DATA NOTE

RVDB-prot, a reference viral protein database and its HMM profiles [version 1; peer review: 2 approved with reservations]

Thomas Bigot¹, Sarah Temmam¹ , Philippe Pérot², Marc Eloit^{2,3}

¹Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS, Institut Pasteur, Paris, 75015, France

²Biology of Infection Unit, Pathogen Discovery Laboratory, Institut Pasteur, Paris, 75015, France

³École Nationale Vétérinaire d'Alfort, Maisons-Alfort, 94700, France

V1 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

Approval Status

	1	2
version 2		
(revision)		
07 Sep 2020		
version 1		
23 Apr 2019		

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

2. **Guy Perriere** , Université 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: © 2019 Bigot T *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), 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 1; peer review: 2 approved with reservations]** F1000Research 2019, 8:530
<https://doi.org/10.12688/f1000research.18776.1>

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

Introduction

Sequence assignation often uses similarity criteria to infer homology, and hence taxonomy and / or protein type. In order to search for this similarity, reliable, accurate and comprehensive databases are required. In the specific field of viruses, several solutions are available yet their ability to provide valid results is highly dependant 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. RefSeq on the other hand, is generally better curated, but it contains only full-length genomes and rarely includes the latest discoveries. Other specialized databases provide only specific groups of taxa for specific purposes, for instance, virus families responsible for infectious diseases like HIV or influenza.

Thus, the need for better, well-annotated and comprehensive public viral databases that can be used for the identification of viruses by high-throughput sequencing lead Goodacre *et al.* to propose their Reference Viral DataBase (RVDB)¹. This database consists of a collection of all currently known viral genomes and virus-related nucleic sequences retrieved from NCBI nr or RefSeq, which includes a specific, both manual and computational reviewing process, as well as four updates of the contents per year. These features make RVDB quite attractive for the virology research community and in fact, in February 2018, version 15.1 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**² allow the building of a HMM profile from a multiple protein sequence alignment. This profile can then be able to recognize proteins based on complex positionspecific models of sequence conservation and evolution, and it does so in a more accurate way than if a classic sequence alignment is used.

Thus, 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, v15.1³ consists of a collection of 2 719 839 nucleic sequences¹. The accession numbers were extracted in order to gather the corresponding database entries in genbank format. From these entries, coding domain sequences

and the description of these sequences were located and copied into the protein collection. 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]>,&br/>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 3 899 699 protein sequence file.

Generation of HMM profiles

The HMM generation rationale was inspired from VFam (the database of profile HMM built from all the viral proteins present in RefSeq, discontinued from 2014)⁴, but was entirely re-coded as a Snakemake pipeline⁵, using different tools for some key steps (clustering, alignment). The proteins sequences were clustered with a 100% identity criteria to duplicates, using **CDHit** 4.7.0⁶. Then, the sequences were processed using **Blast** 2.2.26⁷ performing an all-against-all comparison. These comparisons allow **Silix** 1.2.6⁸ to define clusters of sequences according to the sequence similarity. This step produces a file text in which each sequence is associated to one cluster. The information of each cluster containing at least four sequences was transformed into a fasta file containing all of its sequences. Then, we performed multiple alignment using **Mafft** 7.023⁹ in *auto* mode. The multiple sequence alignments were processed by **HMMER** 3.2.1² in order to obtain the HMM profiles. The HMM profiles were then put together in a single file.

Annotation of HMM profiles

In our pipeline, a cluster consists in a set of sequences, where each sequence belongs to a species, and each sequence is associated with a description. In order to characterize the clusters, these pieces of information and other indicators (such as cluster length and sequence number) 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. These keywords correspond to the union of all the set of sequence names belonging to the cluster, weighted according to their frequencies, and excluding trivial words. For instance, for the cluster number 1, containing 588 sequences, the keywords and their frequencies, are: parvovirus(441), protein(423), Canine(359), capsid(345), VP2(233), virus(89), VP1(83), disease(48), Aleutian(48), mink(48) allowing to describe a cluster composed of Canine parvovirus capsid

Figure 1. Schema of the annotation database.

protein sequences. The database stores all these taxa, using NCBI 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. 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 pipeline⁵, available at Institut Pasteur’s Gitlab.

- Pipeline available from <https://gitlab.pasteur.fr/tbigot/rvdb-prot/>.
- Archived source code at time of publication: <http://doi.org/10.5281/zenodo.2630593>¹⁰
- 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 15.1 described in this manuscript is also available from Figshare.

Figshare: U-RVDBv15.1 <https://dx.doi.org/10.6084/m9.figshare.7745969>³.

This project contains the following underlying data:

- U-RVDBv15.1-prot.fasta (fasta file containing protein features of the original database: -prot.fasta)
- U-RVDBv15.1-prot.fasta-prot.hmm (the HMM profiles, generated with and for hmmer 3.2.1 (from 2019, 3.1b2 before))
- U-RVDBv15.1-prot.fasta-prot-hmm.sqlite (SQLite db containing annotations (please find a documentation below))
- U-RVDBv15.1-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.

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: 14.0 (2018-09), 13.0 (2018-06), 12.2(2018-03), 11.5 (2017-10), 10.2 (2017-04).

Usage HMMER can be used to search for all profiles in a fasta sequence file (sequences.fasta): `hmmsearch U-RVDBv15.`

Table 1. Metrics for release 15.1.

Nucleic sequences	RVDB	2 719 839
Proteins	RVDB-prot	3 899 699
Unique proteins	RVDB-prot	489 207
Clusters	RVDB-prot HMM	86 482

1-prot.fasta-prot.hmm sequences.fasta >
result.out. Additional options are available in HMMER
User's Guide.

Grant information

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

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

- 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](#)
- Eddy SR: **Accelerated Profile HMM Searches.** *PLoS Comput Biol*. 2011; **7**(10): e1002195.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bigot T, Temmam S, Pérot P, *et al.*: **U-RVDBv15.1.** *figshare*. Fileset. 2019.
<http://www.doi.org/10.6084/m9.figshare.7745969.v1>
- 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](#)
- Köster J, Rahmann S: **Snakemake—a scalable bioinformatics workflow engine.** *Bioinformatics*. 2012; **28**(19): 2520–2522.
[PubMed Abstract](#) | [Publisher Full Text](#)
- 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](#)
- 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](#)
- 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](#)
- 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](#)
- Bigot T: **RVDB-prot v15.1.0 (Version 15.1.0).** *Zenodo*. 2019.
<http://www.doi.org/10.5281/zenodo.2630593>

Open Peer Review

Current Peer Review Status: ? ?

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, Université 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 imprecise 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 imprecise 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 of 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 of 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