









Article

VOGDB—Database of Virus Orthologous Groups

Lovro Trgovec-Greif ^{1,2} , Hans-Jörg Hellinger ^{2,3} , Jean Mainguy ⁴ , Alexander Pfundner ^{1,2} ,
Dmitrij Frishman ⁵, Michael Kiening ⁵, Nicole Suzanne Webster ^{6,7,8} , Patrick William Laffy ⁶ ,
Michael Feichtinger ¹  and Thomas Rattei ^{1,*} 

- ¹ Centre for Microbiology and Environmental Systems Science, University of Vienna, 1030 Vienna, Austria
 - ² Doctoral School of Microbiology and Environmental Systems Science, University of Vienna, 1030 Vienna, Austria
 - ³ Armaments and Defence Technology Agency, Austria
 - ⁴ Genoscope, 91000 Evry Cedex, France
 - ⁵ Department of Bioinformatics, School of Life Sciences, Technical University Munich, 85350 Freising, Germany
 - ⁶ Australian Institute of Marine Science, PMB no3 Townsville MC, Townsville 4810, Australia
 - ⁷ Institute for Marine and Antarctic Studies, University of Tasmania, Hobart 7000, Australia
 - ⁸ Australian Centre for Ecogenomics, University of Queensland, Brisbane 4072, Australia
- * Correspondence: thomas.rattei@univie.ac.at

Abstract: Computational models of homologous protein groups are essential in sequence bioinformatics. Due to the diversity and rapid evolution of viruses, the grouping of protein sequences from virus genomes is particularly challenging. The low sequence similarities of homologous genes in viruses require specific approaches for sequence- and structure-based clustering. Furthermore, the annotation of virus genomes in public databases is not as consistent and up to date as for many cellular genomes. To tackle these problems, we have developed VOGDB, which is a database of virus orthologous groups. VOGDB is a multi-layer database that progressively groups viral genes into groups connected by increasingly remote similarity. The first layer is based on pair-wise sequence similarities, the second layer is based on the sequence profile alignments, and the third layer uses predicted protein structures to find the most remote similarity. VOGDB groups allow for more sensitive homology searches of novel genes and increase the chance of predicting annotations or inferring phylogeny. VOGDB uses all virus genomes from RefSeq and partially reannotates them. VOGDB is updated with every RefSeq release. The unique feature of VOGDB is the inclusion of both prokaryotic and eukaryotic viruses in the same clustering process, which makes it possible to explore old evolutionary relationships of the two groups. VOGDB is freely available at vogdb.org under the CC BY 4.0 license.

Keywords: virus genomes; protein families; comparative genomics; orthologous groups; genome annotation; genome analysis



Citation: Trgovec-Greif, L.; Hellinger, H.-J.; Mainguy, J.; Pfundner, A.; Frishman, D.; Kiening, M.; Webster, N.S.; Laffy, P.W.; Feichtinger, M.; Rattei, T. VOGDB—Database of Virus Orthologous Groups. *Viruses* **2024**, *16*, 1191. <https://doi.org/10.3390/v16081191>

Academic Editor: Alexander Gorbalenya

Received: 1 July 2024
Revised: 21 July 2024
Accepted: 23 July 2024
Published: 25 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Viruses are a diverse group of biological entities that share the property of being obligate cellular parasites. Unlike in cellular organisms, no common genes or gene families are shared between all viruses [1]. This raises fundamental questions about virus ancestry and evolution. Moreover, the number of viruses on earth is huge (more than 10^{31} particles) [2,3], and it is estimated they carry between 10^8 and 10^{10} unique genes [4]. Most of the viral diversity is currently unexplored, and for the most sequenced viral genes, little is known about their function [5].

Viral genes not only encode a high number of different functions, which leads to a huge diversity of viral genomes, but also form heterogeneous groups of genes having similar function [6]. Due to the nature of viral lifestyle and their quick replication, mutations and selection, viruses explore the sequence space of genes in less evolutionary time than cellular organisms do [7,8]. Because of the heterogeneity of viral proteins, it is often

difficult to find homologs in databases by traditional bioinformatics, such as pair-wise sequence alignments.

The computational inference of gene homology is valuable for annotating genes that are known from their sequence, but have not been experimentally characterized. Homologous genes have diverged from a common ancestral gene and are likely to have same or similar functions in different organisms. A particularly informative computational observation is gene orthology. Orthologous genes have diverged from a common ancestor by a process of speciation (as opposed to the gene duplication in paralogy). Orthologous genes are more likely to keep the ancestral function [9]. Orthologous genes from multiple organisms form orthologous groups. Homologous relationships are deduced from sequence comparisons due to the assumption that important sequence motifs will stay conserved during evolution [10]. However, due to the absence of universal phylogenetic markers for all viruses and frequent horizontal gene transfers between viruses and viruses as well as viruses and hosts, no universal concepts for the orthology of viral protein families are so far available in bioinformatics.

Due to quick viral evolution, it is often impossible to detect homology by the pair-wise alignment of two protein sequences, especially for proteins that diverged longer ago. However, by building a sequence model based on the group of easily detectable homologs, a conserved pattern becomes discernible, which can be used to connect more distant groups [11]. This approach is widely used by databases that cluster together viral proteins, including pVOG [12], which focuses on prokaryotic viruses, as well as the viral sequences of eggNOG [13]. The PHROGs database [14] clusters phage genomes in two steps, first by grouping them based on the direct sequence comparison and later by clustering group Hidden Markov Models (HMMs) to capture remote homology. However, none of these databases represents the high number and broad diversity of virus genome sequences available to date.

We therefore introduce VOGDB, which is a comprehensive database of virus orthologous groups, virus protein families and virus protein structural folds. VOGDB provides these three layers of homologous groups for all viral proteins from RefSeq genomes [15]. The layers are intended to gather proteins with the increasing evolutionary distance reflected in the higher sequence divergence. Contrary to the prokaryotic genomes from RefSeq, where PGAP [16] is used for the submission and consistent reannotation of genomes, virus genomes in RefSeq may keep their annotation from their GenBank [17] submission. VOGDB, making use of all virus genomes from RefSeq, addresses the potential problem of inconsistent and outdated annotation by filtering and partial reannotation in order to ensure a higher quality of final clusters.

2. Materials and Methods

2.1. General Concept

The first layer of the VOGDB is constructed by all-against-all pair-wise sequence comparisons and represents the easily detectable homologs. The second layer is created by clustering sequence models (HMMs) from the first layer to capture the homology of proteins that diverged beyond the point where homology can be detected by pair-wise alignments. In the third layer, we group together families from the second layer by their shared features within predicted 3D structures. This layer represents remotely homologous groups whose members diverged to a degree that sequence comparison methods cannot detect their similarity anymore. As there is no standard way to validate viral orthologous groups, we suggest an approach based on the homogeneity of functional and structural annotations in terms of SwissProt [18] keywords and SCOPe [19] superfamily labels. The calculation of homogeneity was also applied to other similar databases (pVOG, PHROGs and COG) to compare if VOGDB shows similar homogeneity despite its higher number and wider diversity of genome sequences. pVOG and PHROGs are databases with viral proteins and are directly comparable to the first and second layers from VOGDB. The COG database contains prokaryotic proteins grouped by orthology and was included as a control.

2.2. Preprocessing of Input Data

2.2.1. Input from RefSeq

The input data are all of the complete viral genomes from RefSeq [15], which have at least one protein annotated. Around 98% of records from RefSeq enter the VOGDB pipeline, meaning VOGDB represents almost the entire viral portion of RefSeq. All sequence records with the same taxonomy ID, strain and isolate are considered one genome in VOGDB, which are further called VOGDB genomes.

2.2.2. Polyproteins

Polyproteins are present in DNA viruses and almost all RNA and retroviruses. A polyprotein is translated as a large polypeptide from a single ORF and is later cleaved into functional proteins [20]. At the moment, no general computational strategy exists that would predict the cleavage sites in polyproteins and find the borders of the individual peptides. The iterative approach LAMPA annotates multidomain proteins and addresses the problem that statistical significance is related to the length of domains [21]. We have developed a strategy to annotate the individual peptides from the polyprotein sequence without prior knowledge of conserved domains. First, individual peptides originating from a polyprotein or from RefSeq records that have been validated by the VOGDB team are collected in a peptide reference database. Second, non-annotated or incompletely annotated polyproteins are then reannotated by the best non-overlapping pair-wise sequence alignments against the peptide reference database. Within VOGDB, annotated or reannotated peptides replace the respective segments of their initial polyprotein records and together with the rest of the proteins are called VOGDB proteins.

2.3. Creation of the First-Layer Clusters—VOGs

VOGDB proteins are used as the input to the COGSoft pipeline with the aim of constructing clusters of recently diverged proteins [22]. In short, an all-against-all PSI-BLAST [23] search is conducted followed by the COGtriangles [22] procedure to find orthologous groups. We use the strict clustering, which does not allow for a single protein to be a member of multiple clusters. For each orthologous group, a multiple sequence alignment of all member proteins is calculated using Clustal Omega [24]. Scores according to the minimum reporting standard for multiple sequence alignments are obtained using the program alistat [25]. From the multiple alignment, we calculate Hidden Markov Models (HMMs) using hmmbuild from HMMER [26]. The resulting groups are called VOGs to reflect that they are a viral equivalent to orthologous groups.

2.3.1. Functional Annotation

Annotations of VOGDB clusters are made with the aim of describing most of the cluster members as specifically as possible, and therefore, we are using a consensus of the annotations of the individual proteins as the cluster annotation. During the annotation procedure, we prefer manually curated functional information over computationally inferred annotation. VOGs are functionally annotated, if possible, by deriving functional annotations from hits to the most recent SwissProt [18] database or from the annotations as provided by RefSeq. To retrieve the annotation from SwissProt, we used BLAST [27] to search the SwissProt database with the members of a VOG. For an individual protein from a VOG, we retained the functional annotation of a maximum of 5 hits if the e-value was less than 10^{-10} and the alignment coverage was more than 90%. All annotations of all proteins in a VOG are collected, and the most common annotation string found for a VOG is used as the annotation for that VOG. In cases when it is not possible to obtain the annotation from SwissProt, we collect annotations of proteins in a VOG as they are in RefSeq and use the most common annotation string as the annotation for the VOG.

As an additional step in the annotation process, we maintain a list of SwissProt keywords with which we associate a functional category. Every functional annotation of VOGs belongs to one or more functional categories: virus replication (Xr), virus structure

(Xs), viral protein beneficial for the host (Xh), viral protein beneficial for the virus (Xp) and unknown function (Xu).

2.3.2. Naming

VOG are named with a prefix “VOG” and a number padded with zeroes. To facilitate the comparison of the results between releases, we implemented a stable numbering scheme. VOGs from the older release are compared to the VOGs from the newer release, and the newer VOG receives the name of the largest older VOG for which 50% or more of the proteins are found in the new VOG. For VOGs that do not receive the name from the previous release, a new number is created.

2.4. Creation of the Second Layer Clusters—VFAMs

Clustering Using MCL

To create the second-layer clusters (VFAMs), we first need to align HMMs of VOGs. The alignment is achieved using the `halign` function from HH-Suite [28]. The HMM–HMM alignments are filtered by three different criteria: maximal evalue of 1×10^{-5} , minimal HMM probability value of 85 and minimal coverage for both HMMs of 0.7. The scores of alignments that pass all three criteria are used as input to the MCL clustering algorithm [29] where VOGs are clustered with the inflation value of 2. Clustered sequences are aligned with Clustal Omega [24], assessed with `alistat` [25], and an HMM of the alignment is calculated by the function `hmmbuild` from HMMER [26]. The functional annotation of VFAMs are obtained in the same way as for VOGs. Naming works the same as for VOGs but with a different prefix: “VFAM”.

2.5. Creation of the Third-Layer Clusters—VFOLDS

The third layer of the VOGDB consists of VFOLDS, which are clusters of VFAMs grouped based on the shared structural features. A few experimentally resolved structures of viral proteins are available in the public databases like `pdb` [30]. Therefore, we used `alphafold 2` [31] to predict structures of viral proteins in VFAMs. Since there are more than 500,000 proteins in VFAMs, predicting this number of structures would not be feasible. The strategy was to select one representative for every VFAM and cluster the representatives instead of the whole VFAMs. To select a representative, we have aligned all members of VFAM to the HMM of that VFAM and selected the highest scoring member as a candidate for which the structure would be predicted by `alphafold 2`. After obtaining structure predictions for all representatives, we conducted the clustering using the `FoldSeek` tool [32] with the default settings (`commit 427df8a6b5d0ef78bee0f98cd3e6faaca18f172d`, `command: foldseek easy-cluster`). `FoldSeek` was used to cluster predicted structures from `AlphaFoldDB` [33] and was therefore an appropriate choice for our clustering task. Functional annotations of VFOLDS are obtained in the same way as for VOGs and VFAMs. Starting from VOGDB release 225, we published all predicted protein structures and their pLDDT scores with each VOGDB release.

2.6. Quality Assessment of the Clustering Results

The quality of the clustering was assessed by the homogeneity of functional annotations of cluster members and the structural superfamily membership of cluster members. The homogeneity of clusters from VOGDB was compared to the homogeneity of a random model. We obtained the random model by randomly scrambling functional annotation keywords or structure superfamily labels between annotated proteins and calculated the homogeneity. The randomization step was repeated 1000 times. For functional annotations, we searched `SwissProt` with all protein members of a group, used the keyword of the top-level functional annotation of the hits and calculated the relative frequency of the most common annotation compared to all retrieved annotations. To assess the homogeneity of structural patterns, we used a similar approach, but instead of searching `SwissProt`, we searched the `astral95` database (v2.08) [19] using `cd-hit` [34]. Hits were associated with

protein structural superfamilies as described in the SCOPe database [19]. The homogeneity for superfamilies was calculated as the relative frequency of the most common superfamily per group. Comparison of the homogeneity to the random model was made using the Kolmogorov–Smirnov test.

3. Results

3.1. Database

As the RefSeq database is updated bimonthly, VOGDB is updated with every RefSeq release, and the new release is made available shortly after the newest version of RefSeq is released. The release number of VOGDB is the same as the release number of RefSeq, which was used to build it. As an example in the text, the VOGDB version 221 based on the RefSeq 221 will be used, and it contains 14,974 VOGDB genomes. The polyprotein reannotation step predicted 5499 additional peptides from 995 polyproteins.

3.2. Content

In the VOGDB release 221, 606,019 viral proteins were clustered and produced 59,196 VOGs, 38,576 VFAMs and 30,516 VFOLDS (Figure 1). Due to the clustering, 352,350 (58%) proteins have functional annotation compared to 333,379 (55%) of the initial proteins from RefSeq that were not annotated as hypothetical proteins. The size distribution of the groups from all three layers shows the expected pattern observed in the similar databases where there are many of the smaller groups and a few of the larger groups. The distribution of the VOGs, VFAMs and VFOLDS according to their size is visualized in Figure 2. A feature of VOGs, VFAMs and VFOLDS is the information on the lowest common ancestor (LCA) of the viruses contributing proteins to the groups. Particularly interesting groups are those with LCA “viruses”, which means that proteins from different viral realms were clustered together. There are 2441 such VOGs (4.1%), 1443 VFAMs (3.7%) and 1515 VFOLDS (4.8%). Three files containing the lists of these clusters are available online under <https://vogdb.org/evaluation/vogdb221>.

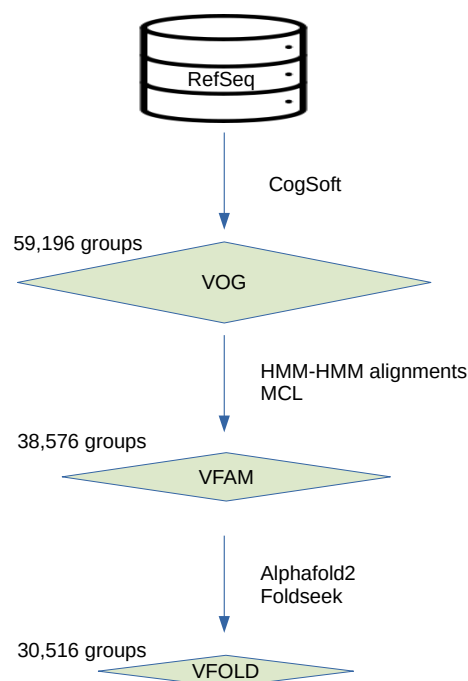


Figure 1. Schema of the layered structure of the database. For each layer, different tools were used to create clusters. Clusters from every next layer are built from the clusters of the previous layer and are connected by more remote similarity.

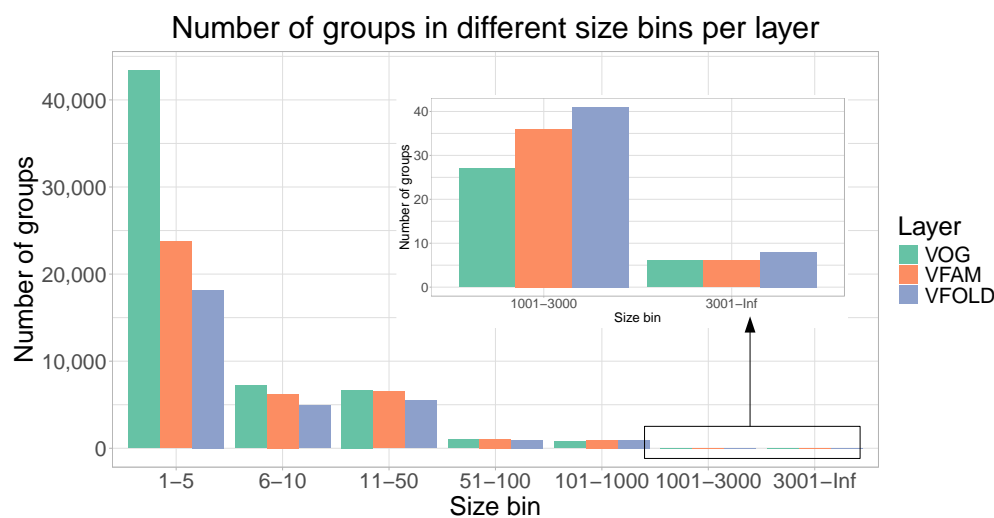


Figure 2. Number of groups per layer in different size bins. Size bins represent the range of the number of proteins for groups in a certain bin. The distribution with many smaller clusters and fewer of the larger ones is what is also observed in the similar databases.

3.3. Quality Assessment

As there is not yet a universal standard procedure to evaluate the clustering of the viral proteins into orthologous groups, we assessed the quality of the VOGDB clusters using the homogeneity of functional annotation and structural classification. If the clustering would perfectly group the proteins by structure and function, all proteins in one cluster would have the same and unique functional and structural annotation. The level of granularity needs to ensure maximal information for the entire database. Too coarse granularity would overestimate the homogeneity, and too fine would underestimate it. We selected the SwissProt keywords and the SCOPe superfamilies as the granularity level at which we calculate the homogeneity. Because there is a limited number of keywords describing the function, we estimated the baseline of the homogeneity from the random model described earlier. Quality assessment based on the homogeneity (Figure 3) shows that both the functional and structural homogeneity of groups from different layers of VOGDB are significantly larger than the baseline for all of the size bins (Kolmogorov–Smirnov test, p -value $< 10^{-5}$).

3.4. Comparison with Similar Databases

To evaluate the homogeneity of functional annotations and structural features, we calculated the homogeneity of the COG database (the release from 2020) [35], the PHROG database (v3) [14] and the pVOG database (May 2016) [12] in the same way as for the VOGDB layers. Clusters in the pVOG database are created similarly as VOGs, and the PHROGs clusters are similar to VFAMs. However, VOGDB has a bigger scope than pVOG and PHROG by including both phages and eukaryotic viruses and therefore needs to cluster more and more diverse proteins. The COG database was included as a control, as it was created using a similar clustering methodology and is manually curated. Figure 4 shows that the homogeneity of VOGDB layers is in the same range as the homogeneity of databases grouping prokaryotic orthologs (COG), phage orthologs (pVOG) and phage remote homologs (PHROG). The homogeneity of clusters from pVOG and VOGs is very similar, which is expected as both are created using COGSoft [22].

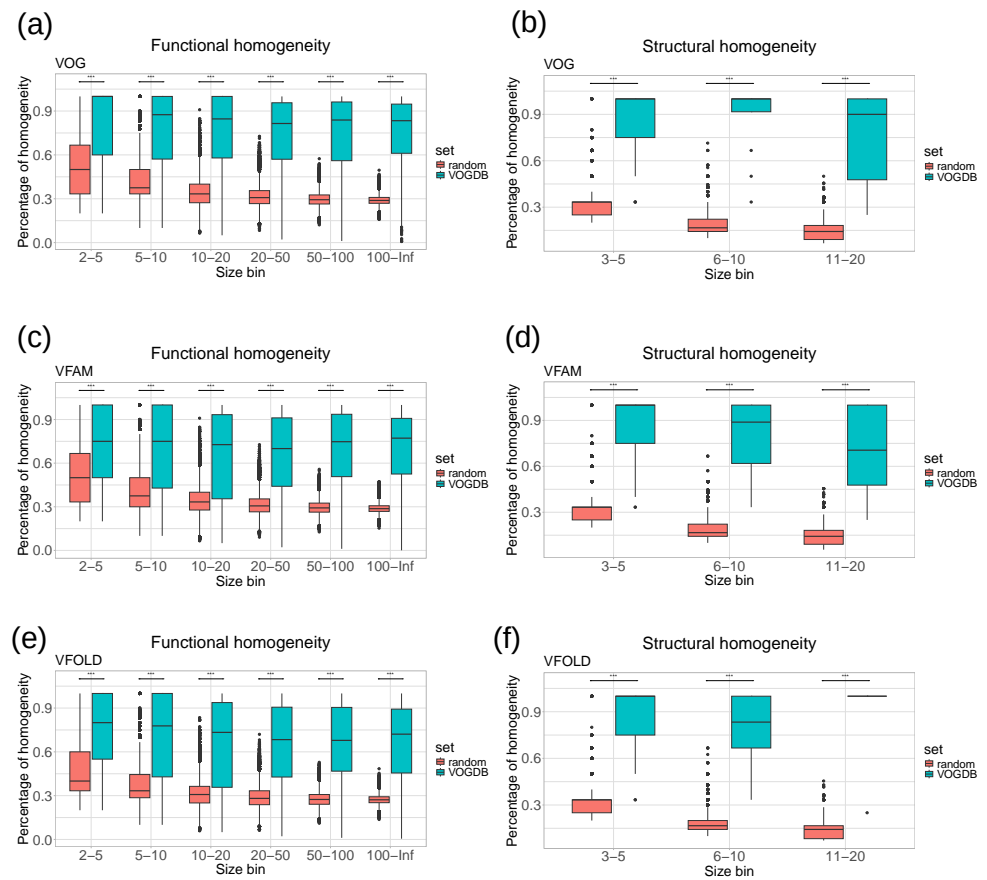


Figure 3. Homogeneity of functional annotations and protein structure classifications in VOGDB layers compared to the random model. (a–f) The groups from each layer are put into size bins based on the number of proteins with functional and structural annotation. The random model is created by randomly redistributing the functional and structural annotation labels between the proteins with respective annotation 1000 times and calculating the overall homogeneity. The results show that groups from VOGDB layers are significantly more homogeneous in terms of SwissProt keywords and structural classifications based on the SCOPe superfamilies (Kolmogorov–Smirnov test, $p < 10^{-5}$).

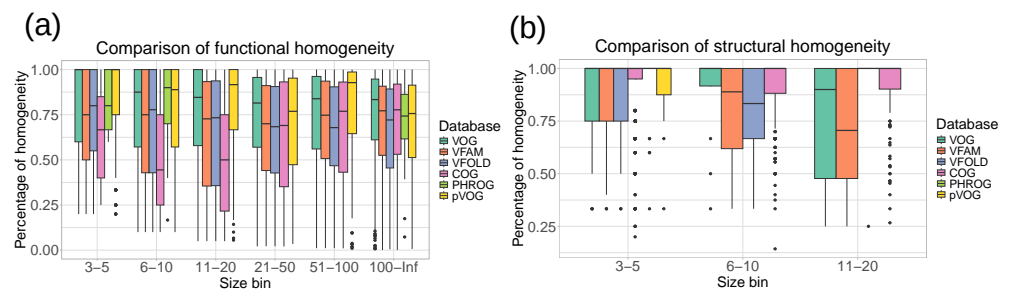


Figure 4. Homogeneity of SwissProt keywords (a) and SCOPe superfamilies (b) for layers from VOGDB and the other databases with orthologous/homologous groups: pVOG (phage orthologous groups), PHROG (phage remote orthologous groups) and COG (prokaryotic orthologous groups). The databases are split into size bins according to the number of proteins with a functional or structural annotation. Bins containing less than 3 proteins are not shown. The results show that the function and structure-based homogeneity of the layers from VOGDB are in the same range as in other similar databases.

3.5. Availability

3.5.1. VOGDB Webpage

VOGDB is accessible online at <https://vogdb.org> where it is possible to browse the clusters and see the statistics of the latest release. The webpage is updated regularly as a new version of VOGDB is calculated. The pre-computed files for the comparison of the VOGDB clusters with clusters from similar databases (see above) are available online under <https://vogdb.org/evaluation/vogdb221>.

3.5.2. VOGDB Release Files

Apart from being accessible via the webpage, we offer all of the resulting files for download. The files offered are formatted similarly to the files offered by the EggNOG database [36]. The most important files offered are HMMs of the clusters and multiple sequence alignments, files with the lowest common ancestry, files with a functional annotation of clusters, an interactive chart of genome taxonomies and predicted structures of VFAM representatives.

4. Discussion

4.1. Limitations

VOGDB is so far the most complete database for virus orthologous groups, virus protein families and virus protein structural similarities. However, it is based on the annotations provided by the underlying RefSeq database. So far, several annotation quality filters and the reannotation of polyproteins are the only means that VOGDB uses to ensure the high accuracy of its input data. A consistent reannotation of all virus genomes is not in the scope of VOGDB. Nevertheless, such reannotation will become increasingly important to sustain the value of comparative genomics of viruses. The VOGDB groups can be a valuable tool toward this aim, e.g., by predicting protein-coding genes that were missed in the original genome annotations.

4.2. Support for Bioinformatic Workflows

Viral hallmark genes [37] could be defined as genes that are found in diverse viruses but have no or only few homologs in cellular organisms and are therefore indicative of the viral origin of a sequence. HMMs of VOGs and VFAMs that represent viral hallmark genes can be used to predict viral sequences from unknown genomes [38] and to estimate the contamination of a viral sequence with bacterial genes [39]. HMMs of groups of viral proteins (either hallmark or not) could be used as input for various other tools. For example, the tool HMM-GraspX [40] uses protein family HMMs to guide the assembly, which is useful if the aim of the analysis is to analyze viruses in samples with a low abundance of viral reads or when the focus on specific families is needed [41]. For VOGDB clusters, we calculate the virus specificity based on the number of hits to cellular organisms based on the HMM–HMM search to the most recent eggNOG database [36]. This database contains selected representatives for cellular species and shows less study bias than genome sequence archives. We therefore approximate virus specificity by allowing for hits in maximally two, three or four cellular genomes with decreasing e-Value thresholds. The virus specificity information can be used to identify clusters representing the viral hallmark genes. Table 1 shows the number of virus-specific VOGs and VFAMs at different stringency criteria, accepting few cellular homologs as expected, e.g., from proviruses.

Table 1. Virus specificity of vFAMs. Virus-specific vFAMs are useful for identifying the viral hallmark genes, the genes definitive for the viral state and with only a very remote similarity to cellular genes. In VOGDB, viral specificity is defined with three stringency levels: strict, medium and low with hits to maximally two, three or four cellular genomes with e-values up to 10^{-4} , 10^{-10} and 10^{-15} .

Layer	Strict	Medium	Low
vOG	38,562	45,613	48,627
vFAM	28,500	32,546	33,951

4.3. Usage for Metagenome Analysis

VOGDB is useful for analyzing metagenomic datasets that intentionally or accidentally contain virus nucleic acid sequences. When pair-wise sequence database searches fail to reveal hits, homology searches with databases containing HMMs, such as from VOGDB, are more sensitive and allow for more proteins to be annotated. In addition to the functional annotation, lineage information of the genome carrying the gene can be inferred. By mapping all genes of a viral contig to VFAMs and using the information about the lowest common ancestor of VFAMs, one can estimate the virus origin of the whole contig. The performance of profile hidden Markov model databases, including VOGDB, for virus identification has recently been evaluated across multiple application scenarios, utilizing both simulated and real metagenomic data [42].

5. Conclusions

VOGDB is a novel resource in the field of virus bioinformatics, and it offers unique features compared to the similar databases and will complement the current toolbox for studying viral genomes. By including both phages and eukaryotic viruses from RefSeq, VOGDB has the biggest scope of all virus orthology databases, and it still ranks similarly with them in terms of the homogeneity of functional annotations and structural classes. The three layers of grouping give the opportunity to analyze the gene clusters connected by the increasingly remote similarity. Downloadable files, including functional annotations of clusters and HMMs, as well as bi-monthly updates that follow the RefSeq releases, make VOGDB a universal tool for downstream workflows in virus bioinformatics.

VOGDB is under constant development, and new knowledge about viruses is quickly implemented (for example, the new phage taxonomy [43]). On the other hand, the stable naming of clusters allows for the comparability of the results obtained by different releases of the database. VOGDB will also be further developed with respect to the user needs and to novel computational algorithms. With release 221, we have replaced the one clustering level with three clustering levels, including structural similarity. This has improved the usability of VOGDB for studying viral protein families without detectable sequence similarity. In the future, we will incorporate improved methods for structure prediction as they become available. We also will improve the reproducibility of VOGDB creation by making the entire database creation workflow open source. Finally, we plan to implement typical VOGDB-driven workflows, such as virus genome annotation or the classification of metagenomic contigs, as web-based open-source pipelines. We invite the user community to share their experience with us and inform us about their needs.

Author Contributions: Conceptualization, H.-J.H., D.F., N.S.W. and T.R.; Data curation, N.S.W., P.W.L. and T.R.; Funding acquisition, D.F. and T.R.; Investigation, H.-J.H., N.S.W., P.W.L. and T.R.; Methodology, D.F., M.K. and T.R.; Project administration, D.F., N.S.W., M.F. and T.R.; Resources, H.-J.H., M.F. and T.R.; Software, L.T.-G., H.-J.H., J.M., A.P., M.K. and T.R.; Supervision, T.R.; Validation, L.T.-G., J.M., M.K., N.S.W., P.W.L. and T.R.; Visualization, L.T.-G. and P.W.L.; Writing—original draft, L.T.-G. and T.R.; Writing—review and editing, L.T.-G., H.-J.H., A.P., D.F., M.K., N.S.W., P.W.L., M.F. and T.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FWF Austrian Science Fund grant number I1303. Lovro Trgovc-Greif was supported by the European Union's Horizon 2020 research and innovation pro-

gram, under the Marie Skłodowska-Curie Actions Innovative Training Networks grant agreement no. 955974 (VIROINF).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository.

Acknowledgments: Open Access Funding by the University of Vienna.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Villarreal, L. Evolution of Viruses. In *Encyclopedia of Virology*; Elsevier: Amsterdam, The Netherlands, 2008; pp. 174–184. [[CrossRef](#)]
2. Hendrix, R.W.; Smith, M.C.M.; Burns, R.N.; Ford, M.E.; Hatfull, G.F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 2192–2197. [[CrossRef](#)] [[PubMed](#)]
3. Mushegian, A.R. Are There 10^{31} Virus Particles on Earth, or More, or Fewer? *J. Bacteriol.* **2020**, *202*, e00052–20. [[CrossRef](#)] [[PubMed](#)]
4. Koonin, E.V.; Krupovic, M.; Dolja, V.V. The global virome: How much diversity and how many independent origins? *Environ. Microbiol.* **2023**, *25*, 40–44. [[CrossRef](#)] [[PubMed](#)]
5. Krishnamurthy, S.R.; Wang, D. Origins and challenges of viral dark matter. *Virus Res.* **2017**, *239*, 136–142. [[CrossRef](#)]
6. Kuchibhatla, D.B.; Sherman, W.A.; Chung, B.Y.W.; Cook, S.; Schneider, G.; Eisenhaber, B.; Karlin, D.G. Powerful Sequence Similarity Search Methods and In-Depth Manual Analyses Can Identify Remote Homologs in Many Apparently “Orphan” Viral Proteins. *J. Virol.* **2014**, *88*, 10–20. [[CrossRef](#)] [[PubMed](#)]
7. Stern, A.; Andino, R. Viral Evolution. In *Viral Pathogenesis*; Elsevier: Amsterdam, The Netherlands, 2016; pp. 233–240. [[CrossRef](#)]
8. Koonin, E.V.; Dolja, V.V.; Krupovic, M. The logic of virus evolution. *Cell Host Microbe* **2022**, *30*, 917–929. [[CrossRef](#)] [[PubMed](#)]
9. Koonin, E.V. Orthologs, Paralogs, and Evolutionary Genomics. *Annu. Rev. Genet.* **2005**, *39*, 309–338. [[CrossRef](#)] [[PubMed](#)]
10. Pearson, W.R. An Introduction to Sequence Similarity (“Homology”) Searching. *Curr. Protoc. Bioinform.* **2013**, *42*, 3.1.1–3.1.8. [[CrossRef](#)] [[PubMed](#)]
11. Yoon, B.J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genom.* **2009**, *10*, 402–415. [[CrossRef](#)]
12. Graziotin, A.L.; Koonin, E.V.; Kristensen, D.M. Prokaryotic Virus Orthologous Groups (pVOGs): A resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **2017**, *45*, D491–D498. [[CrossRef](#)]
13. Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M.C.; Rattei, T.; Mende, D.R.; Sunagawa, S.; Kuhn, M.; et al. eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **2016**, *44*, D286–D293. [[CrossRef](#)] [[PubMed](#)]
14. Terzian, P.; Olo Ndela, E.; Galiez, C.; Lossouarn, J.; Pérez Bucio, R.; Mom, R.; Toussaint, A.; Petit, M.A.; Enault, F. PHROG: Families of prokaryotic virus proteins clustered using remote homology. *NAR Genom. Bioinform.* **2021**, *3*, lqab067. [[CrossRef](#)]
15. Haft, D.H.; Badretdin, A.; Coulouris, G.; DiCuccio, M.; Durkin, A.; Jovenitti, E.; Li, W.; Mersha, M.; O'Neill, K.; Virothaisakun, J.; et al. RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes. *Nucleic Acids Res.* **2024**, *52*, D762–D769. [[CrossRef](#)]
16. Li, W.; O'Neill, K.R.; Haft, D.H.; DiCuccio, M.; Chetvernin, V.; Badretdin, A.; Coulouris, G.; Chitsaz, F.; Derbyshire, M.; Durkin, A.S.; et al. RefSeq: Expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* **2021**, *49*, D1020–D1028. [[CrossRef](#)] [[PubMed](#)]
17. Benson, D.A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Ostell, J.; Pruitt, K.D.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2018**, *46*, D41–D47. [[CrossRef](#)]
18. Boutet, E.; Lieberherr, D.; Tognolli, M.; Schneider, M.; Bansal, P.; Bridge, A.J.; Poux, S.; Bougueleret, L.; Xenarios, I. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods Mol. Biol.* **2016**, *1374*, 23–54. [[CrossRef](#)] [[PubMed](#)]
19. Chandonia, J.M.; Guan, L.; Lin, S.; Yu, C.; Fox, N.; Brenner, S. SCOPe: Improvements to the structural classification of proteins—Extended database to facilitate variant interpretation and machine learning. *Nucleic Acids Res.* **2022**, *50*, D553–D559. [[CrossRef](#)]
20. Yost, S.A.; Marcotrigiano, J. Viral precursor polyproteins: Keys of regulation from replication to maturation. *Curr. Opin. Virol.* **2013**, *3*, 137–142. [[CrossRef](#)]
21. Gulyaeva, A.A.; Sigorskih, A.I.; Ocheredko, E.S.; Samborskiy, D.V.; Gorbalenya, A.E. LAMPA, LARge Multidomain Protein Annotator, and its application to RNA virus polyproteins. *Bioinformatics* **2020**, *36*, 2731–2739. [[CrossRef](#)]
22. Kristensen, D.M.; Kannan, L.; Coleman, M.K.; Wolf, Y.I.; Sorokin, A.; Koonin, E.V.; Mushegian, A. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* **2010**, *26*, 1481–1487. [[CrossRef](#)]

23. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [[CrossRef](#)] [[PubMed](#)]
24. Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T.J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539. [[CrossRef](#)]
25. Wong, T.K.F.; Kalyaanamoorthy, S.; Meusemann, K.; Yeates, D.K.; Misof, B.; Jermini, L.S. A minimum reporting standard for multiple sequence alignments. *NAR Genom. Bioinform.* **2020**, *2*, lqaa024. [[CrossRef](#)] [[PubMed](#)]
26. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195. [[CrossRef](#)] [[PubMed](#)]
27. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)] [[PubMed](#)]
28. Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S.J.; Söding, J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* **2019**, *20*, 473. [[CrossRef](#)] [[PubMed](#)]
29. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* **2008**, *30*, 121–141. [[CrossRef](#)]
30. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P.A.; Crichlow, G.V.; Dalenberg, K.; Duarte, J.M.; et al. RCSB Protein Data Bank (RCSB.org): Delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.* **2023**, *51*, D488–D508. [[CrossRef](#)] [[PubMed](#)]
31. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [[CrossRef](#)]
32. Van Kempen, M.; Kim, S.S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C.L.M.; Söding, J.; Steinegger, M. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **2023**, *42*, 243–246. [[CrossRef](#)]
33. Barrio-Hernandez, I.; Yeo, J.; Jänes, J.; Mirdita, M.; Gilchrist, C.L.M.; Wein, T.; Varadi, M.; Velankar, S.; Beltrao, P.; Steinegger, M. Clustering predicted structures at the scale of the known protein universe. *Nature* **2023**, *622*, 637–645. [[CrossRef](#)] [[PubMed](#)]
34. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [[CrossRef](#)]
35. Galperin, M.Y.; Wolf, Y.I.; Makarova, K.S.; Vera Alvarez, R.; Landsman, D.; Koonin, E.V. COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **2021**, *49*, D274–D281. [[CrossRef](#)]
36. Hernández-Plaza, A.; Szklarzyk, D.; Botas, J.; Cantalapiedra, C.; Giner-Lamia, J.; Mende, D.R.; Kirsch, R.; Rattei, T.; Letunic, I.; Jensen, L.; et al. eggNOG 6.0: Enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* **2023**, *51*, D389–D394. [[CrossRef](#)] [[PubMed](#)]
37. Koonin, E.V.; Senkevich, T.G.; Dolja, V.V. The ancient Virus World and evolution of cells. *Biol. Direct* **2006**, *1*, 29. [[CrossRef](#)]
38. Guo, J.; Bolduc, B.; Zayed, A.A.; Varsani, A.; Dominguez-Huerta, G.; Delmont, T.O.; Pratama, A.A.; Gazitúa, M.C.; Vik, D.; Sullivan, M.B.; et al. VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **2021**, *9*, 37. [[CrossRef](#)] [[PubMed](#)]
39. Nayfach, S.; Camargo, A.P.; Schulz, F.; Eloë-Fadrosh, E.; Roux, S.; Kyrpides, N.C. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **2021**, *39*, 578–585. [[CrossRef](#)]
40. Zhong, C.; Edlund, A.; Yang, Y.; McLean, J.S.; Yooseph, S. Metagenome and Metatranscriptome Analyses Using Protein Family Profiles. *PLoS Comput. Biol.* **2016**, *12*, e1004991. [[CrossRef](#)] [[PubMed](#)]
41. Laffy, P.W.; Wood-Charlson, E.M.; Turaev, D.; Jutz, S.; Pascelli, C.; Botté, E.S.; Bell, S.C.; Peirce, T.E.; Weynberg, K.D.; Van Oppen, M.J.H.; et al. Reef invertebrate viromics: Diversity, host specificity and functional capacity. *Environ. Microbiol.* **2018**, *20*, 2125–2141. [[CrossRef](#)]
42. Yu, R.; Huang, Z.; Lam, T.Y.C.; Sun, Y. Utilizing profile hidden Markov model databases for discovering viruses from metagenomic data: A comprehensive review. *Briefings Bioinform.* **2024**, *25*, bbae292. [[CrossRef](#)]
43. Turner, D.; Shkoporov, A.N.; Lood, C.; Millard, A.D.; Dutilh, B.E.; Alfenas-Zerbini, P.; Van Zyl, L.J.; Aziz, R.K.; Oksanen, H.M.; Poranen, M.M.; et al. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.* **2023**, *168*, 74. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.