9

Rational Design of Profile Hidden Markov Models for Viral Classification and Discovery

Liliane Santana Oliveira[1] • Arthur Gruber[1,2]

[1]Department of Parasitology, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP, 05508-000, Brazil; [2]European Virus Bioinformatics Center, Leutragraben 1, Jena, 07743, Germany

Author for correspondence: Arthur Gruber, Department of Parasitology, Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, SP, Brazil.
Email: argruber@usp.br

Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch9

Abstract: This chapter provides an overview of the theoretical concepts and practical applications of methods for the rational design and application of profile hidden Markov models (profile HMMs) in viral discovery and classification. Profile HMMs are probabilistic models that represent sequence diversity and constitute a very sensitive approach for detecting remote homologs. One of the most relevant and challenging applications of profile HMMs is the discovery of viruses in metagenomic samples, a fundamental task for epidemiological surveillance. In this chapter, publicly available resources of viral profile HMMs are presented and the methods involved in their construction are discussed. Several aspects to be considered for the generation of profile HMMs are presented, including technical pitfalls that should be avoided, and the potential applications of such models for detecting specific viral sequences. This chapter also introduces a bioinformatics application that implements methods to select informative regions of a multiple sequence alignment and build profile HMMs with different taxonomic specificities. Additional programs using profile HMMs for targeted sequence assembly and detection of multigene entities are also presented. Such programs, integrated into a common framework for viral research,

In: *Bioinformatics*. Nakaya HI (Editor). Exon Publications, Brisbane, Australia.
ISBN: 978-0-6450017-1-6; Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021

are discussed in light of several biological issues that involve the classification and discovery of potentially emerging viral pathogens.

Keywords: profile hidden Markov models; viral bioinformatics; viral discovery; viral metagenomics; virus classification

INTRODUCTION

Profile hidden Markov models (profile HMMs) are probabilistic models that capture the diversity of biological sequences. A multiple sequence alignment (MSA) of protein sequences (nucleotide can also be used) is submitted to a position-specific scoring system. The model has a state for every position of the alignment, with each state presenting twenty results, one for each possible amino acid, and two for insertion and deletion (called indels) occurrences (1–3). The Markov chain represents a probabilistic model for the set of states and the transition probabilities between each state. Once a profile HMM is generated from an MSA of sequences belonging to an orthologous group, the patterns observed for this group can be found in other protein sequences of a database. Thus, any query sequence can be traced through the model across all states and then be scored according to the probabilities found for each transition (1).

Viruses constitute a group of highly divergent biological entities, characterized by evolutionary rates that are much higher than those observed in prokaryotes and eukaryotes (4–7). Due to this very high divergence, serological and molecular tests developed for known pathogens do not cross-react with emergent viruses, even when they belong to the same genus (3, 8, 9). More challenging still, viruses do not have universally conserved markers in their genomes that can be used as targets for PCR-based assays, such as 16S rRNA in prokaryotes and 28S rRNA in eukaryotes (3, 8, 10).

Pairwise alignment methods, implemented in programs such as BLAST (11), became a standard for identifying new sequences from sequencing data. However, these methods are not sensitive enough at identifying remote homologs (12), a relatively common situation in viral data. In addition, the scarce amount of viral data in public databases compared to prokaryotes leads many metagenomic sequencing projects to fail in the proper identification of relationships between novel sequences and known viruses (13). Profile-based alignment methods can significantly increase the ability to detect remote homologs (14). In this context, profile HMMs represent an attractive alternative to improve the ability to detect and classify viral sequences and are increasingly being used in viral metagenomic studies.

DATABASES OF VIRAL PROFILE HMMS

In the last decade, several different repositories of profile HMMs constructed from viral proteins became publicly available. Table 1 lists the most relevant and currently used resources. For additional information on older repositories, the reader is referred to a review article from Reyes *et al*. (3). One of the first resources

TABLE 1 Web resources of viral databases that include profile hidden Markov models

Database	Description	Last update	Reference
ClassiPhage	ClassiPhage is a program for phage taxonomic classification. A collection of profile HMMs for four phage families is provided. Source: http://appmibio.uni-goettingen.de/index.php [accessed on 12 Dec 2020]	Current	(27,28)
IMG/VR v3	IMG/VR Viral Resources is a database of viral genome sequences of cultivated and uncultivated viruses Source: https://img.jgi.doe.gov/vr; https://genome.jgi.doe.gov/portal/IMG_VR [accessed on 12 Dec 2020]	Current	(29,30,33)
pVOGs	Prokaryotic Virus Orthologous Groups is a database of orthologous groups built from genomes of viruses that infect bacteria and archaea. Provides accession IDs of viral proteins, lists of orthologous groups, alignments and profile HMMs. Source: http://dmk-brain.ecn.uiowa.edu/pVOGs [accessed on 12 Dec 2020]	2016	(17,18)
RVDB-Prot/ RVDB-Prot-HMM	Reference Viral Databases Source: https://rvdb-prot.pasteur.fr/ [accessed on 12 Dec 2020]	Current	(22,23)
vFam	vFam is a database of profile HMMs built from all viral protein sequences available at RefSeq. Viral protein sequences, annotations and profile HMMs are provided. Source: http://derisilab.ucsf.edu/software/vFam/ [accessed on 12 Dec 2020]	2014	(15)
Viral OGs/ eggNOG v5.0	The viral subset of eggNOG v5.0 is composed of viral sequences, annotations, alignments, trees, and profile HMMs. Source: http://eggnog5.embl.de/#/app/home [accessed on 12 Dec 2020]	Current	(20)
Viral MinionDB	MinionDB is a database of taxonomically defined profile HMMs built from viral protein markers Source: http://www.bioinfovir.icb.usp.br/minion_db/ [accessed on 12 Dec 2020]	Current	Unpublished as of February 2021
VOGDB	VOGDB provides information, interactive access and download for all Viral Orthologous Groups Source: (13) [accessed on 12 Dec 2020]	Current	Unpublished as of February 2021

to provide virus-derived profile HMMs is vFam (15), a collection of 5,585 profile HMMs constructed from 29,655 viral proteins of eukaryotic viruses, last updated in 2014. The construction pipeline used protein sequences annotated as viral (non-phages) from the RefSeq database (16). After sequence collapsing for 80% or greater identity and polyprotein removal, the remaining sequences were submitted to all-versus-all BLASTP searches, Markov sequence clustering, MSAs and profile HMM construction. All profile HMMs are provided with annotation files containing information such as functional annotation of the sequences composing the original MSA and their corresponding viral family and genera.

Another resource of profile HMMs is the Prokaryotic Virus Orthologous Group (pVOGs) (17), an update of the former Phage Orthologous Groups (POGs) database (17, 18). This repository is composed of profile HMMs constructed from orthologous groups of proteins from viruses that infect bacteria and archaea. The available version, last updated in 2016, comprises 9,518 orthologous groups from 296,595 proteins, totaling 18 viral families. Orthologous groups were obtained using RefSeq full-length protein sequences submitted to a clustering step through the identification of symmetric best matches shared between three genomes. Protein sequences used to construct the MSAs and the respective profile HMMs are mapped to taxonomic and functional annotation, which are also provided for download.

ViralOGs (19) was originally a viral subsection of eggNOG, a database of protein orthologous groups from different taxonomic levels associated with functional annotations. The current version of eggNOG (v5.0) (20) was upgraded from 325 (v4.5) to 2,502 viral proteomes, obtained from the Uniprot (21). In total, 8,318 profile HMMs derived from viral protein are available, together with functional, phylogenetic and taxonomic data.

As reported in a previous review on the use of profile HMMs for viral research (3), vFam, eggNOG and pVOGs present some limitations, such as the low representation of sequences used in the construction of each model, highly biased representation of viral families and the lack of a direct relationship of the models with specific taxonomic groups. For example, orthologous groups can be composed of protein sequences derived from multiple taxa, sometimes belonging to different viral families. Therefore, assigning taxonomy classification of new sequences based on similarity to profile HMMs of these databases is possible, but results should be taken with caution and ideally be complemented by additional evidence.

RVDB-prot (22) is a protein version of Reference Viral DataBase (RVDB) (23) constructed with a pipeline similar to that used for the construction of vFAM database (15). Viral proteins are submitted to a CD-HIT (24) step to remove duplicated sequences, followed by all-versus-all BLASTP (11) searches and, finally, to a clustering step using SiLiX (25), a similarity-based clustering tool. The current distributed version (v.20) comprises 13,621 profile HMMs. The database is updated on a regular basis, following changes of the RVDB database.

VOGDB provides a web front end of the Viral Orthologous Groups (VOG) database. Version vog202 contains 26,224 VOGs derived from 8,745 genomes and the corresponding profile HMMs. Virus-specific VOGs are defined according to different stringencies specified by the e-value, and correspond to 20,785, 22,430 and 23,149 virus-specific VOGs for high, medium and low stringency, respectively. The pipeline for VOG construction uses RefSeq data, quality/completeness annotation, pairwise alignments and choice of bidirectional best hits,

multiple sequence alignments, all-versus-all HHalign (26) searches, remote homology clustering and functional annotation of the groups. The method is unpublished, but a description of the entire methodology is available on the web site. Users can perform searches and VOGs are provided with functional annotation (when available), together with last common ancestor information. A plethora of files are also offered for download, including profile HMMs corresponding to the respective MSA of each VOG.

Chibani *et al*. (27) recently described ClassiPhage, a method for phage taxonomic classification using profile HMMs. This methodology builds phage family-specific profile HMMs using annotated phage genomes and employs refinement protocols to ensure model specificity. To validate the proposed methodology, the authors constructed profile HMMs for four phage families: *Myoviridae*, *Podoviridae*, *Siphoviridae*, and *Inoviridae*. The program and profile HMMs for the four phage families listed above is available for download (27). This work was extended to create ClassiPhage 2.0, an updated method that uses profile HMMs derived from phage sequences to train an Artificial Neural Network (ANN) to classify the phages into one of 12 different families (28). The generated models showed very high specificity for all families, but the observed sensitivity was very low, with only 3 out of the 12 families showing values above 50% and six families presenting sensitivity below 20%.

IMG/VR v3, the Integrated Microbial Genomes/Viral Resources v3 (29, 30), is the newest and largest resource of viral sequences, gathered, reconstructed and integrated from a large variety of sources, including tens of thousands of metagenomic datasets. IMG/VR is now the main resource and analysis framework of uncultivated virus genomes (UViGs), a conceptual definition supported and described by a series of standardized metadata, and widely adopted by the scientific community (31, 32). The whole IMG/VR v3 database, comprising 2,302,702 distinct UViGs (as of December 2020), is available for download, together with a collection of 25,281 curated profile HMMs derived from viral protein families (VPFs) (33–35).

A ROADMAP FOR THE RATIONAL DESIGN OF PROFILE HMMS

As presented in the previous section, different resources containing viral family groups and associated profile HMMs are publicly available. All these databases use a variety of methods to eliminate or reduce sequence redundancy and are orthology-oriented in the sense that viral sequence groups are obtained by a step of all-versus-all pairwise comparisons, followed by a clustering method. Sequences of each cluster are then submitted to a multiple sequence alignment and profile HMM construction. In addition, viral clusters are mapped to functional and taxonomic annotations.

Although some of these repositories offer extremely rich and diverse sets of viral data, the accompanying profile HMMs should be used with caution for viral detection and discovery in metagenomic datasets. In fact, before universalizing the use of profile HMMs, it is fundamental to know the potential limitations and risks of using such models. First, profile HMMs are built on top of an MSA and

their ability to be used as representatives of a profile of sequences, rather than a single sequence, is directly dependent on the quality of the training set; that is, the set of sequences used in the MSA. An ideal set should (i) reflect the main polymorphisms of members of the viral taxon; (ii) avoid imbalanced representation, such as over- and under-sampling of some variant groups; and (iii) avoid very distant orthologs that may disturb the alignment by introducing a large number of indels.

Another important issue to bear in mind is that profile HMMs are constructed by calculating position-specific amino acid occurrence probabilities and, in addition, indel probabilities are also computed (1–3). This feature resembles how most phylogenetic methods work, but without the use of evolutionary models. Thus, it is quite straightforward to build models specific to a group of sequence representing a monophyletic clade. Conversely, an MSA containing paraphyletic sequences could hardly yield a profile HMM that is specific to a particular monophyletic clade. Therefore, in an ideal method, viral taxa should first be analyzed by phylogenetic methods and only monophyletic groups associated with taxonomic taxa should be used to build taxon-specific profile HMMs. Orthologous clusters are often obtained from all-versus-all sequence comparisons and rational choices based on best reciprocal hits, resembling to some extent the groups deduced by true phylogenetic methods. However, without the use of appropriate evolutionary models, this assumption can be misleading. The relatively common finding of sequences belonging to different viral families being shared in the same orthologous groups corroborates this fear.

Profile HMMs usually show higher sensitivity or recall rates in comparison to pairwise alignment methods. In fact, more than two decades ago Brenner *et al*. (12) showed that pairwise comparison methods fail to detect half of the relationships between distantly related proteins, sharing 20-30% identity. Conversely, methods based on multiple-sequence profiles detect three times more remote homologs than pairwise alignment methods (14). Skewes-Cox *et al*. (15) showed for a set of best-performing profile HMMs of the `vFam` database that these models are more sensitive than `BLAST` for detecting sequences from distantly related viruses in real metagenomic datasets. However, `BLAST` showed better recall than profile HMMs for more similar viral sequences. This result prompted the authors to propose the combined use of profile HMMs and pairwise alignment methods to obtain more sensitive viral detection and discovery in metagenomic data. Also, the authors found that the main factors contributing to higher recall were the number of sequences used to build the model and the lack of non-viral homologs in the set of viral sequences used to build the models.

The higher sensitivity of profile HMMs, compared to pairwise alignment methods such as `BLAST`, comes with a price. A typical profile HMM is constructed from sequences that are specific to a particular orthology cluster/taxonomic group. Nonetheless, the fact that it is built from an MSA that is representative of many different viral variants makes this model much more tolerant to sequence variability in a similarity search than any individual sequence in a pairwise alignment. The counterpart of such a wide detection range is the high probability of detecting sequences belonging to other taxa. Even though publicly available profile HMMs are commonly associated with annotation and taxonomic information, there is no common recipe on how to use them in viral metagenomic research. The most commonly used approach is to arbitrarily define an e-value cut-off and assume that alignments falling above this canonical value are false positives.

In a recent study, Bzhalava *et al*. (36) used a collection of models from the vFAM database to screen human metagenomic datasets from many different tissues, assuming as viral those sequences presenting an arbitrary e-value of less than 1e-5. More than 500 viral sequences missed by conventional BLAST-based similarity searches were successfully detected, showing the potential impact of using profile HMMs in viral discovery studies. Nevertheless, the authors reported an overall rate of 96% of true positive sequences, but within the *Mimiviridae* family, the true positive rate dropped 3%. According to the authors, this result could be explained by the fact that these viruses encode genes whose orthologs can be found in cellular organisms. This finding exemplifies how risky the use of generalized arbitrary cut-offs can be when surveying wide and diverse groups of viruses. In fact, viral metagenomic datasets can contain variable amounts of nucleic acids from both prokaryotes and eukaryotes contaminants. Many viral genomes contain genes that have orthologous counterparts in cellular organisms. For instance, uracil-DNA glycosylase gene is ubiquitously found in herpesviruses, prokaryotes and eukaryotes. Hence, profile HMMs constructed from viral sequences of this protein can potentially detect genes originating from cellular organisms, rather than from viruses. A large number of other proteins resembling their prokaryotic and/or eukaryotic orthologs can make viral surveys a complex challenge.

To avoid unspecific detection by profile HMMs, Pagnuco *et al*. (37) developed the HMMER Cut-off Threshold Tool (HMMERCTTER), an interesting approach in which a user-provided phylogeny is used as input for the construction of profile HMMs for each protein cluster, assigning specific cut-off thresholds. TABAJARA program (38) goes a step further, identifying informative sites that are conserved in all orthologous sequences belonging to a given clade (synapomorphies) or specific to orthologs of a particular taxon or subgroup of sequences (autapomorphies), thus resembling phylogenetic reconstruction methods. Next, the program determines the most informative regions of the MSA; that is, those stretches that are rich in informative sites. By using only preselected regions of the MSA, TABAJARA avoids protein domains that could result in cross-specificity those regions. Finally, like HMMERCTTER, TABAJARA performs a series of validation tests to discard non-specific models and establishes cut-off scores that allow to obtain a good balance between specificity and recall rates.

RATIONAL DESIGN OF PROFILE HMMS

As stated, profile HMMs should ideally find viral sequences without detecting non-viral orthologs and, additionally, should present viral-taxon specificities. To meet such requirements, a novel approach must be taken for the design and use of such models. As previously commented, the publicly available viral profile HMMs are built from full-length proteins. This strategy implies that very divergent regions containing a large number of gaps are used together with highly conserved stretches of sequences. The former feature introduces noise into the models, while the latter makes the model suitable for detecting a wide range of taxa, but not for discriminating lower taxonomic levels of viral groups. A possible

way to overcome these pitfalls was developed in an integrated set of algorithms implemented on TABAJARA, a publicly available tool for the rational design of profile HMMs (38).

Figure 1 depicts a summarized diagram of TABAJARA's pipeline. Starting from a multiple sequence alignment (MSA), TABAJARA is able to find blocks that are either conserved across all sequences (Conservation execution mode) or discriminative for two specific groups of sequences (Discrimination execution mode). For the identification of regions conserved in all protein sequences of an MSA, TABAJARA implements an information-theoretic algorithm (39) based on Jensen–Shannon divergence method (40) to estimate sequence conservation across all sequences and assign position-specific scores along the entire MSA. To find group-discriminative blocks, TABAJARA uses a combination of Mutual Information (41, 42) and Sequence Harmony (43, 44). Once position-specific scores are determined, the program uses a sliding window to screen the whole alignment and delimit top-scoring regions. This is a particularly interesting feature if regions highly specific to a particular subset of sequences are sought. The program automatically extracts the selected alignment blocks, discards identical sequences, eliminates gap-only columns, and builds the corresponding profile HMMs.

After building the profile HMMs, TABAJARA executes a series of validation steps to discard models that do not meet a set of quality criteria. The models are submitted to similarity searches using `hmmsearch` program from HMMER package (45) against all sequences of the training set. TABAJARA then inspects the results and checks whether they fulfill the quality control criteria defined by the user. When executed in Conservation mode, TABAJARA determines for each profile HMM whether a minimum percentage of sequences are successfully detected. Finally, the program inserts a cut-off score tag in the profile HMM's header, using a value corresponding to 80% of the score obtained for the last hit of the training set. When running in Discrimination mode, TABAJARA also validates the models following a set of criteria. Given an MSA containing two groups of sequences, TABAJARA implements a heuristic to create cut-off scores optimized for each profile HMM to discriminate the group of interest with high specificity while maintaining a good recall. Finally, using the assigned cut-off values of each model, TABAJARA verifies whether the percentage of detected sequences of the chosen group meet a minimum sensitivity value.

Typical profile HMMs generated by TABAJARA are constructed from short, selected regions varying from 20 to 60 bp, and represent specific signatures of different protein families or viral groups. Such short models, named Minions, show a lower recall than profile HMMs built from full-length protein sequences, but with much higher specificity. To overcome this limitation, multiple Minions can be used in combination to increase recall without sacrificing specificity. Finally, TABAJARA also implements the construction of full-length models. In this case, models are built from the entire MSA and submitted to similar validation steps. Once appropriate cut-off scores are assigned, such models can present, at least for some viral taxa, specificity rates similar to those obtained for Minions. Nonetheless, the use of full-length models to screen sequencing databases can be tricky, since cut-off scores are calculated using full-length protein sequences from the training sets, which tend to yield long alignment blocks with relatively high alignment scores. When using such long models to screen short-read sequencing

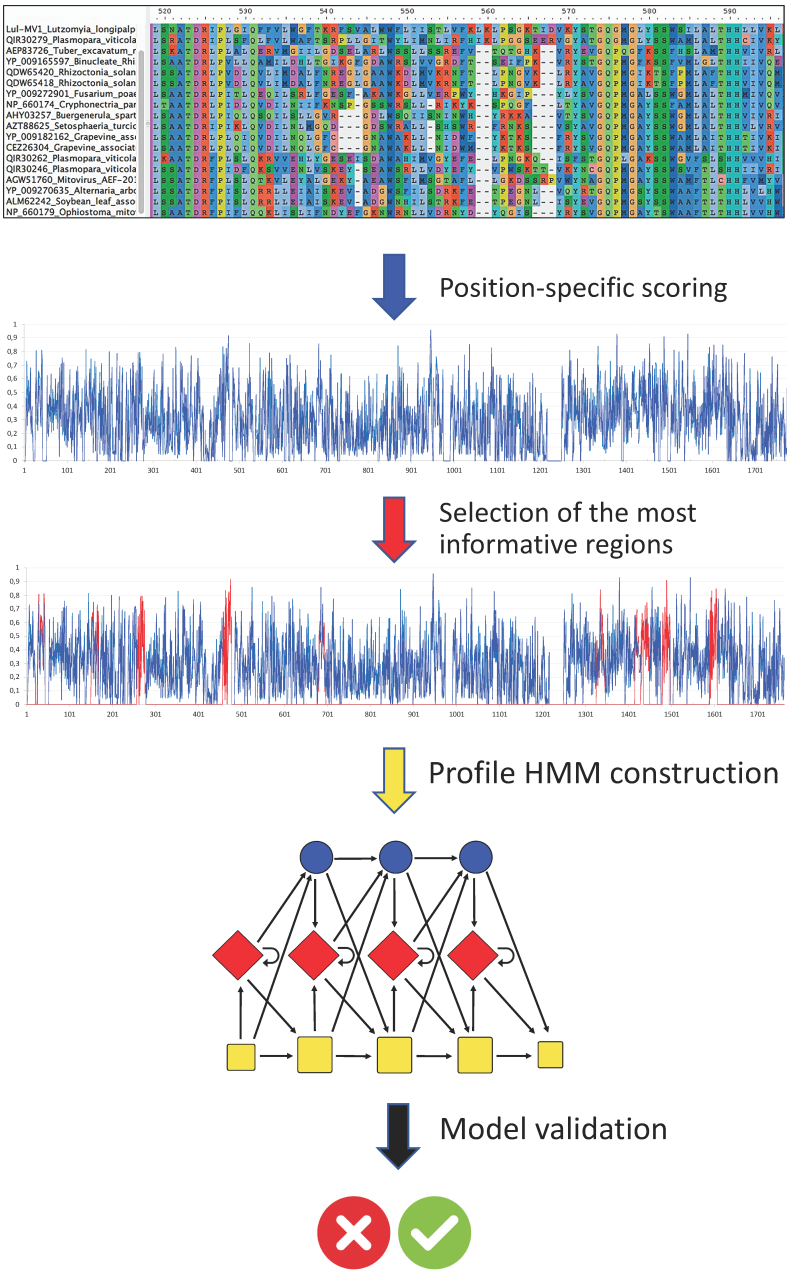

Figure 1. Pipeline of TABAJARA program. TABAJARA uses a multiple sequence alignment as an input training set. In Conservation mode, alignment blocks are selected for regions conserved in all sequences, using Shannon entropy for nucleic acid and Jensen-Shannon divergence for protein sequences. In Discrimination mode, TABAJARA uses a combination of Mutual Information and Sequence Harmony to assign position-specific scores and select the most discriminative regions. Selected alignment blocks are extracted and used to build profile HMMs with `hmmbuild` program, which in turn are submitted to validation tests and stored.

datasets, no matter how good the obtained alignments are, they will never present very high scores due to the limit imposed by the length of the sequencing reads.

TABAJARA is publicly available (https://github.com/gruberlab/tabajara [accessed on 12 Dec 2020]) and is fully documented. The program's site provides an extensive tutorial and datasets that illustrate how to produce conserved and discriminative profile HMMs. Real-life metagenomic datasets, configuration files and a step-by-step tutorial are provided for two viral groups: phages of the *Microviridae* family and eukaryotic viruses of the *Flavivirus* genus. The tutorial covers the design of conserved models capable of detecting any member of these viral groups or, alternatively, discriminative models that specifically detect *Alpavirinae* and *Gokushovirinae* subfamilies (*Microviridae*) or different viral species belonging to the *Flavivirus* genus, such as *Dengue virus*, *Zika virus* and *Yellow fever virus*.

SCREENING SEQUENCING DATASETS WITH PROFILE HMMS

Once profile HMMs with different specificities and detection ranges are available, it becomes feasible to interrogate genomic and metagenomic sequencing datasets for specific protein coding genes or to search for all known viruses. Small or very large collections of models can be used for this task. Some studies employing viral profile HMMs to screen metagenomic data have been reported in the literature (36). To discriminate viral from non-viral sequences, as well as to classify sequences into viral taxonomic groups, the most common solution is the use of arbitrary scores or e-value cut-offs in the `hmmsearch` execution. Such a common recipe is certainly capable of delivering reasonable results, but from a biological viewpoint is far from modeling the different evolutionary rates observed in distinct viral families. For example, cut-off scores effective for the discrimination of dsDNA viral families may be too restrictive for the differentiation of RNA viruses due to the much higher divergence rate observed in the latter group (4–7). Also, the use of common arbitrary cut-off e-values does not take into account the variable size of the sequence datasets, a feature that directly impacts this measure. Finally, profile HMMs are constructed from protein sequences of different lengths, influencing the maximum size of alignment blocks that can theoretically be obtained. This is even more critical when using profile HMMs constructed on top of short protein sequences, as is the case with Minions.

As discussed, TABAJARA allows the construction of profile HMMs from either short or full-length sequences, with recall and specificity governed by cut-off scores automatically customized for each model. Each discriminative profile HMM is validated using a training set composed of sequences of the viral taxonomic group of interest and its sister groups and, therefore, the assigned cut-off scores reflect the evolutionary rates of these specific taxa. A large-scale use of profile HMMs for similarity searches, with customized cut-off scores, can easily be performed with `HMM-Prospector`, a publicly available program (https://github.com/gruberlab/hmmprospector [accessed on 12 Dec 2020]). The workflow of the program (Figure 2) starts with the conversion of input

data, when necessary, from FASTQ (a format produced by most sequencing platforms) into FASTA format, and then the 6-frame conceptual translation to protein sequences. If the profile HMMs contain cut-off scores inserted in their header (as produced by `TABAJARA` – see previous section), `HMM-Prospector`

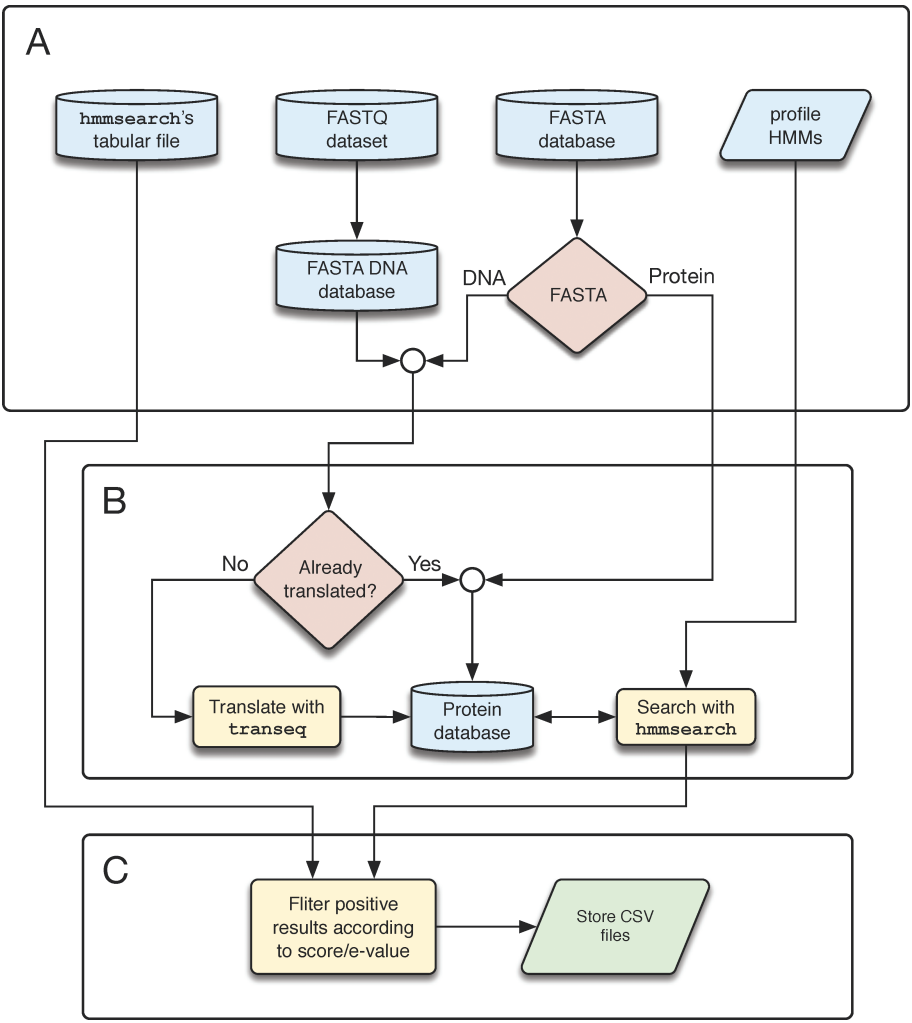

Figure 2. Workflow of `HMM-Prospector` program. The input consists of a profile HMM file and a dataset in either FASTQ or FASTA (DNA or protein sequences) formats. A pre-run `hmmsearch`'s tabular result file is also accepted **(A).** If necessary, `HMM-Prospector` invokes `transeq` to translate the nucleotide sequences into the six possible reading frames. Profile HMMs are then used as queries in similarity searches against the translated dataset using `hmmsearch` **(B).** In the next phase **(C),** `HMM-Prospector` lists all sequences containing positive results, according to user-defined cut-off values (score or e-value) and stores all results into CSV spreadsheet files.

invokes the `hmmsearch` program to perform each similarity search using these custom values. This produces searches optimized for every model, maximizing both sensitivity and specificity. As a final execution step, `HMM-Prospector` produces tabulated files that can be imported into any spreadsheet program. Among other results, these files list all tested profile HMMs with the respective number of positive reads. Thus, a single run with a batch of profile HMMs can unveil which protein sequences are encoded by dataset reads, as well as determine which viral groups are present in the metagenomic sample. `HMM-Prospector` is provided with comprehensive documentation, including a tutorial with an accompanying dataset to perform a survey of *Microviridae* phages using profile HMMs against a metagenomic dataset derived from virus-like particles isolated from human fecal samples (46).

USING PROFILE HMMS FOR TARGETED SEQUENCE RECONSTRUCTION

Metagenomic assembly is a complex and challenging task due to the heterogeneity and abundance of viral communities that often result in many fragmented assemblies and the potential risk of creating chimeric sequences, among other pitfalls. Also, the choice of assembly software can drastically impact the final results (8, 47). If the purpose of the study is to pursue specific viral groups/targets, rather than conducting a comprehensive survey of the virome, then a more sensible approach is to perform a target-specific assembly (3). A seed-driven progressive assembly algorithm was developed and implemented in the `GenSeed` program (48) and, some years later, in other programs applied to the assembly of specific viral genomes (49). In this method, a short nucleotide or protein sequence is used to recruit reads from a sequencing dataset and these reads are then assembled. Short end sequences of the resulting contigs are then extracted and used as extension seeds in an iterative process that generates progressively longer contigs with each cycle. More recently, `GenSeed-HMM` implemented the use of profile HMMs as seeds and a case study showed that this approach could be used successfully for sequence reconstruction and viral discovery from metagenomic data (3, 50). The main advantage of using profile HMMs to recruit reads is the possibility of specifically nucleating multiple growing contigs representing distinct viral genomes each. Like the `HMM-Prospector` program, `GenSeed-HMM` is also capable of using cut-off scores inserted into the headers of the profile HMMs, thus allowing a high specificity in the process of read recruitment. Thus, viral sequences from a metagenomic dataset can be easily reconstructed in a specific way for any taxon, as long as effective profile HMMs are available. This method can also be extended to non-viral sequences such as plasmids, organellar genomes and gene families of cellular organisms, among other targets. `GenSeed-HMM` is publicly available (https://sourceforge.net/projects/genseedhmm/ [accessed on 12 Dec 2020]) and is provided with an instruction guide and tutorial using the dataset described for the progressive assembly of *Alpavirinae* phage genomes (50).

FINDING MULTIGENE ELEMENTS IN CELLULAR ORGANISMS WITH PROFILE HMMS

Many different multigene regions are potentially interesting targets for comparative studies, including prophages, endogenous retroviruses, transposable elements, operons, etc. Nevertheless, some of these entities often present gene copies outside their specific chromosome location, hampering the distinction between element genes and sparse gene copies, especially in partially assembled genomes. Also, if divergence rate is high, as is usually the case with prophages, similarity searches may not detect evolutionarily distant sequences. Some solutions to characterize multigene elements were reported in the literature using `BLAST` similarity searches (51) and profile HMMs (52). As previously discussed, profile HMMs can detect remote homologs with higher sensitivity than conventional pairwise alignment methods. Detection of multigene elements would also benefit if the genome screening included several genes within a defined syntenic context. An approach integrating the use of profile HMMs to interrogate assembled sequencing datasets, allowing for the identification of multigene elements based on gene composition and order, was implemented in the software `e-Finder`. Figure 3 shows a diagram of the program's processing steps. Single or multiple profile HMMs can be used for each protein encoded in the element. In the first step, `transeq` program is invoked to perform a 6-frame translation of all input sequences. Next, `e-Finder` executes `hmmsearch` to perform similarity searches of the models against the translated sequences, and then checks whether pre-defined synteny criteria have been met. Each sequence must contain a minimum number of genes within a proper range of intergenic distances. Element sequences are then extracted, their ORFs identified and conceptually translated into full-length protein sequences. In the final phase, `e-Finder` stores all sequences, together with a CSV file listing all elements and respective features. The program `e-Finder` is freely available and is fully documented (https://github.com/gruberlab/efinder [accessed on 12 Dec 2020]).

AN INTEGRATED APPROACH FOR VIRAL RESEARCH USING PROFILE HMMS

Several bioinformatics tools and databases of viral profile HMMs were presented and discussed in the previous sections. Recently, a new generation of software has become available, composed of tools conceived to be components of an integrated solution for the rational design and use of profile HMMs in viral research (Figure 4). The main aspects involved in the rational design of profile HMMs were implemented in the `TABAJARA` program, including a heuristic approach to calculate cut-off scores customized for each model. Profile HMMs with a defined range of viral taxa detection can be used to interrogate genomic and metagenomic sequencing data. For this task, `HMM-Prospector` can use multiple models with the proper cut-off scores to specifically identify viral sequences of known and emerging viruses in assembled or unassembled metagenomic/genomic datasets.

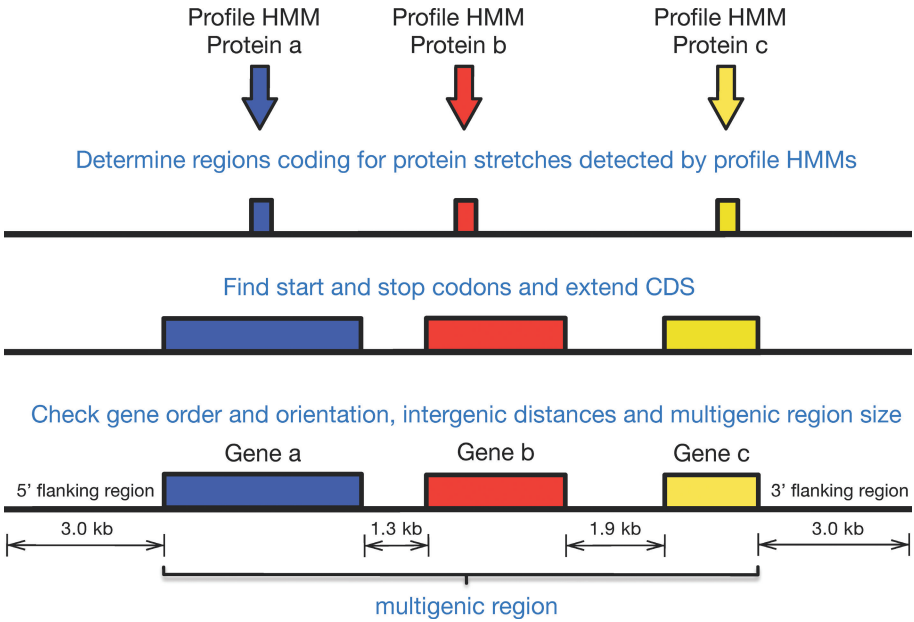

Figure 3. Diagram of the processing steps of `e-Finder` program to find multigene elements in a syntenic context. In this example, the program is used to search a multigene element composed of four genes. First, `e-Finder` performs similarity searches using profile HMMs of each protein against a dataset of translated sequences. The program checks if the sequence contains the minimum number of genes specified by the user. Next, the program identifies the coding sequences of each gene and determines whether the order and distance between the genes meet the set of criteria defined by the user. The element sequences are then extracted together with their respective flanking regions according to a specified length.

Positive datasets for selected profile HMMs can be used for downstream analyses in two alternative analysis pipelines. For unassembled data, `GenSeed-HMM` can use single or multiple profile HMMs to perform seed-driven progressive assemblies and automatically reconstruct the genomes of specific viral taxa. In the case of assembled prokaryotic or eukaryotic genomic data, `e-Finder` can use profile HMMs derived from multiple proteins to identify multigene entities in a proper syntenic context, such as proviruses, transposable elements and operons.

MINIONDB - A DATABASE OF VIRAL PROFILE HMMS

The available repositories of viral profile HMMs share a common characteristic; the sequences used to compose the MSA and build the models are selected from previously generated orthologous groups. While such approaches result in clusters of sequences that usually share functions and a last common ancestor, these sequences do not necessarily belong to the same viral taxa; that is, they may be members of different genera or families. If specific markers for different viral taxa

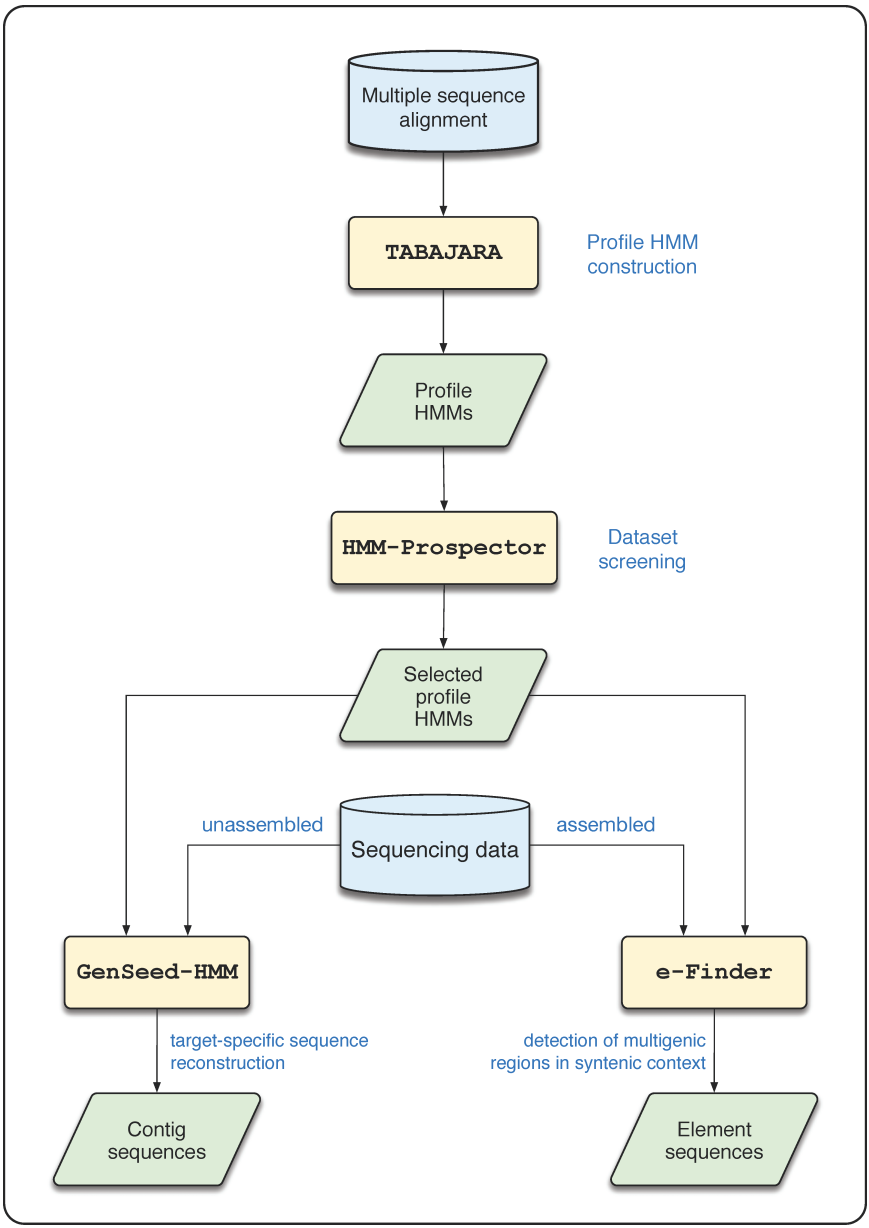

Figure 4. Workflow of the integrated approach for viral bioinformatics studies. Using a multiple sequence alignment as input, `TABAJARA` program can construct profile HMMs using a variety of execution modes. The generated models can be used to screen genomic or metagenomic sequencing data with `HMM-Prospector` program. Models displaying the most relevant results can be used as seeds by `GenSeed-HMM` program to perform a seed-driven progressive assembly using unassembled sequencing data. Alternatively, profile HMMs can be also be used by the program `e-Finder` to identify multigene elements in a syntenic context using assembled data.

are targeted, then the sequences should be selected on a taxonomic basis, rather than orthology.

