

RVDB-prot, a reference viral protein database and its HMM profiles [version 2; peer review: 2 approved]

Thomas Bigot, Sarah Temmam, Philippe Pérot, Marc Eloit

▶ **To cite this version:**

Thomas Bigot, Sarah Temmam, Philippe Pérot, Marc Eloit. RVDB-prot, a reference viral protein database and its HMM profiles [version 2; peer review: 2 approved]. F1000Research, 2020, 8, pp.530. 10.12688/f1000research.18776.2 . pasteur-02946473

HAL Id: pasteur-02946473

https://pasteur.hal.science/pasteur-02946473

Submitted on 23 Sep 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Check for updates

DATA NOTE

REVISED RVDB-prot, a reference viral protein database and its HMM profiles [version 2; peer review: 2 approved]

Thomas Bigot[1], Sarah Temmam[2], Philippe Pérot[2], Marc Eloit[2,3]

[1]Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France
[2]Pathogen Discovery Laboratory, Institut Pasteur, Paris, 75015, France
[3]École Nationale Vétérinaire d'Alfort, Maisons-Alfort, 94700, France

v2 **First published:** 23 Apr 2019, **8**:530
https://doi.org/10.12688/f1000research.18776.1
Latest published: 07 Sep 2020, **8**:530
https://doi.org/10.12688/f1000research.18776.2

Abstract

We present RVDB-prot, a database corresponding to the protein equivalent of the nucleic acid reference virus database RVDB. Protein databases can be helpful to perform more sensitive protein sequence comparisons. Similarly to its homologous public repository, RVDB-prot aims to provide reliable and accurately annotated unique entries, while including also an Hidden Markov Model (HMM) protein profiles database for distant protein searching.

Keywords

virus, genomes, proteins, hmm, clusters, annotations, database

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 07 Sep 2020	✓ report	✓ report
version 1 23 Apr 2019	? report	? report

1. **Philippe le Mercier**, Swiss Institute of Bioinformatics, Geneva, Switzerland

2. **Guy Perriere**, Universite Claude Bernard - Lyon 1, Villeurbanne, France

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Thomas Bigot (thomas.bigot@pasteur.fr)

Author roles: Bigot T: Conceptualization, Data Curation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Temmam S**: Conceptualization, Data Curation, Writing – Review & Editing; **Pérot P**: Conceptualization, Data Curation, Writing – Review & Editing; **Eloit M**: Conceptualization, Data Curation, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2020 Bigot T *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Bigot T, Temmam S, Pérot P and Eloit M. **RVDB-prot, a reference viral protein database and its HMM profiles [version 2; peer review: 2 approved]** F1000Research 2020, **8**:530 https://doi.org/10.12688/f1000research.18776.2

First published: 23 Apr 2019, **8**:530 https://doi.org/10.12688/f1000research.18776.1

REVISED **Amendments from Version 1**

This second version takes into account the comments of the reviewers. We also updated the metrics and the files available to reflect the version 19.0 of RVDB. The code of the pipeline itself was updated. In the introduction, we exemplified usages of the database, introduced UniProtKB virus database, and explained why the curation of RVDB is so important. In the methods, we detailed where the protein sequences coded by nucleic sequences are found. We specified that we used the default parameters of two programs used by the pipeline (Silix and HMMER). Regarding the database annotation, we clarified how each cluster is described, and we added a new feature: annotation keywords are not only gathered from sequences descriptions, but now also from PFAM clusters matching to the cluster sequences.

Any further responses from the reviewers can be found at the end of the article

Introduction

Sequence assignation often uses similarity criteria to infer homology, and hence taxonomy and / or protein function. In order to search for this, similarity, reliable, accurate and comprehensive databases are required. When trying to characterize sequences present in a metagenomics sample, searching first for related sequences in a viral database can lead to identify rapidly a known virus (high identity between the query sequence and the one in the database), or identify potential new species (low identity with any known sequence). Such hits must be further characterized on more comprehensive databases to increase the robustness of taxonomic assignations.

In the specific field of viruses, several solutions are available but their ability to provide valid results is highly dependent on the goal of the study and on the available computer resources. Using a database with a high number of sequences, such as NCBI nr/nt may seem appropriate, but it implies an increased computation time and annotation quality is not always optimal. Similarly, UniProtKB[1] contains numerous viral sequences (4 497 049 in total, including 17 008 (0.38%) reviewed) that could, as for NCBI/nr, increase computation time when thousands of sequences have to be analyzed concomitantly, which is routinely practiced in metagenomics analyses. RefSeq, on the other hand, is generally better curated but contains only full-length genomes, which reduces the diversity of available sequences, and also rarely includes the latest discoveries. RefSeq contains 13 180 virus sequences. Other specialized databases provide only specific groups of taxa for specific purposes, for instance, virus families responsible for infectious diseases like HIV or influenza viruses.

Thus, the need for better, well-annotated and comprehensive public viral database that can be used for the identification of viruses by high-throughput sequencing led Goodacre *et al.* to propose their Reference Viral DataBase (RVDB)[2]. This database consists of a collection of all currently known viral genomes and virus-related nucleic sequences retrieved from NCBI/nr or RefSeq and includes a specific, both manual and computational reviewing process, as well as four updates of the contents per year. The reviewing process eliminates a great quantity of unwanted non-viral sequences like: cloning vectors, endogenous sequences, sequences that were wrongly annotated as virus but were actually of cellular origin, etc. This high level of curation makes RVDB quite attractive for the virology research community and in fact, in June 2020, version 19.0 was released.

Since viral genomes mainly consist of coding sequences, the need for an equivalent reference database that provides the protein version of these sequences may prove quite advantageous.

Indeed, protein sequences are useful when searching for distant homologs: their substitution rates are much lower than nucleic sequences. Additionally, proteins can also be efficiently clustered according to their similarity, and the resulting clusters can then be used to build Hidden Markov Model (HMM) profiles in order to identify more evolutionary distant proteins. In fact, programs like HMMER[3] allow the building of HMM profiles from a multiple sequence alignment of proteins. This profile can then help recognizing proteins based on complex position-specific models of sequence conservation and evolution, and it does so in a more accurate way than if a classic sequence alignment is used.

Therefore, we propose a protein sequence version of RVDB whose update will be synchronized with the original nucleotide RVDB release. Here we describe the conversion from the nucleotide version of RVDB to the protein version RVDB-prot, as well as the clustering process leading to the HMM profiles.

Methods

Conversion from RVDB nucleic database to RVDB-prot

The current version of RVDB, v19.0[4] consists of a collection of 3 084 319 nucleic sequences[2]. The accession numbers were extracted in order to gather the corresponding database entries in Genbank format. From these entries, the corresponding coding domain protein sequences, description, and protein accession numbers were automatically recognized and copied into the protein collection. The process relies on the amino-acid sequences and information provided initially in the nucleic entry annotations. The resulting protein file contains the nucleic sequence reference, for traceability purposes. The sequence names are formatted in the following way:

`>acc|<p_bank>|<p_acc>|<n_bank>|<n_acc>|<descr[sp]>`, where:

`p_bank` is the bank in which the protein can be found

`p_acc` is the accession number corresponding to the protein sequence

`n_bank` is the bank in which the original nucleotide sequence was found

`n_acc` is the original information found in the nucleic database

`descr` is the description of the protein sequence as found in the database entry

`sp` is the species name.

This process produces a 4 705 359 protein sequence file.

