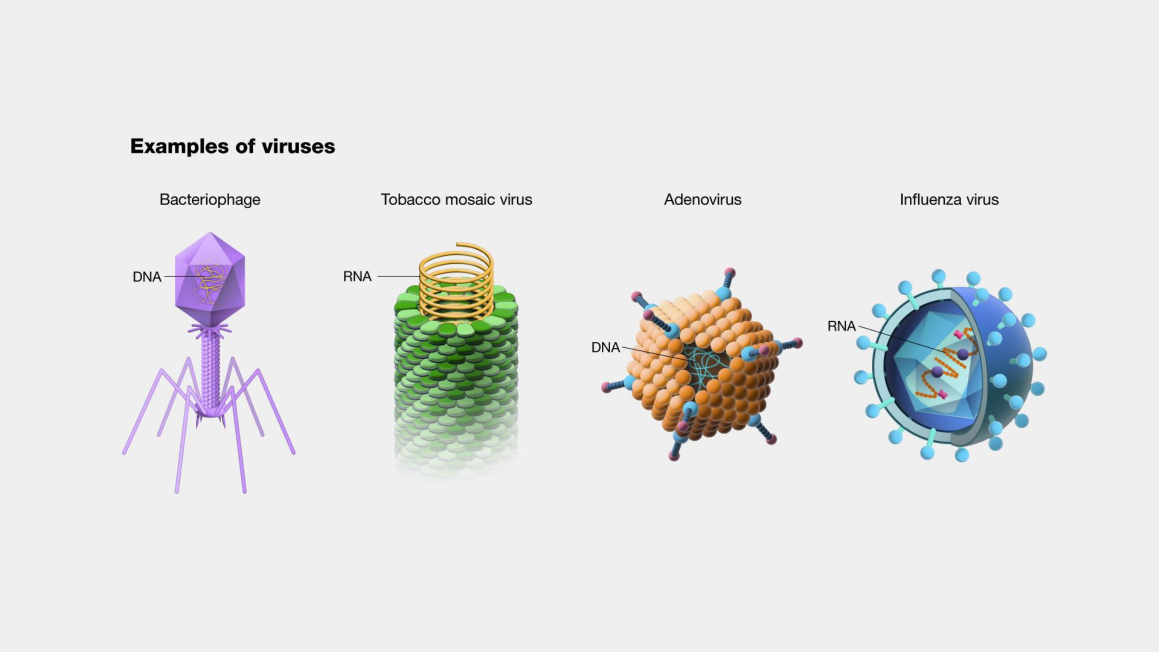
PHROG: families of prokaryotic virus proteins clustered using remote homology

Viruses

A virus is a microscopic parasite (generally much smaller than a bacterium) consisting of a segment of nucleic acid (either DNA or RNA) surrounded by a protein coat.

Well-known examples of viruses causing human diseases such as AIDS, COVID-19, measles and smallpox.



Viruses are abundant and diverse biological entities; their diversity is high, both in terms of the number of different protein families encountered and in the sequence heterogeneity of each protein family.

The recent increase in sequenced viral genomes constitutes a great opportunity to gain new insights into this diversity and consequently urges the development of annotation resources to help functional and comparative analysis.

Viruses infecting prokaryotes (1) out-number eukaryotic viruses in some ecosystems (*) and (2) represent the great majority of the viruses found in viral metagenomes (*).

Viral proteins clustering

The growing amount of viral sequences produced nowadays, especially from metagenomes, calls for a need in resources that are able to assign functions to viral proteins in order to improve functional and comparative analyses.

As determining information for newly identified genes in sequences relies on finding an annotated homologous sequence using similarity searches, a set of well-annotated proteins that can be used for future work is particularly important.

In order to build such a reference set of proteins, a classical way is to organize the proteins into homologous groups and to annotate these groups.

Several methods have been used to cluster viral proteins into either homologous or orthologous groups.

Homolog or homologue. A gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of speciation (**ortholog**) or to the relationship betwen genes separated by the event of genetic duplication (paralog).

Viral proteins clustering

1. an approach based on the identification of genome-specific best hits that are joined to form clusters of orthologs that has been used for *the pVOGs database (Prokaryotic Virus Orthologous Groups).*

2. a similar approach based on best-hit triangles has been implemented recently in *eggNOG (evolutionary genealogy of genes*), which integrates an additional step of in-paralogs detection and the identification of fused genes.

3. clustering-based approaches have also been used to compute groups of homologous viral proteins

Despite using different clustering strategies, all these methods rely on similarity search results generated by BLAST or faster equivalent such as MMseqs or DIAMOND, and none integrate remote homology detection.

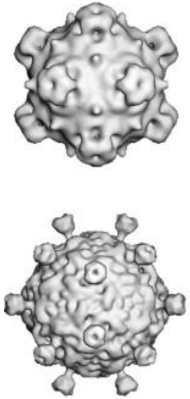
MMseqs (Many-agains-Many sequence searching) is a software suite to search and cluster huge protein and nucleotide sequence sets (https://github.com/soedinglab/MMseqs2).
Diamond is a sequence aligner for protein and translated DNA searches, designed for high performance analysis of big sequence data (https://github.com/bbuchfink/diamond).

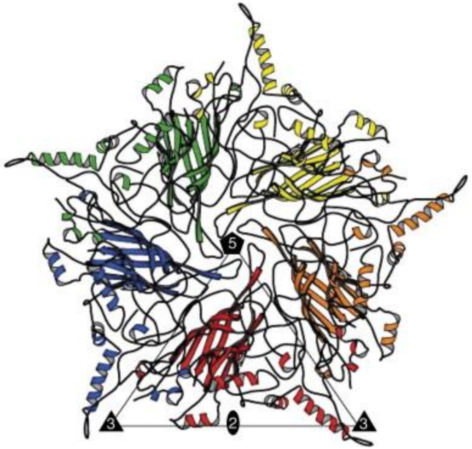
Remote homology

Any pair of proteins within the same superfamily classification was considered as a homology. Remote homologues can be defined as pairs of homologous proteins having a pairwise sequence identity <=25% as calculated using structurally equivalent regions only.

Viral proteins clustering problem

Viruses are known to have distant evolutionary relationships, resulting in distant sequence similarities not always captured by sequence comparison tools.





This can be illustrated by the *Microviridae* family, whose members all encode homologous major capsid and replication initiation proteins in their ~5 kb genomes.

As the origin of the family is ancient, homologs are difficult to detect using sequence similarity search tool such as BLAST.

Example:
For two distantly related *Microviridae*, Spiroplasma phage 4 and Enterobacteria phage phiX174 belonging respectively to the *Gokushovirinae* and *Bullavirinae* sub-families respectively,
the best BLASTp hit between their capsid proteins only exhibit a bit-score of 38.5 (corresponding to an E-value of 0.14 on a ~1 million protein database) and an alignment coverage below 30% for the two proteins.

Bit score is an important measure that gives an indication about the statistical significance of an alignment. In simple terms, the higher the bit score, the more similar the two sequences are. Bit scores below 50 are generally assumed to be untrustworthy.
E-value represents the expectation of finding that sequence by random chance. So if you search a short sequence you are likely to have a lot more hits with high e-value (low significance), and if you search a long sequence you are likely to have fewer hits with lower e-value (greater significance).

Solution of the problem and the main idea of article

Clustering viral proteins into homologous groups in two steps:

1) proteins are first gathered based on similarity search results (score >30 and coverage >80%).

2) HMM profiles generated for each protein cluster are then compared to each other and grouped (coverage >60%, probability >90 %) into super-clusters termed as PHROGs (Prokaryotic Virus Remote Homologous Groups).

Profile Hidden Markov Models

They are one of the computational algorithms used for predicting protein structure and function, identifies significant protein sequence similarities allowing the detection of homologs and consequently the transfer of information, i.e. sequence homology-based inference of knowledge.

Profile HMMs are probabilistic models that encapsulate the evolutionary changes that have occurred in a set of related sequences (i.e. a multiple sequence alignment). To do so, they capture position-specific information about how conserved each amino acid is in each column of the alignment (see Figure on the next slide).

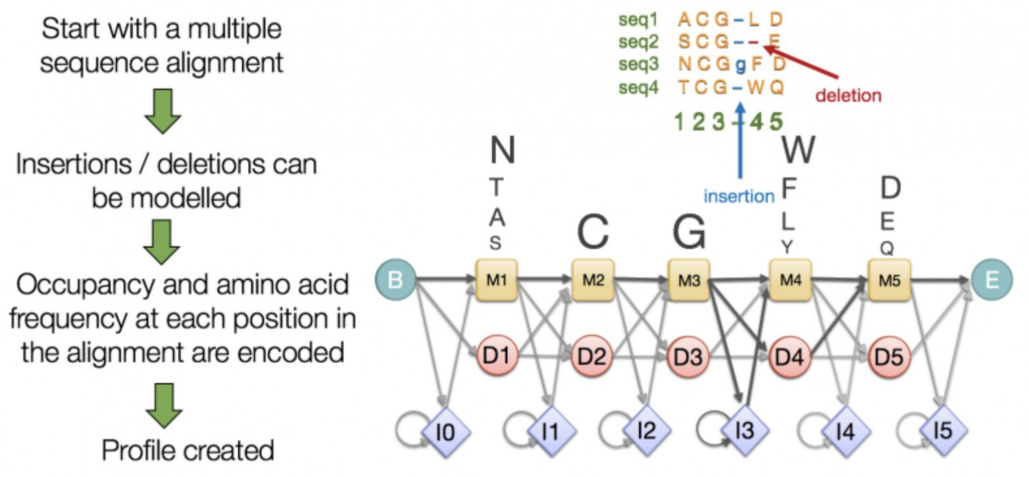
The model also captures important information such as where gaps and insertions have occurred. Unlike other sequence homology detection algorithms, profile HMMs use position dependent gap penalties and substitution probabilities which better reflect biological reality.

Profile Hidden Markov Models

The boxes in yellow are the match states (M).
In the M state the probability distribution is the frequency of the amino acids in that position.

The row of diamond shaped states are insert states (I) which are used to model highly variable regions in the alignment.

The circular states are delete states (D).
These are called silent states since they do not match any residues, and they are there merely to make it possible to jump over one or more columns in the alignment.


Start with a multiple sequence alignment

Insertions / deletions can be modelled

Occupancy and amino acid frequency at each position in the alignment are encoded

Profile created

A profile HMM modelling a multiple sequence alignment.

The final probabilistic model conveys the estimation of the observed frequencies of the amino acids in each position, as well as the transitions between the amino acids derived from the observed occupancy of each position in a multiple sequence alignment.

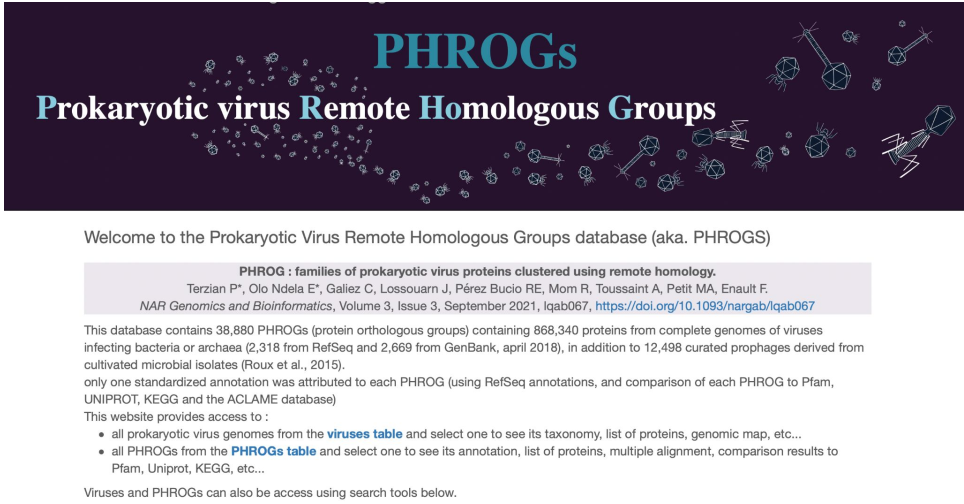
Retrieved from: https://www.ebi.ac.uk/training/online/courses/pfam-creating-protein-families/what-are-profile-hidden-markov-models-hmms/

