A database of viral taxon-specific profile HMMs for the detection and classification of viral sequences

Wendel Hime Lima Castro, Arthur Gruber, Liliane Santana Oliveira Kashiwabara

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ, UNIVERSIDADE DE SÃO PAULO

Abstract

Viruses are the most abundant and diverse biological entities on nature, characterized by high rates of replication and evolution. The highly variable nature of their genomes, composed of either single- or double-stranded RNA or DNA molecules, implies that different enzymes are required for replication. No single gene is shared by all viruses and, therefore, there are no universal markers for viral phylogenetic analysis and taxonomic classification. Thus, different viral markers must be established to study distinct taxonomic groups. Profile HMMs are statistical models that incorporate the diversity of a set of primary sequences and constitute a very sensitive approach for detecting known and emerging viruses. Publicly available resources of viral profile HMMs are based on orthologous clusters, which may include sequences from different taxonomic groups, implying that models can often detect a wide and unpredictable range of taxa. We have recently developed TABAJARA, a program for the rational design of profile HMMs. TABAJARA uses a multiple sequence alignment (MSA) as input and is able to find short blocks that are either (1) conserved across all sequences or (2) discriminative for two specific groups of sequences. The program then runs validation tests against the training set and automatically define cutoff scores customized for each developed model. Alternatively, full-length sequences are used for model construction and highly specific profile HMMs may be obtained when using the assigned cut-off scores. This approach has been successfully used to develop models for some taxonomically specific viral groups, including Microviridae phages and viruses of the genus Flavivirus. Given the utmost importance of developing bioinformatic resources and tools for novel virus detection and classification, we decided to extend our approach to all viral taxa represented with protein sequences on the NCBI's Identical Protein Groups (IPG) database. We implemented a Python pipeline to perform the following steps: (1) automatic sequence retrieval from IPG database, according to a pre-defined list of queries; (2) data organization and storage; (3) multiple sequence alignment for each query; (4) profile HMM construction; (5) model validation; and (6) report generation. We obtained 20.749 models, comprising prokaryotic and eukaryotic viruses. Relational searches can be performed on a web front end according to the taxonomic name or ID, protein name, type of model (short or full-length) and host (prokaryotic or eukaryotic), among other parameters. The selected models can then be downloaded. We expect this database will become an important resource for the scientific community.

Funding:

Link to Video: