viruses

MDPI

Article

VIRIDIC—A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses

Cristina Moraru [1,*], Arvind Varsani [2,3] and Andrew M. Kropinski [4]

[1] Institute for Chemistry and Biology of the Marine Environment, Carl-von-Ossietzky-Str. 9–11, D-26111 Oldenburg, Germany
[2] The Biodesign Center for Fundamental and Applied Microbiomics, Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ 85287-5001, USA; Arvind.Varsani@asu.edu
[3] Structural Biology Research Unit, Department of Integrative Biomedical Sciences, University of Cape Town, Observatory, Cape Town 7701, South Africa
[4] Departments of Food Science, and Pathobiology, University of Guelph, Guelph, ON N1G 2W1, Canada; phage.canada@gmail.com
[*] Correspondence: liliana.cristina.moraru@uni-oldenburg.de

Received: 30 September 2020; Accepted: 4 November 2020; Published: 6 November 2020

Abstract: Nucleotide-based intergenomic similarities are useful to understand how viruses are related with each other and to classify them. Here we have developed VIRIDIC, which implements the traditional algorithm used by the International Committee on Taxonomy of Viruses (ICTV), Bacterial and Archaeal Viruses Subcommittee, to calculate virus intergenomic similarities. When compared with other software, VIRIDIC gave the best agreement with the traditional algorithm, which is based on the percent identity between two genomes determined by BLASTN. Furthermore, VIRIDIC proved best at estimating the relatedness between more distantly-related phages, relatedness that other tools can significantly overestimate. In addition to the intergenomic similarities, VIRIDIC also calculates three indicators of the alignment ability to capture the relatedness between viruses: the aligned fractions for each genome in a pair and the length ratio between the two genomes. The main output of VIRIDIC is a heatmap integrating the intergenomic similarity values with information regarding the genome lengths and the aligned genome fraction. Additionally, VIRIDIC can group viruses into clusters, based on user-defined intergenomic similarity thresholds. The sensitivity of VIRIDIC is given by the BLASTN. Thus, it is able to capture relationships between viruses having in common even short genomic regions, with as low as 65% similarity. Below this similarity level, protein-based analyses should be used, as they are the best suited to capture distant relationships. VIRIDIC is available at viridic.icbm.de, both as a web-service and a stand-alone tool. It allows fast analysis of large phage genome datasets, especially in the stand-alone version, which can be run on the user's own servers and can be integrated in bioinformatics pipelines. VIRIDIC was developed having viruses of *Bacteria* and *Archaea* in mind; however, it could potentially be used for eukaryotic viruses as well, as long as they are monopartite.

Keywords: nucleotide-based intergenomic similarity; nucleotide-based intergenomic distance; viruses; phages; VIRIDIC

1. Introduction

Intergenomic comparisons are useful in determining how viruses are related to each other. Indeed, the primary classification technique used by the International Committee on Taxonomy of Viruses (ICTV), Bacterial and Archaeal Viruses Subcommittee is based upon overall nucleic acid sequence identity. For a number of years, a crude method of estimating this was derived from BLASTN searches at NCBI, by multiplying the "query cover" by the "per. ident" values. The subcommittee established thresholds for the demarcation of viruses into species (95%) and into genera (~70%). While this technique is useful for undertaking pairwise comparisons, it is not convenient for comparisons of larger datasets.

There are a number of online tools and stand-alone software packages that have been used to compare viral genomes, including with the purpose of taxonomic classification. These include average nucleotide identity (ANI and ANI Calculator) [1–3], OrthoANI [4], EMBOSS Stretcher [5], Gegenees [6], JSpeciesWS [7], KI-S tool (https://f1000research.com/posters/7-147), pairwise sequence comparison (PASC) [8,9], Sequence Demarcation Tool (SDT) [10], Simka (https://arxiv.org/abs/1604.02412), and Yet Another Similarity Searcher (YASS) [11]. Some of the tools not only calculate, but also offer a visual of the comparison of the nucleic acid sequence relatedness (reviewed in [12]). To these we can add progressiveMauve [13] and VICTOR [14], which are focused on visualization of the alignments/genome relatedness, without explicitly giving access to the similarity values themselves.

The alignment algorithms used to calculate intergenomic relatedness vary from those based on the Needleman-Wunsch global alignment (Stretcher, SDT, PASC), to those based on BLASTN, either with previous genome fragmentation (Gegenees, OrthoANI) or without (PASC, VICTOR). With the exception of PASC and VICTOR, which can normalize the intergenomic identities to the whole genome length, most of the other tools normalize the intergenomic identities to the length of the alignment. This can lead to artificially high similarity values. The differences in algorithms can result in significant differences between the similarities reported by the different tools and can lead to inconsistencies in virus classification.

Here, with the purpose of offering a standardized and high-throughput tool for comparing viral genomes, we developed Virus Intergenomic Distance Calculator (VIRIDIC). VIRIDIC builds and improves on the traditional BLASTN method used by Bacterial and Archaeal Viruses Subcommittee from ICTV, to both calculate and visualize virus intergenomic relatedness. VIRIDIC is available as a web-service and as a stand-alone program for Linux, both accessible at viridic.icbm.de. It reports either intergenomic similarities or intergenomic distances.

2. Materials and Methods

2.1. VIRIDIC—Development and Workflow

VIRIDIC was developed in R 3.5 programming language [15]. The web interface was developed under the shiny web application framework (https://cran.r-project.org/web/packages/shiny/index.html, RStudio, MA, USA). The stand-alone tool for Linux was wrapped in a container using the Singularity v. 3.5.2 software (https://sylabs.io/, Sylabs.io, CA, USA). This VIRIDIC version can be deployed on any systems running the Singularity software, without any additional installation and configuration steps.

VIRIDIC's work flow consists of four steps. First, each viral genome is aligned against all other genomes in the dataset, using BLASTN 2.9.0+ from the BLAST+ package [16] with the core parameters "-evalue 1 -max_target_seqs 10,000 -num_threads 6". The default alignment parameters are "-word_size 7 -reward 2 -penalty -3 -gapopen 5 -gapextend 2". The user can choose between 3 other parameter sets: "-word_size 11 -reward 2 -penalty -3 -gapopen 5 –gapextend 2", "-word_size 20 -reward 1 -penalty -2", and "-word_size 28 -reward 1 -penalty -2".

Second, the BLASTN output is used to calculate pairwise intergenomic similarities. For one genome pair, the number of identical nucleotide matches reported by BLASTN is summed up for all aligned genomic regions. In the case of overlapping alignments, the overlapping part is removed from one of the aligned regions, such that, at the end, the different genome regions are represented only once in the alignments. The intergenomic similarity or distance is calculated as described below, as previously proposed [17].

$$\mathrm{simAB} = ((\mathrm{idAB} + \mathrm{idBA}) * 100) / (\mathrm{lA} + \mathrm{lB}), \tag{1}$$

$$\text{distAB} = 100 - \text{simAB}, \tag{2}$$

where

$$\begin{aligned} \text{idAB} &= \text{ identical bases when genome A is aligned to genome B,} \\ \text{idBA} &= \text{ identical bases when genome B is aligned to genome A,} \\ \text{lA} &= \text{length genome A,} \\ \text{lB} &= \text{length genome B,} \\ \text{simAB} &= \text{ intergenomic similarity between genomes A and B,} \\ \text{distAB} &= \text{intergenomic distance between genomes A and B,} \end{aligned}$$

The intergenomic similarity algorithm has been implemented to run on multiple central processing unit cores using the future v. 1.17.0 R package (https://github.com/HenrikBengtsson/future).

In the second step, VIRIDIC also calculates for each genome pair three other indicators related to the alignment: the aligned fraction for genome 1, the length ratio between genome 1 and genome 2, and the aligned fraction for genome 2.

$$\begin{aligned} \text{aligned fraction genome 1} &= \text{number of aligned bases for genome 1/length of genome 1,} \\ \text{genome length ratio} &= \text{smaller genome length/bigger genome length,} \\ \text{aligned fraction genome 2} &= \text{number of aligned bases for genome 2/length of genome 2,} \end{aligned}$$

Third, VIRIDIC performs a hierarchical clustering of the intergenomic similarity values. For this, the intergenomic similarities are clustered using the fastcluster v. 1.1.25 R package (https://cran.r-project.org/web/packages/fastcluster/index.html) [18]. For clustering, VIRIDIC uses by default the "complete" agglomeration method (see hclust function, fastcluster package). Several other agglomeration methods from the fastcluster package can be given as parameters.

Fourth, VIRIDIC graphically represents the intergenomic similarity values, the aligned ratios 1 and 2, and the genome length ratios as a heatmap, using the ComplexHeatmap v. 2.5.3 R package [19]. The heatmap is ordered based on the genome clustering by their similarity values.

VIRIDIC outputs an ordered similarity/distance matrix (tab-separated text format), a heatmap (pdf format), and a table with the viral genomes tentatively clustered at the species or genus level (tab-separated text format). Additionally, the stand-alone tool offers access to several intermediary files, both in RDS (file storing for an R object) and tab-separated text format, containing further information about the alignments. These files could eventually be integrated in bioinformatics pipelines.

2.2. Benchmarking VIRIDIC

The dataset used for benchmarking consisted of 60 T7-like phages genomes, from the Autographviridae family, downloaded from the GenBank RefSeq database [20]. These genomes were chosen because they are related, colinear and have an average genome size of 39.4 kbp (range: 31.5–41.7) and G + C mol% content of 50.7 (range: 42.6–61.8; Table S1). The testing dataset also contained the Pelagibacter phage HTVC011P genome, used as outlier for the T7-like phages. For this dataset of 61 phages, the intergenomic similarities were calculated with the following tools: Sequence Demarcation Tool (SDT) [10], pairwise sequence comparison (PASC) [8], OrthoANI [4], Gegenees [6], and VIRIDIC.

Additionally, two *Salmonella* phages (GE_vB_N5 and FE_vB_N8) were used for the illustration of genome and alignment length differences. The genome of the K155 strain of the T7 phage was used to test the effect of genome permutations and reverse complementarity on the intergenomic distances. Lastly, two artificial DNA sequences were generated by (i) scrambling the T7 genome with Shuffle DNA, part of the Sequence Manipulation Suite [21] and (ii) using Vladimír Čermák's Random DNA Sequence Generator at http://www.molbiotools.com/randomsequencegenerator.html to generate a 39,937 bp (48.4% GC) sequence.

3. Results and Discussion

VIRIDIC calculates intergenomic similarities between pairs of viral genomes based on BLASTN alignments. Because these alignments depend on the BLASTN parameters used, we have tested four sets of such parameters (Figure 1). These ranged from "relaxed" (-word_size 7 -reward 2 -penalty -3 -gapopen 5 -gapextend 2) to "very stringent" (-word_size 28 -reward 1 -penalty -2). All four sets of parameters performed similarly for genomes with a higher degree of similarity. However, for more distant genomes, the calculated similarity values were significantly lower for the "very stringent" parameters, compared with the "relaxed" ones (see Figure 1). This difference was expected, because at "very stringent" parameters, BLASTN produces alignments only for highly similar genomic regions, and thus the regions of lower similarity are not taken into account when calculating the intergenomic similarity values.

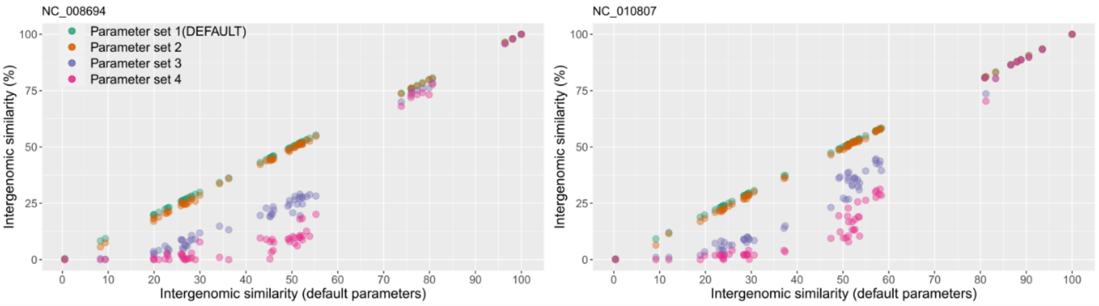

Figure 1. Comparison between the intergenomic similarity values produced with the default BLASTN alignment parameters (parameter set 1: -word_size 7 -reward 2 -penalty -3 -gapopen 5 -gapextend 2) and parameter sets of increasing stringency. Parameter set 2: "-word_size 11 -reward 2 -penalty -3 -gapopen 5 -gapextend 2". Parameter set 3: "-word_size 20 -reward 1 -penalty -2". Parameter set 4: "-word_size 28 -reward 1 -penalty -2". For illustration, the similarity values between two viral genomes (NCBI accession NC_008694 and NC_010807) and all the other genomes in the benchmarking dataset were chosen. On the X axis are plotted the intergenomic similarity values as calculated with the parameter set 1. On the Y axis are plotted the intergenomic similarity values as calculated with each of the four parameter sets. The plot was generated with the ggplot2 R package [22].

Taking the above findings into consideration, we have chosen the "relaxed" parameter set 1 as the default for VIRIDIC. The other, more stringent parameter datasets are made available for the user because they significantly decrease computational times, which can be most advantageous when desiring to cluster at high similarity thresholds (e.g., 90–100%) a large number of viral genomes, as for example found in viral metagenomic studies. On a benchmarking dataset of 61 phage genomes, VIRIDIC needed 270 s with the default parameters, and only 56 s with the most stringent parameters. Because in the range 90–100% intergenomic similarity, the "very stringent" parameters produced only a small decrease in similarity (see Figure 1), these parameters could be used in viral metagenomic datasets to enable clustering of highly related genomes (also discussed below). However, if clustering at lower similarity thresholds is desired, the relaxed parameter set 1 should be used.

Further, we have compared the intergenomic similarity values produced by VIRIDIC with those calculated manually from BLASTN alignments (the "traditional" method used by ICTV for phage classification) and those calculated by other different tools (see Figure 2). VIRIDIC showed the highest agreement with the manually calculated similarity values, being able to correctly indicate genome pairs with low similarity. In contrast, most of the other tools either gave artificially high similarity values for distant genomes (OrthoANI, SDT, EMBOSS Stretcher), they significantly deviated from the traditional method (PASC, Gegenees BLASTN 5%), or they were not linear with respect to a type species (Gegenees BLASTN 0%). The artificially high similarity values were likely due to their calculation only for the aligned part of the genomes. When the intergenomic similarity is normalized only to the alignment length, even if only a small region is aligned between two genomes in a pair, the outputted similarity can be high. Instead, VIRIDIC normalizes the number of aligned bases between the two genomes in a pair to the lengths of both genomes, and thus estimates better the similarity between distant genomes. One such example is the pair between the genomes of Escherichia coli T7 phage and Pelagibacter phage HTVC011P. When visualizing the alignment between these two genomes (see Figure 3), it is clear that only a small portion of their genomes align. VIRIDIC reported a 0.34% similarity for this genome pair. However, other tools reported similarity values of 18.13% (PASC) and even >42% (SDT and Emboss Stretcher), see Table S1. ANI tools have been extensively used in bacterial classification [1,4], and to a certain degree in phage classification [23,24], because their results mimic those of nucleic acid hybridizations. In our study, OrthoANI gave a similarity value of zero between the T7 phage and the Pelagibacter phage HTVC011P or the two artificially generated sequences. However, for the rest of the T7 phages, the OrthoANI values plateaued at an artificially high 62% similarity, in agreement with the previous observation that ANI values below 75% are meaningless [25].

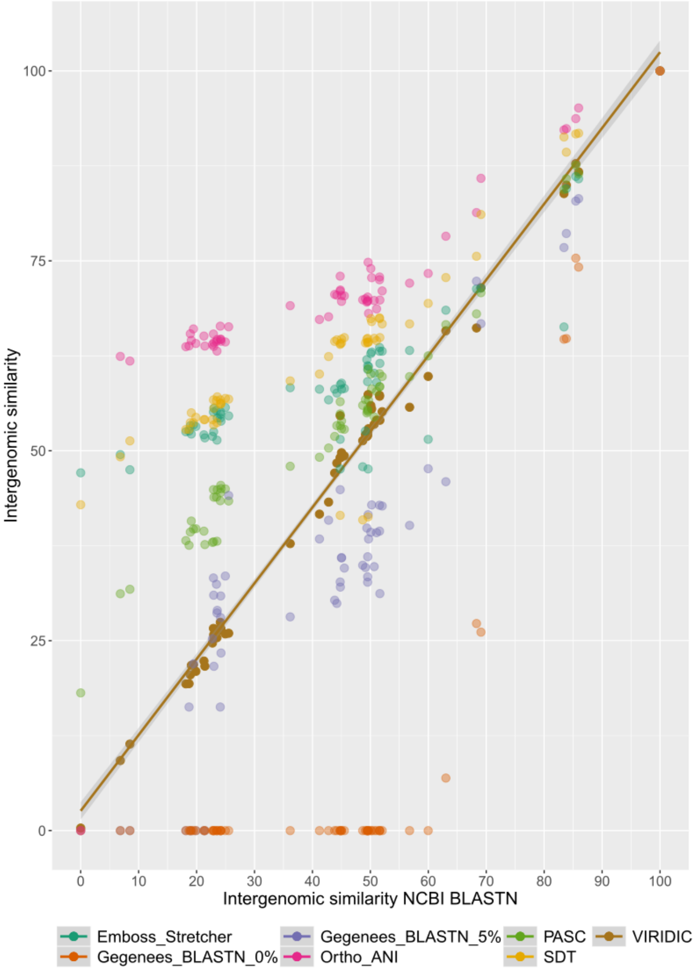

Figure 2. Plot comparing intergenomic similarity values generated by different tools (on the Y axis) with those generated by the traditional method used by ICTV (on the X axis). The plot was generated with the ggplot2 R package [22]. Data used for this plot are found in Table S1.

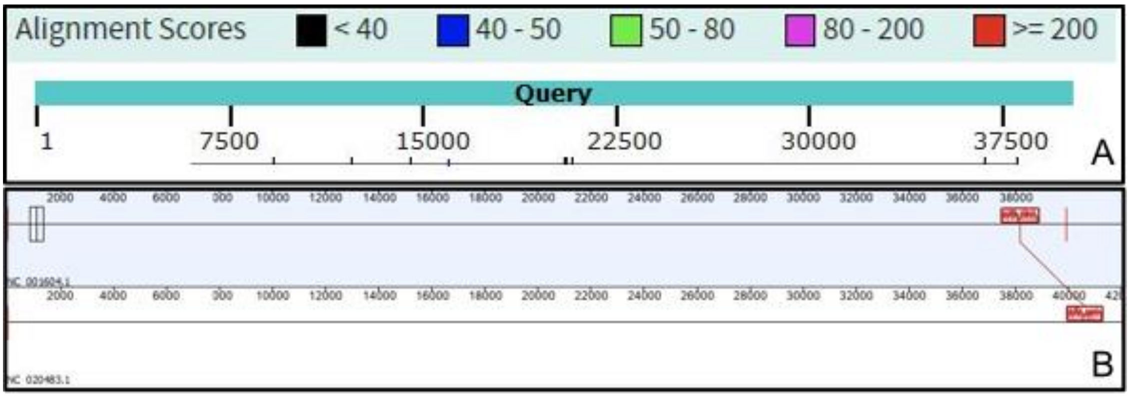

Figure 3. Genome alignments of the Escherichia coli T7 phage (NC_001604.1) and Pelagibacter phage HTVC011P (NC_020483.1) using (**A**) NCBI BLASTN, with the T7 genome as query; and (**B**) progressiveMAUVE plugin from Geneious software [26].

Following the calculation of the intergenomic similarities, VIRIDIC clusters and graphically represents these in a heatmap visualization (see Figure 4). Due to the color-coding, groups of related phages can easily be recognized visually. Furthermore, if the results of the clustering are not satisfying, different clustering methods can be tested without recalculating the intergenomic similarities, which are the most time consuming. This is especially easy in the web-service version of VIRIDIC, which provides access through a graphical interface to many parameters for clustering and heatmap visualization. It is important to note that, although VIRIDIC performs a hierarchical clustering, the resulting tree is not a representation of the evolutionary paths and evolutionary distances between the different phages. To avoid such confusions, the tree resulting from clustering is only used to generate the heatmap and is not visualized along its side. To reconstruct the phylogeny between or within the different virus clusters identified with VIRIDIC, further complementary phylogenetic analyses (e.g., core protein phylogeny) should be performed [27].

In the heatmap, for display purposes, the similarity values have been rounded to the first decimal. This rounding can hide minute differences between almost identical phages. These differences will be visible however in the similarity table, another output of VIRIDIC, where the similarity values are represented up to the third decimal.

A third output of VIRIDIC is a cluster table, in which the phage genomes are grouped into putative species and genera, based on user-set similarity thresholds (default 95% for species and 70% for genus). In the case of genomes having similarity values at the threshold border, the clusters in this table could represent sub-clusters of species or genus level clusters identified by eye in the heatmap. In such cases, the user should decide which are the best clusters. In conjunction with the stand-alone VIRIDIC version, the cluster table can be used for de-replication of large datasets of viral genomes.

The thresholds that should be used for grouping viruses at different taxonomic levels is an ongoing discussion in virology and is viral family-specific in most cases. A 95% threshold has been proposed for the species level, similarly to the threshold for bacterial and archaeal species [28]. Depending how this threshold is calculated, it can signify different degrees of similarity. For example, several metagenomic studies grouped viral contigs into populations using an ANI of >95% over at least 80% of their genes [29,30]. Empirical evidence from a large marine metagenomic study suggests indeed that at this threshold, dsDNA viruses form distinct genotypic clusters. Within these clusters, the frequency of homologous recombination between the individual viruses is presumably higher than with individuals from other similar groups, being thus consistent with the biological species definition (see Gregory et al. [29]). This threshold is the equivalent of 76% intergenomic similarity, as calculated by VIRIDIC and by the traditional ICTV method.

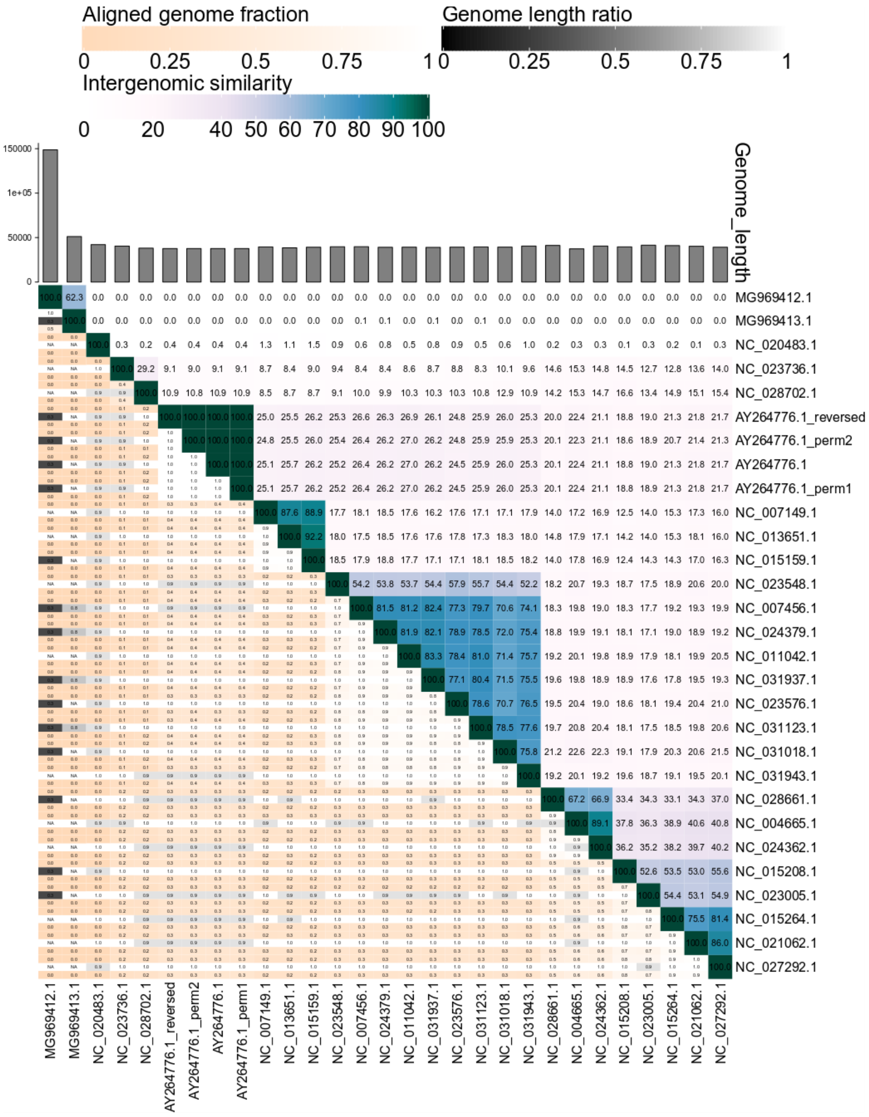

Figure 4. VIRIDIC generated heatmap incorporating intergenomic similarity values (right half) and alignment indicators (left half and top annotation). In the right half, the color-coding allows a rapid visualization of the clustering of the phage genomes based on intergenomic similarity: the more closely-related the genomes, the darker the color. The numbers represent the similarity values for each genome pair, rounded to the first decimal. In the left half, three indicator values are represented for each genome pair, in the order from top to bottom: aligned fraction genome 1 (for the genome found in this row), genome length ratio (for the two genomes in this pair) and aligned fraction genome 2 (for the genome found in this column). The darker colors emphasize low values, indicating genome pairs where only a small fraction of the genome was aligned (orange to white color gradient), or where there is a high difference in the length of the two genomes (black to white color gradient). The aligned genome fractions are expected to decrease with increasing the distance between the phages. Therefore, darker colors should correspond to genome pairs with low similarity values, and whiter colors to genome pairs with higher similarity values. Similarly, more closely-related viruses are expected to have similar lengths. Therefore, if low genome length ratios correspond to genome pairs with high similarity (e.g., MG969412.1 and MG969413.1 have a 62.4% similarity, but only 0.3 genome length fraction), this signals that the pair needs to be investigated further before being classified. The genome of the K155 strain of the T7 phage (AY264776.1) and its permuted (AY264776.1_perm1 and AY264776.1_perm2) and reversed complemented (AY264776_reversed) variants presented no significant differences between their intergenomic similarity values.

When comparing viral genomes of different length, additional information can help to better interpret the intergenomic similarity values. This is the case especially when the two genomes in a pair share a high degree of similarity, as for example the pair between a complete and a partial genome, or between one smaller genome which is very similar to a region of a much bigger genome. For this purpose, VIRIDIC calculates three additional indicators of the alignment ability to capture the relatedness between viruses—the aligned fraction for genome 1, the length ratio between genome 1 and genome 2, and the aligned fraction for genome 2. Then it displays the three indicators in a color-coded manner in the heatmap, as a visual aid for the user to spot genome pairs of different lengths or partial alignments (see Figure 4). Even more, the length of each genome is plotted as an annotation along the columns of the heatmap.

The intergenomic similarity values calculated by VIRIDIC are not influenced either by genome permutations, or the genomes being in different directions (see Figure 4). However, the values will be influenced by the use of draft genomes at the scaffold level, which contain long stretches of "N", because BLASTN is ignoring these regions. Therefore, it is not recommended to use such genomes. The intergenomic similarities will be influenced (underestimated) by the presence of repeats, because overlapping alignments are de-replicated, but the score is still normalized to the whole genome length. Therefore, for phages with repeats, for example with long terminal repeats, it is important to compare genomes with a single repeat copy.

In terms of sensitivity, VIRIDIC can align genomic regions having as low as 65% similarity and as little as 140 bases in length when using the BLASTN parameter set 1. Shorter regions, but of higher similarity (e.g., 30 bases length and 87% similarity) are also detected. Therefore, VIRIDIC will not capture the relationships between those viruses which have regions of similarity of less than 65%, a limitation inherent to BLASTN. Generally, nucleotide-based alignments are unable to capture similarity lower than 50% (two random DNA sequences can produce alignments of 50% similarity). A protein bases analysis is thus recommended to clarify phylogenetic relationships between distantly-related phages.

The VIRIDIC web-service provides a graphical interface for running VIRIDIC remotely and it is meant for small- to medium-sized projects, ideally not bigger than 200–300 viral genomes. The stand-alone program can be run from the command line in Linux and thus, it can be integrated into bioinformatics pipelines. Furthermore, depending on the configuration of the computational resources, it can analyze a significantly larger number of viral genomes than the web-service. In the VIRIDIC workflow, there are two computational intensive steps, the BLASTN step and the calculation of intergenomic similarity matrix step. The computational requirements of the second step increase exponentially with the number of viral genomes to be compared. For example, we have run two projects, one with 169 and the other with 1236 viral genomes (size range 30–150 kb), on a Linux server with 40 central processing unit (CPU) cores and 256 GB RAM memory. The first project finished in 10 min. The second project finished in 19.5 h. For the BLASTN step, the number of CPU cores to be used can be controlled via a command line parameter. For the calculation of intergenomic similarity matrix, all available CPU cores will be used.

VIRIDIC offers several advantages compared to other similar tools. First, it provides a better estimation of the similarity between phage genomes, especially for the more distantly-related ones. Second, it can be used in a high-throughput manner, allowing the analysis of datasets containing hundreds (the web-service) and even thousands (the stand-alone version) of phage genomes. Third, it generates an informative heatmap, which incorporates not only the similarity values, but also information about the genome lengths and aligned genome fraction, useful for evaluating the ability of the similarity values to capture the virus relatedness.

Supplementary Materials: The following are available online at http://www.mdpi.com/1999-4915/12/11/1268/s1, Table S1: the complete list of viral genomes used for benchmarking VIRIDIC, their genome properties and intergenomic similarity values produced by different tools.

Author Contributions: Conceptualization, C.M. and A.M.K.; methodology, C.M.; software, C.M.; validation, C.M. and A.M.K.; formal analysis, C.M., A.V., and A.M.K.; data curation, C.M. and A.M.K.; writing—original draft preparation, C.M. and A.M.K.; writing—review and editing, C.M., A.V., and A.M.K.; visualization, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We are thankful to Matthias Schroeder (Institute for Chemistry and Biology of the Marine Environment, Oldenburg) for excellent IT support through the project and for the preparation of the VIRIDIC singularity container.

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

References

1. Goris, J.; Konstantinidis, K.T.; Klappenbach, J.A.; Coenye, T.; Vandamme, P.; Tiedje, J.M. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **2007**, *57*, 81–91. [CrossRef] [PubMed]
2. Yoon, S.-H.; Ha, S.-M.; Lim, J.; Kwon, S.; Chun, J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Leeuwenhoek* **2017**, *110*, 1281–1286. [CrossRef] [PubMed]
3. Han, N.; Qiang, Y.; Zhang, W. ANItools web: A web tool for fast genome comparison within multiple bacterial strains. *Database* **2016**, *2016*. [CrossRef] [PubMed]
4. Lee, I.; Ouk Kim, Y.; Park, S.-C.; Chun, J. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.* **2016**, *66*, 1100–1103. [CrossRef]
5. Ceyssens, P.-J.; Miroshnikov, K.; Mattheus, W.; Krylov, V.; Robben, J.; Noben, J.-P.; Vanderschraeghe, S.; Sykilinda, N.; Kropinski, A.M.; Volckaert, G.; et al. Comparative analysis of the widespread and conserved PB1-like viruses infecting Pseudomonas aeruginosa. *Environ. Microbiol.* **2009**, *11*, 2874–2883. [CrossRef]
6. Agren, J.; Sundström, A.; Håfström, T.; Segerman, B. Gegenees: Fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. *PLoS ONE* **2012**, *7*, e39107. [CrossRef]
7. Richter, M.; Rosselló-Móra, R.; Oliver Glöckner, F.; Peplies, J. JSpeciesWS: A web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* **2016**, *32*, 929–931. [CrossRef]
8. Bao, Y.; Chetvernin, V.; Tatusova, T. PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses. *Viruses* **2012**, *4*, 1318–1327. [CrossRef] [PubMed]
9. Bao, Y.; Chetvernin, V.; Tatusova, T. Improvements to pairwise sequence comparison (PASC): A genome-based web tool for virus classification. *Arch. Virol.* **2014**, *159*, 3293–3304. [CrossRef]
10. Muhire, B.M.; Varsani, A.; Martin, D.P. SDT: A virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS ONE* **2014**, *9*, e108277. [CrossRef]
11. Noé, L.; Kucherov, G. YASS: Enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* **2005**, *33*, W540–W543. [CrossRef]
12. Mahadevan, P. An Analysis of Adenovirus Genomes Using Whole Genome Software Tools. *Bioinformation* **2016**, *12*, 301–310. [CrossRef]
13. Darling, A.E.; Mau, B.; Perna, N.T. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **2010**, *5*, e11147. [CrossRef]
14. Meier-Kolthoff, J.P.; Göker, M. VICTOR: Genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* **2017**, *33*, 3396–3404. [CrossRef] [PubMed]
15. R Core Team. R: A Language and Environment for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 12 September 2020).
16. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinf.* **2009**, *10*, 421. [CrossRef]
17. Meier-Kolthoff, J.P.; Auch, A.F.; Klenk, H.-P.; Göker, M. Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinf.* **2013**, *14*, 60. [CrossRef]
18. Müllner, D. Fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *J. Stat. Soft.* **2013**, *53*. [CrossRef]

19. Gu, Z.; Eils, R.; Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef]
20. Haft, D.H.; DiCuccio, M.; Badretdin, A.; Brover, V.; Chetvernin, V.; O'Neill, K.; Li, W.; Chitsaz, F.; Derbyshire, M.K.; Gonzales, N.R.; et al. RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **2018**, *46*, D851–D860. [CrossRef]
21. Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **2000**, *28*, 1102–1104. [CrossRef]
22. Hadley, W. *Ggplot2. Elegrant Graphics for Data Analysis*, 2nd ed.; Springer: Cham, Switzerland, 2016; ISBN 978-3-319-24277-4.
23. Accetto, T.; Janež, N. The lytic *Myoviridae* of *Enterobacteriaceae* form tight recombining assemblages separated by discontinuities in genome average nucleotide identity and lateral gene flow. *Microb. Genom.* **2018**, *4*. [CrossRef]
24. Oliveira, H.; Sampaio, M.; Melo, L.D.R.; Dias, O.; Pope, W.H.; Hatfull, G.F.; Azeredo, J. Staphylococci phages display vast genomic diversity and evolutionary relationships. *BMC Genom.* **2019**, *20*, 357. [CrossRef]
25. Rodriguez, R.L.M.; Konstantidinis, K.T. Bypassing Cultivation to Identify Bacterial Species. *Microbe* **2014**, *9*, 211–218. [CrossRef]
26. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649. [CrossRef] [PubMed]
27. Barylski, J.; Enault, F.; Dutilh, B.E.; Schuller, M.B.; Edwards, R.A.; Gillis, A.; Klumpp, J.; Knezevic, P.; Krupovic, M.; Kuhn, J.H.; et al. Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages. *Syst. Biol.* **2020**, *69*, 110–123. [CrossRef]
28. Konstantinidis, K.T.; Tiedje, J.M. Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2567–2572. [CrossRef]
29. Gregory, A.C.; Zayed, A.A.; Conceição-Neto, N.; Temperton, B.; Bolduc, B.; Alberti, A.; Ardyna, M.; Arkhipova, K.; Carmichael, M.; Cruaud, C.; et al. Marine DNA viral macro- and microdiversity from Pole to Pole. *Cell* **2019**. [CrossRef]
30. Brum, J.R.; Ignacio-Espinoza, J.C.; Roux, S.; Doulcier, G.; Acinas, S.G.; Alberti, A.; Chaffron, S.; Cruaud, C.; Vargas, C.d.; Gasol, J.M.; et al. Patterns and ecological drivers of ocean viral communities. *Science* **2015**, *348*, 1261498. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.