing (NGS) brought with it a remarkable growth of drafts genomes in public databases and problems such as short length reads and high repetition rates, making the genome assembly process more complex, and some regions cannot be represented after the assembly: gaps. Thus, a computational-web tool able to finish the assembly improving the representation of the bases based on unmapped raw data was developed, in order to reduce the number of gaps in the assemblies. The tool Smaps was developed using the pipeline structure, a process of running multiple programs in sequence, consisting of de novo assembly without reference of the short reads, using the SPAdes software, generating the sequences that will be used, in the first moment, to be annotated in the Prokka, aiming at primary identification of genes, as are used as reference to map the reads through the software bowtie2 and samtools, to identify the reads that were not used in the genome assembly process, that is, the unmapped reads. The reads that were not mapped were used in the scaffolding process of the primary assembly using SSPACES, which consists in using the unmapped reads with the contigs of the assembly, to do the extension of the contigs. With the extended contigs, there is again the use of Prokka, for the identification of genes with extended assembly. The Prokka software generates the GenBank files with the annotation, and since there were two distinct uses of the tool, one with annotations of the primary assembly and another with extended assembly, they are used to a comparison between them, consisting in verifying the possible existence of a greater number of genes annotated in the Genbank file of the extended assembly. To verify the accuracy of the tool, two samples of Escherichia coli, SRR1424625 and SRR2000272, were used. The results were that in SRR1424625 there was an extension of 156 bases in 14 contigs, of the 190 existing ones, however, there was no increase in the number of annotated genes. In the sample SRR2000272, there was an extension of 3172 bases in 85 contigs, out of the 209 existing, having an increase of 2 genes in the annotation. The front-end of this pipeline, based on a web interface was also developed using PHP, embedded within HTML and JavaScript web development languages. Through this interface, it is possible to query information from the database previously created with the genes mapped and extended in the process.

Keywords: scaffolding, extend contigs, assembly improvement

Development Agency: CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico, CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior