

Profile HMMs as auxiliary tools for the taxonomic classification of viruses: a case study using Spounarivinae phages

Liliane Santana Oliveira Kashiwabara, Miriã Nunes Guimarães, Wendel Hime Lima Castro, Arthur Gruber

USP

Abstract

Profile HMMs are probabilistic models that are much more sensitive to detect remote orthologs than conventional pairwise alignment methods. We have recently developed TABAJARA, a program for the rational design of profile HMMs. In this work, we report the development and application of profile HMMs for the detection and taxonomic classification of phages of the subfamily Spounavirinae. We obtained a dataset of bona fide Spounavirinae sequences from the NCBI's Identical Protein Groups (IPG) database, restricting the query to complete sequences of terminase and tail sheath protein (TSP) associated to txid857473. Protein sequences were aligned with MUSCLE and the resulting alignments processed by TABAJARA to produce profile HMMs specific at the levels of genus and subfamily. Specificity and sensitivity of all models were assessed by similarity searches against the training set. As a final validation procedure, we tested all models against a dataset of Spounavirinae-depleted Myoviridae sequences. These tests revealed the detection of 60 unique sequences of terminase and 66 of TSP, corresponding to 72 non-redundant viral genomes. To clarify whether these sequences represented false positives detected by the models or corresponded to misclassified sequences, we inspected their taxonomic assignment on the IPG and NCBI Taxonomy databases. With no exception, all sequences belonged to orphan genera (not included within any subfamily) or to unclassified viruses. Terminase and TSP sequences of this group, together with representatives of all Myoviridae subfamilies, including the Spounavirinae subfamily, were used in ML phylogenetic reconstructions with FastTree program. Trees derived from both protein datasets revealed that all sequences identified by our models constituted a monophyletic group, including sequences originally classified as Spounavirinae, as well as sequences from the orphan genera Cp51virus, B4virus, Bastillevirus, Wphvirus, Bc431virus, Nit1virus, Agatevirus and Sep1virus, and some previously unclassified Myoviridae viruses. This wide group of taxa corresponds to Herelleviridae, a new virus family recently proposed by Barylski et al. (Syst Biol. 2019 May 25. pii: syz036) using a variety of phylogenetic analyses based on genomic, proteomic and marker gene-based data. Similar results were also observed by Aiewsakun et al. (J. Gen. Virol. 99: 1331-1343, 2018), using a genetics-based platform that computes sequence relatedness between viruses. Altogether, this body of evidence suggests that our models, rather than detecting false-positives, had indeed identified mis- or unclassified Myoviridae sequences. We conclude that profile HMMs may be used as auxiliary tools for the taxonomic classification of known and emergent viruses.

Funding: CAPES