# Rational design of profile HMMs for viral detection, classification and discovery

Liliane Santana Oliveira Kashiwabara[1], Dolores U. Mehnert[1], Paolo M. A. Zanotto[1], Alan Durham[1], Alejandro Reyes[1], Arthur Gruber[1],

*1 USP*

**Abstract**

Some of the most devastating pandemic diseases have arisen through the transmission of emerging viruses that have not been detected before the tragic consequences of their dissemination. The detection of novel viruses is a challenging task due to the their high evolutionary rates. Profile HMMs are a powerful way of modeling sequence diversity and constitute a very sensitive approach to detect emerging viruses. In this work, we report the development and implementation of TABAJARA, a tool for rational design of profile HMMs. Starting from a multiple sequence alignment (MSA), TABAJARA is able to find blocks that are either (1) conserved across all sequences or (2) discriminative for two specific groups of sequences. For the identification of regions conserved across all protein sequences of an MSA, we implemented a previously described algorithm based on Jensen–Shannon divergence. In the case of nucleotide sequences, TABAJARA can use Shannon entropy or, alternatively, different substitution matrices to define position-specific scores. To find group-discriminative blocks, the program uses Mutual Information or Sequence Harmony for both, DNA or protein sequences. Once position-specific scores have been determined, TABAJARA uses a sliding-window to screen the whole alignment and delimit top-scoring regions. The program automatically extracts the selected alignment blocks, discards identical sequences, and builds the corresponding profile HMMs, which can then be used for many potential applications. To validate such models for viral detection, classification and discovery, we used two different viral taxonomic groups: phages of the Microviridae family and viruses of the Flavivirus genus. Using different metagenomic datasets, we observed that profile HMMs generated by TABAJARA can successfully be used as seeds to reconstruct genome sequences with GenSeed-HMM program. In both viral groups, we were able to obtain wide-range seeds (generic for all members of Microviridae or Flavivirus); and narrow-range seeds, exclusive to specific Microviridae subfamilies (Alpavirinae and Gokushovirinae) or to particular flaviviruses (e.g. DENV, ZIKV or YFV). The approach proposed here, using short and specific sequences to build profile HMMs, represents a radical change, compared to viral models from public databases such as vFam and pVOGs, built from MSAs derived from full-length protein sequences. Narrow-range seeds can be used to detect and classify already known viruses, whereas more generic seeds are useful for detecting wider viral groups, as well as distantly related viruses that could represent potentially emerging pathogens.

Funding: PhD fellowship from CAPES (LSO)