Novel bioinformatic approaches for viral discovery from NGS data

Liliane S. Oliveira[1], João M. P. Alves[1], Dolores U. Mehnert[2], Alan M. Durham[3], Paolo M. A. Zanotto[2] and Arthur Gruber[1]*

[1]*Dept. of Parasitology and* [2]*Dept. of Microbiology, Institute of Biomedical Sciences, USP, São Paulo, Brazil;*
[3]*Dept. of Computer Sciences, Institute of Mathematics and Statistics, USP, São Paulo, Brazil.*
*Correspondence: argruber@usp.br

Viruses are the most abundant biological entities and play an important role in defining the composition of microbial communities. Some of the most devastating pandemic diseases have arisen through the transmission of viruses originally infecting wild and domestic animals. Thus, a systematic surveillance for emerging viruses with new computational tools is of utmost importance. In this work, we report the development and implementation of some bioinformatic approaches using profile HMMs (pHMMs) for viral discovery. Profile HMMs are used in a variety of bioinformatic applications and the most relevant publicly available databases for viruses are vFam and viralOGs (a subset of eggNOG). We developed virDB-Pack, a suite of programs to quantify, manipulate and select pHMMs from vFams and viralOGs databases. The package allows using the selected pHMMs to screen sequencing datasets for known and potentially emergent viruses. A preliminary survey using 506 selected pHMMs against a metagenomic dataset from raw sewage revealed a variety of novel viruses, including sequences from smacovirus, densovirus and circovirus. Another dataset, composed of genomic reads of *Aedes aegypti*, allowed us to identify a large number of viral sequences integrated into the mosquito genome (see abstract by E. Aguiar & A. Gruber – X-Meeting 2016) and a 9-kb genome segment from a recently described Phasi Charoen Like-virus (PCLV). We are also developing an algorithm for the construction of pHMMs based on an iterative enrichment of viral sequence representation. The method involves an initial screen of the sequencing dataset with a pHMM constructed from a few sequences, and recruitment of positive reads and sequence reconstruction using GenSeed-HMM (Alves *et al.* - Front Microbiol. 7:269, 2016), a tool recently developed by our group. Resulting contigs are then translated and aligned, with the most conserved blocks being automatically selected and used to construct distance trees. Sequence and taxonomic redundancy are eliminated by clustering using patristic distance and user-defined parameters. The selected representative sequences are used to build a new pHMM that covers a higher viral sequence diversity. Preliminary results demonstrate that increasing the number of non-redundant sequences used to generate pHMMs leads to a better ability to detect viral diversity until a saturation point is attained. The approach proposed here, when applied to a wide number of viral families, will allow the detection of viruses phylogenetically distant from those already known, with potential application in viral discovery studies and epidemiological surveillance. Support: Fellowships from CNPq and CAPES.