PHROG (Prokaryotic Virus Remote Homologous Groups)

is *a library of viral protein families* generated using *a new clustering approach based on remote homology detection by HMM profile-profile comparisons.*

Data:
- 17 473 reference viruses of prokaryotes
- 868 340 of the total 938 864 proteins were grouped into 38 880 clusters

Advantages:
a 2-fold deeper clustering than using a classical strategy based on BLAST-like similarity searches



PHROGs

Prokaryotic virus Remote Homologous Groups

Welcome to the Prokaryotic Virus Remote Homologous Groups database (aka. PHROGS)

PHROG : families of prokaryotic virus proteins clustered using remote homology.
Terzian P*, Olo Ndela E*, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, Toussaint A, Petit MA, Enault F.
NAR Genomics and Bioinformatics, Volume 3, Issue 3, September 2021, lqab067, https://doi.org/10.1093/nargab/lqab067

This database contains 38,880 PHROGs (protein orthologous groups) containing 868,340 proteins from complete genomes of viruses infecting bacteria or archaea (2,318 from RefSeq and 2,669 from GenBank, april 2018), in addition to 12,498 curated prophages derived from cultivated microbial isolates (Roux et al., 2015).
only one standardized annotation was attributed to each PHROG (using RefSeq annotations, and comparison of each PHROG to Pfam, UNIPROT, KEGG and the ACLAME database)
This website provides access to :

- all prokaryotic virus genomes from the **viruses table** and select one to see its taxonomy, list of proteins, genomic map, etc...
- all PHROGs from the **PHROGs table** and select one to see its annotation, list of proteins, multiple alignment, comparison results to Pfam, Uniprot, KEGG, etc...

Viruses and PHROGs can also be access using search tools below.

> “Hopefully, PHROG will be a useful tool to better annotate future prokaryotic viral sequences thus helping the scientific community to better understand the evolution … of these entities.”

https://phrogs.lmge.uca.fr

Step-by-step (library generation)

1. Two datasets of archaeal and bacterial viruses formation.

2318 reference sequences of viruses infecting prokaryotes were retrieved in RefSeqVirus genomes (as of April 2018). Viruses included in the pVOGs database and complete virus genomes from GenBank were also downloaded.

The resulting 4975 completely sequenced viruses (2315 from RefSeqVirus, 686 from pVOGs and 1986 from GenBank) will be refered as RefVirus. To this dataset, 12 498 previously published curated viral sequences derived from cultivated microbial isolates were added (is here termed ProVirus).

