MARVEL: A pipeline for recovery and analysis of viral genomes from metagenomic shotgun sequencing data

Deyvid Amgarten[1], Aline Maria da Silva[2], João Carlos Setubal[2]

1 USP - DEPARTAMENTO DE QUIMICA
2 USP

Abstract

The study of the viral diversity in environmental samples has become increasingly important due to the recognition of key roles played by these organisms in diverse ecosystems. Recent works provide evidence that viruses of bacteria (bacteriophages) are key players in biogeochemical cycles of large ecosystems, such as oceans and forests. Viruses may also be determinant in the flux of genes among microbial populations and in the plasticity of microbial communities, helping these communities to deal with environmental stresses. Knowing the genomes of viruses that are present in diverse environments can thus help the understanding of the microbial ecology and evolution of these environments. Here we describe the MARVEL pipeline for recovery and analysis of viral genomes from metagenome shotgun sequencing data. The main steps in this pipeline are: sequence quality control, metagenome assembly, similarity searches against wide and hallmark-protein databases of viruses, removal of false positives, and multisample contig binning. At the end, MARVEL generates an automatically curated set of contigs that correspond to draft and complete genomes of environmental viruses present in the analyzed sample. We have applied MARVEL to metagenomic datasets obtained in two environments (composting and a reservoir) of the Sao Paulo Zoo. We obtained 37 viral genomes from reservoir samples and 36 viral genomes from composting. Most of these genomes have low or no similarity with viral genomes in public databases. Therefore these results are a contribution for shedding light on the gigantic viral dark matter that exists in our planet. MARVEL can be applied to any shotgun metagenomic dataset for which Illumina reads are available.

Funding: Funding for this research is provided by FAPESP (2014/16450-8) and CAPES