Generation of HMM profiles

The HMM generation rationale was inspired from vFam (the database of HMM profiles built from all the viral proteins present in RefSeq, discontinued from 2014)[5], but was entirely re-coded as a Snakemake pipeline[6], using different tools for some key steps (clustering, alignment). The proteins sequences were clustered with a 100% identity criterion to remove duplicates using CD-Hit 4.7.0[7]. Then, the sequences were processed using Blast 2.2.26[8] performing an all-against-all comparison. These comparisons allowed Silix 1.2.6[9] (using default parameters) to define clusters of sequences according to their similarity. This step produced a text file in which each sequence was associated to one cluster. The information of each cluster (containing at least four sequences) was transformed into a fasta file containing all the sequences within the cluster. Then, sequences were aligned using Mafft 7.023[10] in auto mode. The multiple sequence alignments were processed by HMMER 3.2.1[3] (hmmbuild, using default parameters) in order to obtain the HMM profiles. The HMM profiles were finally grouped into a single file.

Annotation of HMM profiles

A cluster is defined as a set of sequences, among which each sequence is characterized by its taxonomy (*i.e.* a virus species) and eis associated with a description of its putative function, when it is known. In order to describe the different clusters, these information and other indicators (such as the cluster length and number of sequences) are combined into an annotation database, in SQLite format. The schema of this database is shown in Figure 1.

The first type of data associated to a cluster is a set of keywords describing the putative function of the proteins present in a given cluster. These keywords correspond to the union of all names of the significant sequences found in PFAM[11] (with --cut_ga parameter which tells HMMER to trust the cutoff defined by PFAM) using all the sequences of the cluster as queries, weighted according to their frequencies, and excluding trivial words. We also produce a complementary word frequency count using sequence descriptions. These keywords are stored

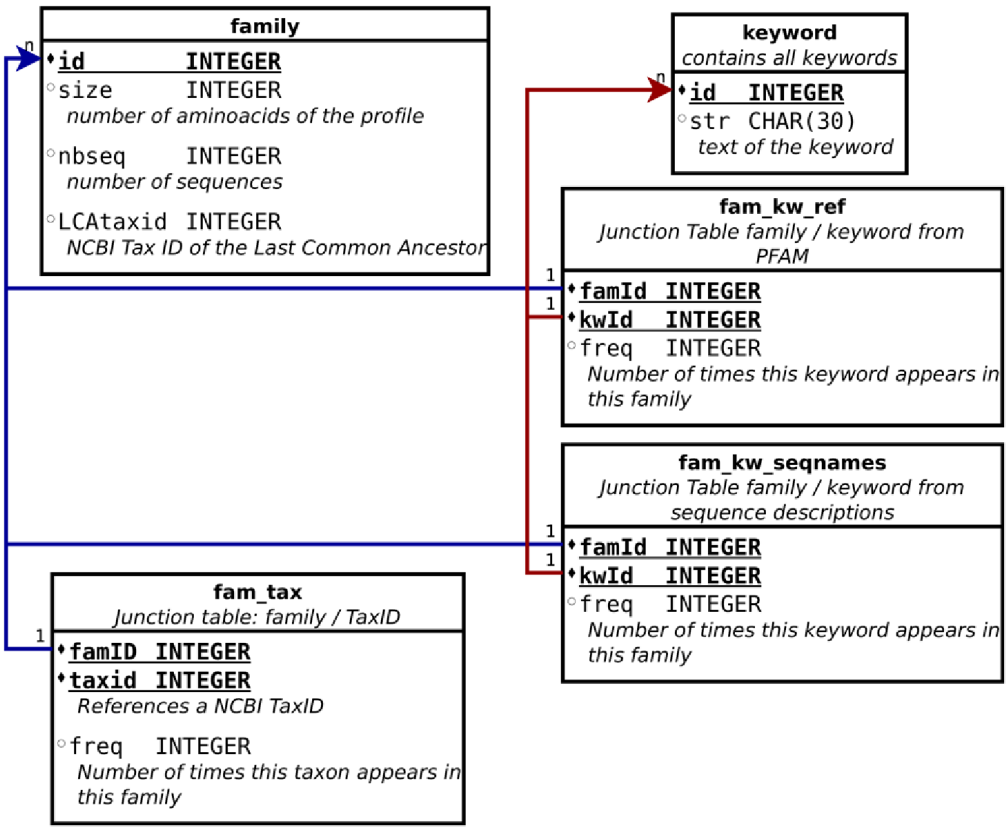


Figure 1. Schema of the annotation database.

separately from the PFAM ones. Despite the fact that sequence descriptions can be vague or inaccurate, they are a good fallback in case the cluster had no match with any PFAM one.

Here is an example of different keywords, using annotations from the cluster number 1, containing 18 sequences; the keywords and their frequencies from PFAM, are: RNA(58), viral(18), dependent(18), polymerase(18), helicase(17), cysteine(11), picorna(11); the first ten keywords from the sequence descriptions are: virus(17), protein(13), hypothetical(10), like(9), picorna(9), polyprotein(5), RNA(4), Wenzhou(4), Beihai(4), non-structural(2). Altogether, these keywords allow to describe a cluster composed of RNA-dependent RNA polymerase of picorna-like viruses. The complementarity of these two annotations is well illustrated here since the simple list of keywords would not have allowed to identify the function of this cluster (here the viral polymerase) without PFAM.

In addition to the protein description, the database stores the virus taxonomy associated to all the taxa, referring to tNCBI TaxIDs. For each cluster, the taxonomic information is summarized by a Last Common Ancestor (LCA) that corresponds to the taxon in the tree of life to which all the sequence taxa belong; this LCA can be close to the root of the tree (Viruses), but is usually specific to a family.

Finally, the database also provides the length (number of amino acids of the multiple sequences alignment) and the number of sequences in each cluster.

This database is available in SQLite format, and to provide more direct access, flat text files are proposed. A text file for each cluster, identified with its cluster number, contains all the information related to it.

Software availability

The different steps explained above are performed using a Snakemake pipeline[6], available at Institut Pasteur's Gitlab.

- Pipeline available from https://gitlab.pasteur.fr/tbigot/rvdb-prot/.
- Archived source code at time of publication: https://doi.org/10.5281/zenodo.4001989[12]
- Licence: GNU GPL v3.0

Several tools are needed to run the pipeline, including: Python, Mafft, Golden, HMMER, Snakemake, Silix, Blast+. The versions of these tools compatible with the pipeline are listed in the README file.

Data availability

Underlying data

Database files are available at https://rvdb-prot.pasteur.fr/. Release 19.0 described in this manuscript is also available from Zenodo.

Zenodo: U-RVDBv19.0 https://doi.org/10.5281/zenodo.4002051[4].

This project contains the following underlying data:

- U-RVDBv19.0-prot.fasta (fasta file containing protein features of the original database: -prot.fasta)
- U-RVDBv19.0-prot.fasta-prot.hmm (the HMM profiles, generated with and for hmmer 3.2.1 (from 2019, 3.1b2 before))
- U-RVDBv19.0-prot.fasta-prot-hmm.sqlite (SQLite db containing annotations (please find a documentation below))
- U-RVDBv19.0-prot.fasta-annot.txt (a directory of annotations with plain text files (one per protein family))

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Table 1 shows some summary metrics for the entries of this release and the different resources.

Table 1. Metrics for release 19.0.

Nucleic sequences	RVDB	3 084 319
Proteins	RVDB-prot	4 705 359
Unique proteins	RVDB-prot	674 970
Clusters	RVDB-prot HMM	13 201

Updates are manually curated each time a new release of the main database (nucleic RVDB) is announced, i.e., four times a year. The following older versions are also available online: 18.0 (2020–03), 17.0 (2019–11), 16.0 (2019–06), 15.1 (2019–02), 14.0 (2018–09), 13.0 (2018–06), 12.2(2018–03), 11.5 (2017-–0), 10.2 (2017–04).

Usage HMMER can be used to search for all profiles in a fasta sequence file (sequences.fasta): `hmmsearch U-RVDBv15.1-prot.fasta-prot.hmm sequences.fasta > result.out`. Additional options are available in HMMER User's Guide.

Acknowledgements

We would like to thank Peter Skewes-Cox, Jr., author of vFAM database for kindly providing his scripts which were an inspiration for the earlier versions of RVDB-prot. We thank Natalia Pietrosemoli for her help in the editing of the manuscript. This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris.

References

1. UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Goodacre N, Aljanahi A, Nandakumar S, *et al.*: **A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection.** *mSphere.* 2018; **3**(2): pii: e00069-18.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS Comput Biol.* 2011; **7**(10): e1002195.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Bigot T, Temmam S, Pérot P, *et al.*: **RVDB-prot, a reference viral protein database and its HMM profiles.** *Zenodo.* Fileset. 2020.
http://www.doi.org/10.5281/zenodo.4002051
5. Skewes-Cox P, Sharpton TJ, Pollard KS, *et al.*: **Profile hidden Markov models for the detection of viruses within metagenomic sequence data.** *PLoS One.* 2014; **9**(8): e105067.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Köster J, Rahmann S: **Snakemake--a scalable bioinformatics workflow engine.** *Bioinformatics.* 2012; **28**(19): 2520–2522.
PubMed Abstract | Publisher Full Text
7. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics.* 2006; **22**(13): 1658–1659.
PubMed Abstract | Publisher Full Text
8. Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text
9. Miele V, Penel S, Duret L: **Ultra-fast sequence clustering from similarity networks with SiLiX.** *BMC Bioinformatics.* 2011; **12**(1): 116.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Katoh K, Misawa K, Kuma K, *et al.*: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res.* 2002; **30**(14): 3059–3066.
PubMed Abstract | Publisher Full Text | Free Full Text
11. El-Gebali S, Mistry J, Bateman A, *et al.*: **The Pfam protein families database in 2019.** *Nucleic Acids Res.* 2019; **47**(D1): D427–D432.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Bigot T: **RVDB-prot pipeline v19.0.1 (Version 15.1.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4001989

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 15 September 2020

https://doi.org/10.5256/f1000research.29158.r70931

© **2020 Perriere G.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Guy Perriere**
Laboratoire de Biométrie et Biologie Evolutive, CNRS, UMR5558, Universite Claude Bernard - Lyon 1, Villeurbanne, 69622, France

I have no other comments to add as the authors adressed all the concerns I raised in my previous review.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Phylogeny, molecular evolution, comparative genomics, sequence databases

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 07 September 2020

https://doi.org/10.5256/f1000research.29158.r70930

© **2020 le Mercier P.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Philippe le Mercier**
Swiss-Prot Group, CMU, Swiss Institute of Bioinformatics, Geneva, Switzerland

No further comments.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Virus proteomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 20 May 2019

https://doi.org/10.5256/f1000research.20570.r47713

© 2019 Perriere G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Guy Perriere
Laboratoire de Biométrie et Biologie Evolutive, CNRS, UMR5558, Universite Claude Bernard - Lyon 1, Villeurbanne, 69622, France

This is an interesting resource that can be of use for people dealing with comparative genomics in viruses. There are some points that need to be clarified before this paper can be indexed though.

1. In order to ease reproducibility, the parameters used for the different programs (e.g. HMMER, SiLiX) of the pipeline should be provided.
2. Why is it required to locally translate the Coding DNA Sequences (CDS) from the original RVDB nucleotide database instead of downloading them from the resource?
3. I have a question for the taxonomic assignation to the Last Common Ancestor (LCA) when building the clusters. How are handled the possible contradictions within a cluster? More exactly, what is done exactly if sequences that belong to distantly related taxa are clustered together? If a strict LCA rule is applied, then it would be possible to have a really inprecise assignation (something like "virus" and that's it).

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Phylogeny, molecular evolution, comparative genomics, sequence databases

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 26 Aug 2020

Thomas Bigot, Institut Pasteur, Paris, France

We would like to thank the Reviewer. Please find below our line-by-line responses.

In order to ease reproducibility, the parameters used for the different programs (e.g. HMMER, SiLiX) of the pipeline should be provided.

Done. Parameters are now specified in the new version of the manuscript. Actually, for both of these programs, we use default parameters.

Why is it required to locally translate the Coding DNA Sequences (CDS) from the original RVDB nucleotide database instead of downloading them from the resource?

We have clarified this point in the new version: actually, we use translations provided in the entry of each nucleic sequence. They are provided in the raw data of the original database (Genbank, RefSeq) along with the accession number. What we do amounts to retrieve all protein sequences
corresponding to a nucleic sequence from protein database with accession numbers, but doing it directly from the nucleic database is faster.

I have a question for the taxonomic assignation to the Last Common Ancestor (LCA) when building the clusters. How are handled the possible contradictions within a cluster? More exactly, what is done exactly if sequences that belong to distantly related taxa are clustered together? If a strict LCA rule is applied, then it would be possible to have a really inprecise assignation (something like "virus" and that's it).

Indeed, we use naïve LCA assignation, and it can lead to imprecise assignation (some clusters can be tagged as Viruses). As we do not have other information about the cluster we characterize, we chose not to avoid this possibility. We have added a precision about this case in the new version of the manuscript: “For each cluster, the taxonomic information is summarized by a Last Common Ancestor (LCA) that corresponds to the taxon in the tree of life to which all the sequence taxa belong; this LCA can be close to the root of the tree (Viruses), but is usually specific to a family.”

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 May 2019

https://doi.org/10.5256/f1000research.20570.r47715

© 2019 le Mercier P. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Philippe le Mercier
Swiss-Prot Group, CMU, Swiss Institute of Bioinformatics, Geneva, Switzerland

In this article, the authors present a RVDB-prot, a reference viral protein database and its HMM profiles. The purpose of this approach is providing a complete reference database of viral proteins to identify new sequences. Their database is based on nucleotide Reference Viral Database (RVDB). The rationale of this work is that protein sequences can be better than nucleotides for searching distant homologs.

In brief, their approach was as follows: RVDB database was converted to proteins, thereby creating a new dataset of 3,899,699 proteins. The protein were clustered, and these clusters used to create HMM. Words frequently present in sequence names of a cluster were used to annotate HMM profiles. The software, pipeline and final database are all available.

The final data are of good quality, will hopefully be maintained along with RVDB and offer a new approach for protein virus reference.

While the article is well written and the method is well described, there are a number of issues that need to be addressed:

- The database is described as facilitating sequence assignation. This seems a bit vague, a sentence describing possible applications could help.
- The introduction may describe better the current state of research in the field. UniProtKB should be cited in “existing databases” for viral proteins, and authors may add citations of its use in virus detection. (ex: UniRef90 used with success to create synthetic human virome [1] (PMID: 26045439)). This would also highlight new potential applications for RVDB-prot.
- UniProtKB provides data for 3,972,271 viral proteins, a bit more than RVDB-Prot (3,899,699). RVDB-prot data is based similarity gathering of sequences with viral RefSeq, which has the advantage to ignore any taxonomical issues. On the other hand, many RefSeq are provisional, and those are not free or errors. The authors may discuss the advantage of their method over existing protein dataset.
- Similarly, UniRef90 contains 577,105 clusters of proteins, which could be compared to the 489,207 unique proteins of RVDB-prot. Further discussion may help understanding the advantages of these two datasets.
- The paper could provide more details on parameters used for defining clusters with Silix.

Minor remark:

The naming system may be perfected. Although imaginative and automatic, it seems to be limited. For example cluster 77 in the -prot-hmm-txt.zip (v 15.1) contains 398 sequences, which are obviously rep proteins for ssDNA viruses circo, gemini and their satellites, but the names fished out by author’s method are not clear. This can be problematic for sequence assignation.

Keywords for cluster 77:

protein 329
replication 296
virus 183
alphasatellite 154
associated 138
putative 120
viral 103
CRESS 78
leaf 50
curl 49
initiator 49
Rep 41
yellow 39
Circoviridae 39

Actually, the name used to create RVDB-pro keywords is not clearly defined. The name of GenBank protein entry are not very consistent and this may explain these problems. Maybe using pfam or any other method of identification over the clusters may help naming them in a more consistent way.

References

1. Xu GJ, Kula T, Xu Q, Li MZ, et al.: Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome.*Science*. 2015; **348** (6239): aaa0698 PubMed Abstract | Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Partly

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Virus proteomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 26 Aug 2020

Thomas Bigot, Institut Pasteur, Paris, France

We would like to thank the Reviewer. Please find below our line-by-line responses.

The database is described as facilitating sequence assignation. This seems a bit vague, a sentence describing possible applications could help.

We have added the following sentence to exemplify possible applications: “When trying to characterize sequences present in a metagenomics sample, searching first for related sequences in a viral database can lead to identify rapidly a known virus (high identity between the query sequence and the one in the database), or identify potential new species (low identity with any known sequence). Such hits must be further characterized on more comprehensive databases to increase the robustness of taxonomic assignations.”

The introduction may describe better the current state of research in the field. UniProtKB should be cited in “existing databases” for viral proteins, and authors may add citations of its use in virus detection. (ex: UniRef90 used with success to create synthetic human virome1 (PMID: 26045439)). This would also highlight new potential applications for RVDB-prot. UniProtKB provides data for 3,972,271 viral proteins, a bit more than RVDB-Prot (3,899,699). RVDB-prot data is based similarity gathering of sequences with viral RefSeq, which has the advantage to ignore any taxonomical issues. On the other hand, many RefSeq are provisional, and those are not free or errors. The authors may discuss the advantage of their method over existing protein dataset.

We have updated the introduction, introducing UniProtKB viral sequences: “UniProtKB11 contains numerous viral sequences (: 4 497 049 in total, including 17 008 (0.38%) reviewed ones) that could, as for NCBI/nr, increase computation time when thousands of sequences have to be analyzed concomitantly, which is routinely practiced in metagenomics analyses.” We have
also updated the description of RefSeq (with updated contents) and better exemplified the benefit of RVDB over these two databases.

Similarly, UniRef90 contains 577,105 clusters of proteins, which could be compared to the 489,207 unique proteins of RVDB-prot. Further discussion may help understanding the advantages of these two datasets.

We stressed on the first asset of RVDB: the curation that is done on the sequences is unique and allows to raise confidence in the fact that all the sequences of the database are real viral sequences.

The paper could provide more details on parameters used for defining clusters with Silix.

Done. We used the default parameters of Silix.

The naming system may be perfected. Although imaginative and automatic, it seems

to be limited. For example cluster 77 in the -prot-hmm-txt.zip (v 15.1) contains 398 sequences, which are obviously rep proteins for ssDNA viruses circo, gemini and their satellites, but the names fished out by author's method are not clear. This can be problematic for sequence assignation. Actually, the name used to create RVDB-pro keywords is not clearly defined. The name of GenBank protein entry are not very consistent and this may explain these problems. Maybe using pfam or any other method of identification over the clusters may help naming them in a more consistent way.

We are grateful for this remark which helped us make the naming system clear. Indeed, the pipeline does now query PFAM to annotate sequences. As explained in the new version of the manuscript, for each cluster, we query PFAM with every sequences of this cluster, using --cut_ga option of HMMER (this option makes HMMER trust PFAM GA bitscore cutoff defined for each cluster). We kept the original system (using sequences descriptions) despite the fact they are inaccurate, since sometimes, we do not find homologs clusters in PFAM.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research