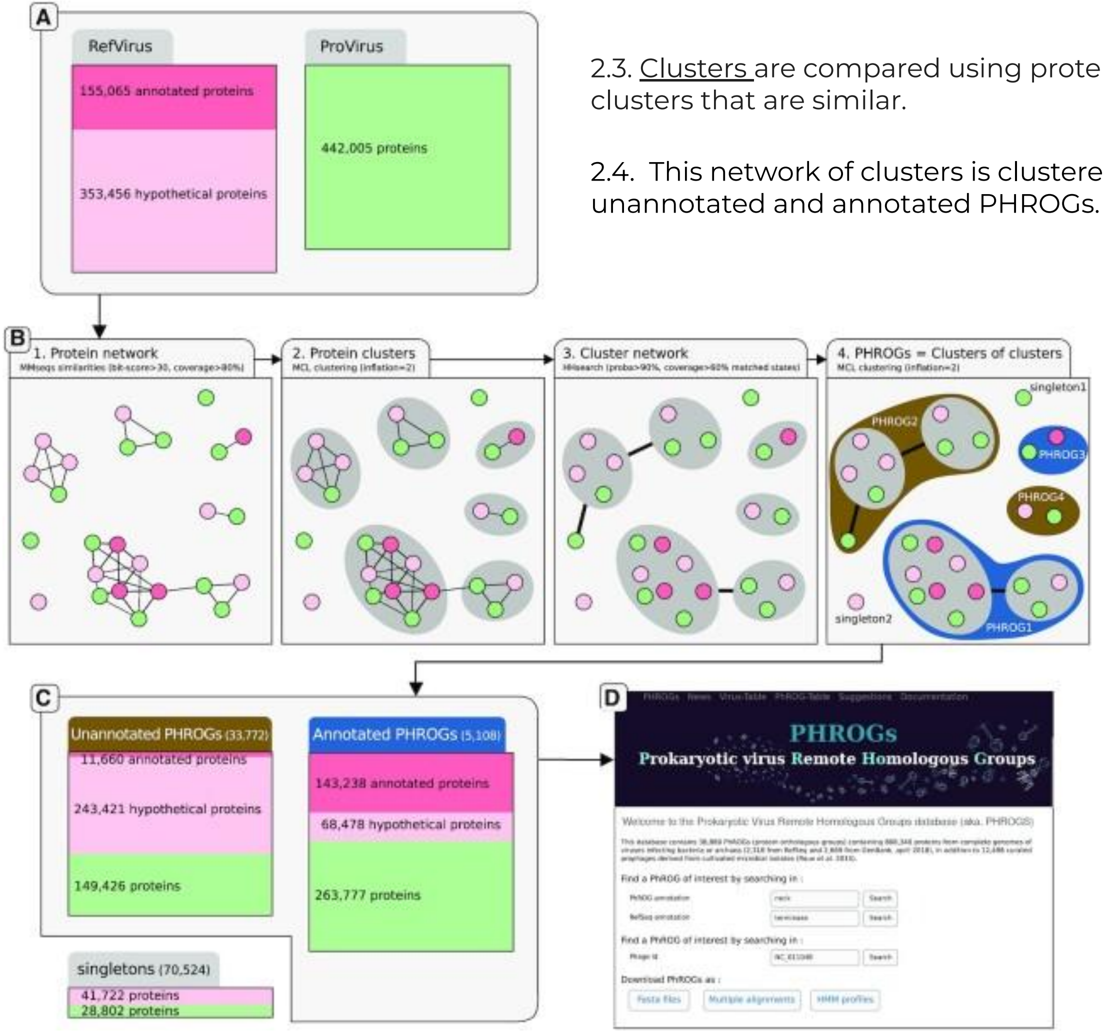
In total, 496 859 proteins from complete viral genomes (i.e. RefVirus) and 442 005 proteins from (pro)viruses (i.e. ProVirus) were collected.

2. Clustering procedure.

2.1. Protein network built from pairwise sequence similarities, each dot/vertex representing a protein (green for ProVirus proteins, red for annotated RefVirus proteins and pink for unannotated RefVirus), linked by edges if the two proteins are similar.

2.2. Protein clusters are identified by applying MCL on to this network, and clusters are depicted as gray circle.

MCL algorithm is short for the **Markov Cluster Algorithm**, a fast and scalable unsupervised cluster algorithm for graphs (also known as networks) based on simulation of (stochastic) flow in graphs. It has found usage in bioinformatics and other disciplines (https://micans.org/mcl/)


A
RefVirus
155,065 annotated proteins
353,456 hypothetical proteins
ProVirus
442,005 proteins
B
1. Protein network
MMseqs similarities (bit-score>30, coverage>80%)
2. Protein clusters
MCL clustering (inflation=2)
3. Cluster network
HHsearch (proba>90%, coverage>60% matched states)
4. PHROGs = Clusters of clusters
MCL clustering (inflation=2)
singleton1
PHROG2
PHROG3
PHROG4
singleton2
PHROG1
C
Unannotated PHROGs (33,772)
11,660 annotated proteins
243,421 hypothetical proteins
149,426 proteins
Annotated PHROGs (5,108)
143,238 annotated proteins
68,478 hypothetical proteins
263,777 proteins
singletons (70,524)
41,722 proteins
28,802 proteins
D
PHROGs News Virus-Table PHROG-Table Suggestions Documentation
PHROGs
Prokaryotic virus Remote Homologous Groups
Welcome to the Prokaryotic Virus Remote Homologous Groups database (aka. PHROGS)
Find a PhROG of interest by searching in :
PhROG annotation
neck
Search
RefSeq annotation
terminase
Search
Find a PhROG of interest by searching in :
Phage id
NC_011048
Search
Download PhROGs as :
Fasta files
Multiple alignments
HMM profiles

2.3. Clusters are compared using protein profiles and edges are drawn for pairs of protein clusters that are similar.

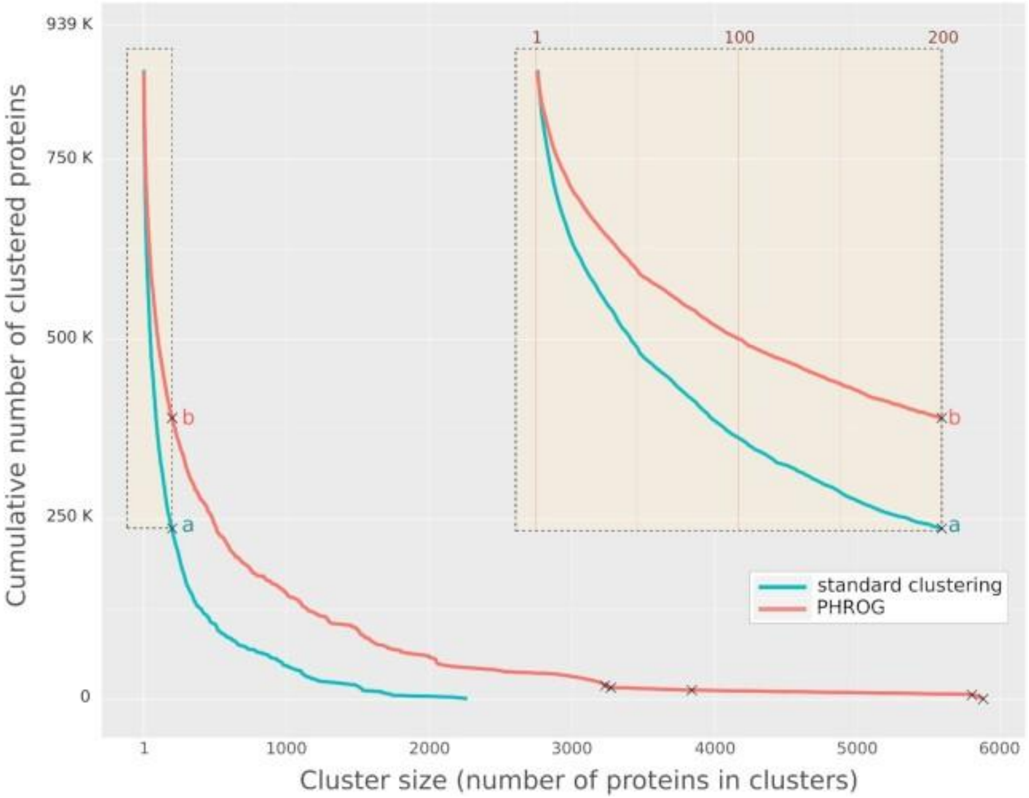
2.4. This network of clusters is clustered into PHROGs, depicted as dark brown and blue for unannotated and annotated PHROGs.

3. Description of the number of annotated and unannotated PHROGs with the number and origin of proteins involved (red and green for RefVirus and ProVirus, respectively).

4.. PHROGs website main page where users can search for viruses of interest.