`Viral MinionDB` is a new repository of profile HMMs covering both prokaryotic and eukaryotic viruses, including short (Minions) and full-length models. All models were created with `TABAJARA` following the guidelines discussed in the previous sections. `Viral MinionDB` was planned assuming some premises: (i) viral taxonomy is dynamic and is continuously and rapidly changing; (ii) official viral taxonomy, as released by the International Committee on Taxonomy of Viruses (ICTV), is distinct from orthology-based clusters, available in the different repositories of viral orthologous groups; and (iii) NCBI Taxonomy is being regularly updated in conformity with up-to-date classifications released by the ICTV.

Thus, instead of running a pipeline for orthology-based clustering, followed by taxonomy mapping, `Viral MinionDB` uses a dump file from the NCBI Taxonomy database to construct a local relational database with all entries, following the original database schema. The NCBI's Taxonomy Browser is updated in real time and the corresponding database dump files are updated hourly. This means that, at any time, a local program can download an updated database dump file from the NCBI's FTP site. With all taxonomic identifiers on hand, viral proteins can be obtained from the NCBI's Identical Protein Groups (IPG) database (https://www.ncbi.nlm.nih.gov/ipg/) using taxon-associated queries. These sets of sequences, selected according to their taxonomic classification, are then aligned and used to construct profile HMMs designed as specific markers for the different viral taxa. `Viral MinionDB` models constructed for a wide range of taxa (viral families) incorporate a higher diversity than models designed for a narrower group of viruses (viral genera). All profile HMMs are built and validated independently using the sequences from the corresponding MSAs as training sets. Also, all profiles' HMMs incorporate custom cut-off scores, a feature that allows to perform searches with optimized stringency.

`Viral MinionDB` was recently released to the public (http://www.bioinfovir.icb.usp.br/miniondb [accessed on 12 Dec 2020]) and its version 1.0 (as for December 2020) contains a collection of 2,415 profile HMMs for prokaryotic viruses (312 full-length and 2,103 short [Minion] models) and 18,334 profile HMMs for eukaryotic viruses (1,173 full-length and 17,161 short [Minion] models). The complete repository covers 27 viral families of prokaryotic viruses and 120 families of eukaryotic viruses. Instructions for using the web interface and downloading the models are available on the web site.

CONCLUSION

Unraveling the dark viral matter (34, 53) is one of the main challenges in biological research. The deluge of data provided by metagenomics has revealed new challenges for viral sequence detection and classification. Traditional viral classification based on morphological and compositional criteria is being rapidly expanded with the incorporation of molecular criteria. Viral taxonomic classification will likely change, but new bioinformatic tools must be developed to easily reflect such changes and permit a fast and reliable classification of viruses. Profile HMMs are gradually gaining space to model sequence diversity within new taxa

and are becoming valuable tools to help researchers to determine the viral content and the diversity of different biomes, as well as for novel virus discovery. Many challenges persist - mainly the development of proper methods to select sequences for profile HMM construction, and, as discussed above, the incorporation of more refined criteria to discriminate positive from negative similarity search results.

The paradigm of diagnosis relies on the fact that detection is achieved by identifying previously known features. For example, the detection and quantification of antibody response depends on the use of specific antigens. Similarly, PCR-based assays amplify a specific target whose flanking sequences are used to design the specific primers of the reaction. This paradigm represents a limitation for viral discovery using metagenomic datasets. In classical virology, novel viruses are identified by their association to clinical symptoms of a disease, characterized by cytopathic effects and particle morphology, and followed by multiplication in cell cultures, or inoculation in animal models to replicate these viruses and reproduce the disease. All these features are not attainable when using metagenomics for viral discovery (13). Since profile HMMs can be constructed with various levels of specificity, covering wide and narrow taxonomic ranges of detection, these models can be incorporated in metagenomic processing pipelines. Narrow-range models can be used to detect and reliably classify currently known viruses, whereas wide-range models are capable of detecting evolutionary distant viruses. Although these tools do not completely break the paradigm of diagnosis, they certainly improve our ability to directly detect the unknown viruses in a "*de novo*" diagnosis. In the coming years, when viral taxonomic classification is likely to be based almost entirely on genomic data (31, 54) using frameworks such as the `GRAViTy` platform (55, 56), programs like `TABAJARA` can become useful tools for large-scale automated production of profile HMMs that will cover all taxonomically defined viral groups. With such models, the entire process of analyzing metagenomic data, as well as detecting emerging viruses, will become much simpler than in the present.

The ability to detect known and unknown viruses more efficiently, using profile HMMs, will impact epidemiological surveillance programs. Critical locations such as hospitals, sewage treatment stations, animal production farms, and environments seasonally colonized by migratory birds are potential targets for such programs. Detection of emerging viruses on such sites may alert health authorities in time to prevent or mitigate the effects of potentially devastating diseases.

Conflict of Interest: The authors declare no potential conflicts of interest with respect to research, authorship and/or publication of this chapter.

REFERENCES

1. Gollery M, editor. Handbook of Hidden Markov Models in Bioinformatics. New York, USA: Chapman and Hall/CRC; 2008. 176 p. https://doi.org/10.1201/9781420011807

2. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov Models in Computational Biology. J Mol Biol. 1994;235(5):1501–31. https://doi.org/10.1006/jmbi.1994.1104
3. Reyes A, Alves JMP, Durham AM, Gruber A. Use of profile hidden Markov models in viral discovery: current insights. Adv Genom Genet. 2017;7:29–45. https://doi.org/10.2147/AGG.S136574
4. Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S. Rapid evolution of RNA genomes. Science. 1982;215(4540):1577–85. https://doi.org/10.1126/science.7041255
5. Drake JW. Rates of spontaneous mutation among RNA viruses. P Natl Acad Sci USA. 1993;90(9):4171–5. https://doi.org/10.1073/pnas.90.9.4171
6. Peck KM, Lauring AS. Complexities of Viral Mutation Rates. J Virol. 2018;92(14):e01031–17. https://doi.org/10.1128/JVI.01031-17
7. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral Mutation Rates. J Virol. 2010;84(19):9733–48. https://doi.org/10.1128/JVI.00694-10
8. Fancello L, Raoult D, Desnues C. Computational tools for viral metagenomics and their application in clinical research. Virology. 2012;434(2):162–74. https://doi.org/10.1016/j.virol.2012.09.025
9. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, et al. A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases. N Engl J Med. 2008;358(10):991–8. https://doi.org/10.1056/NEJMoa073785
10. Brüssow H. The not so universal tree of life or the place of viruses in the living world. Phil Trans R Soc B. 2009;364(1527):2263–74. https://doi.org/10.1098/rstb.2009.0036
11. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402. https://doi.org/10.1093/nar/25.17.3389
12. Brenner SE, Chothia C, Hubbard TJP. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. P Natl Acad Sci USA. 1998;95(11):6073–8. https://doi.org/10.1073/pnas.95.11.6073
13. Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. Curr Opin Virol. 2012;2(1):63–77. https://doi.org/10.1016/j.coviro.2011.12.004
14. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol. 1998;284(4):1201–10. https://doi.org/10.1006/jmbi.1998.2221
15. Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL. Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. PLoS ONE. 2014;9(8):e105067. https://doi.org/10.1371/journal.pone.0105067
16. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45. https://doi.org/10.1093/nar/gkv1189
17. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45(D1):D491–8. https://doi.org/10.1093/nar/gkw975
18. Kristensen DM, Waller AS, Yamada T, Bork P, Mushegian AR, Koonin EV. Orthologous Gene Clusters and Taxon Signature Genes for Viruses of Prokaryotes. J Bacteriol. 2013;195(5):941–50. https://doi.org/10.1128/JB.01801-12
19. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44(D1):D286–93. https://doi.org/10.1093/nar/gkv1248
20. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):D309–14. https://doi.org/10.1093/nar/gky1085
21. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–15. https://doi.org/10.1093/nar/gky1049
22. Bigot T, Temmam S, Pérot P, Eloit M. RVDB-prot, a reference viral protein database and its HMM profiles. F1000Res. 2020;8:530. https://doi.org/10.12688/f1000research.18776.2
23. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. mSphere. 2018;3(2):e00069–18. https://doi.org/10.1128/mSphereDirect.00069-18

24. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2. https://doi.org/10.1093/bioinformatics/bts565
25. Miele V, Penel S, Duret L. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics. 2011;12(1):116. https://doi.org/10.1186/1471-2105-12-116
26. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21(7):951–60. https://doi.org/10.1093/bioinformatics/bti125
27. Chibani CM, Farr A, Klama S, Dietrich S, Liesegang H. Classifying the Unclassified: A Phage Classification Method. Viruses. 2019;11(2):195. https://doi.org/10.3390/v11020195
28. Chibani CM, Meinecke F, Farr A, Dietrich S, Liesegang H. ClassiPhages 2.0: Sequence-based classification of phages using Artificial Neural Networks [Internet]. Bioinformatics; 2019 [cited 2020 Dec 17]. Available from: http://biorxiv.org/lookup/doi/10.1101/558171 https://doi.org/10.1101/558171
29. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res. 2020;49(D1):D764–5. https://doi.org/10.1093/nar/gkaa946
30. Paez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Szeto E, et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Res. 2016;45(D1):gkw1030. https://doi.org/10.1093/nar/gkw1030
31. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15(3):161–8. https://doi.org/10.1038/nrmicro.2016.177
32. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol. 2019;37(1):29–37. https://doi.org/10.1038/nbt.4306
33. Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. Nat Protoc. 2017;12(8):1673–82. https://doi.org/10.1038/nprot.2017.063
34. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536(7617):425–30. https://doi.org/10.1038/nature19094
35. Paez-Espino D, Roux S, Chen I-MA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. Nucleic Acids Res. 2019;47(D1):D678–86. https://doi.org/10.1093/nar/gky1127
36. Bzhalava Z, Hultin E, Dillner J. Extension of the viral ecology in humans using viral profile hidden Markov models. PLoS ONE. 2018;13(1):e0190938. https://doi.org/10.1371/journal.pone.0190938
37. Pagnuco IA, Revuelta MV, Bondino HG, Brun M, ten Have A. HMMER Cut-off Threshold Tool (HMMERCTTER): Supervised classification of superfamily protein sequences with a reliable cut-off threshold. PLoS ONE. 2018;13(3):e0193757. https://doi.org/10.1371/journal.pone.0193757
38. Ibrahim B, Arkhipova K, Andeweg A, Posada-Céspedes S, Enault F, Gruber A, et al. Bioinformatics Meets Virology: The European Virus Bioinformatics Center's Second Annual Meeting. Viruses. 2018;10(5):256. https://doi.org/10.3390/v10050256
39. Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics. 2007;23(15):1875–82. https://doi.org/10.1093/bioinformatics/btm270
40. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory. 1991;37(1):145–51. https://doi.org/10.1109/18.61115
41. Adami C. Information theory in molecular biology. Phys Life Rev. 2004;1(1):3–22. https://doi.org/10.1016/j.plrev.2004.01.002
42. Cover TM, Thomas JA, editors. Elements of information theory. 2nd ed. Hoboken, N.J: Wiley-Interscience; 2006. 748 p. https://doi.org/10.1002/047174882X
43. Feenstra KA, Pirovano W, Krab K, Heringa J. Sequence harmony: detecting functional specificity from alignments. Nucleic Acids Res. 2007;35(suppl_2):W495–8. https://doi.org/10.1093/nar/gkm406
44. Pirovano W, Feenstra KA, Heringa J. Sequence comparison by sequence harmony identifies subtype-specific functional sites. Nucleic Acids Res. 2006;34(22):6540–8. https://doi.org/10.1093/nar/gkl901
45. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011;7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195
46. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature. 2010;466(7304):334–8. https://doi.org/10.1038/nature09199

47. Sutton TDS, Clooney AG, Ryan FJ, Ross RP, Hill C. Choice of assembly software has a critical impact on virome characterisation. Microbiome. 2019;7(1):12. https://doi.org/10.1186/s40168-019-0626-5
48. Sobreira TJP, Gruber A. Sequence-specific reconstruction from fragmentary databases using seed sequences: implementation and validation on SAGE, proteome and generic sequencing data. Bioinformatics. 2008;24(15):1676–80. https://doi.org/10.1093/bioinformatics/btn283
49. Smits SL, Osterhaus AD. Virus discovery: one step beyond. Curr Opin Virol. 2013;3(2):e1–6. https://doi.org/10.1016/j.coviro.2013.03.007
50. Alves JMP, de Oliveira AL, Sandberg TOM, Moreno-Gallego JL, de Toledo MAF, de Moura EMM, et al. GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in Alpavirinae Viral Discovery from Metagenomic Data. Front Microbiol. 2016;7. https://doi.org/10.3389/fmicb.2016.00269
51. Medema MH, Takano E, Breitling R. Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. Mol Biol Evol. 2013;30(5):1218–23. https://doi.org/10.1093/molbev/mst025
52. Abby SS, Néron B, Ménager H, Touchon M, Rocha EPC. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. PLoS ONE. 2014;9(10):e110726. https://doi.org/10.1371/journal.pone.0110726
53. Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res. 2017;239:136–42. https://doi.org/10.1016/j.virusres.2017.02.002
54. Simmonds P, Aiewsakun P. Virus classification - where do you draw the line? Arch Virol. 2018;163(8):2037–46. https://doi.org/10.1007/s00705-018-3938-z
55. Aiewsakun P, Adriaenssens EM, Lavigne R, Kropinski AM, Simmonds P. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. J Gen Virol. 2018;99(9):1331–43. https://doi.org/10.1099/jgv.0.001110
56. Aiewsakun P, Simmonds P. The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification. Microbiome. 2018;6(1):38. https://doi.org/10.1186/s40168-018-0422-7