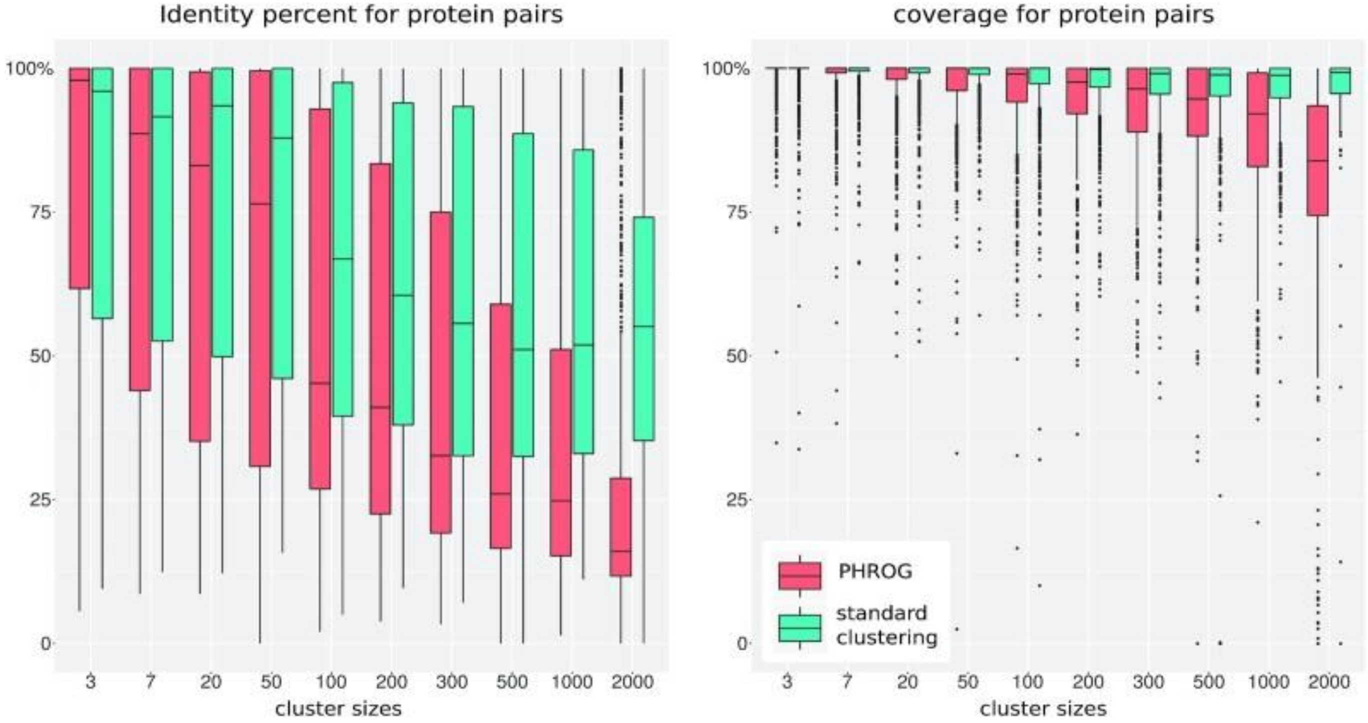
Comparing PHROGs to the standard clustering method

To estimate the performance of PHROGs, the same set of proteins were clustered using a classical approach:
sequence similarities were searched for all protein pairs using MMseqs and clustered using Markov clustering algorithm.



Cumulated number of clustered proteins. For example, point *a* means that for the standard clustering procedure, ~234 000 proteins are in clusters that contain at least 200 proteins, whereas for the PHROG procedure, ~390 000 proteins are in clusters >200 (point *b*).

Comparing PHROGs to the standard clustering method



Identity percent (**A**) and coverage (**B**) for protein pairs in the same clusters.
The clusters were separated according to interval of size, the first value «3» representing clusters that contain 3, 4, 5 or 6 proteins, «7» being clusters containing between 7 and 19 proteins, and the last interval «2000» being clusters >2000 proteins.

CONCLUSIONS

The PHROG database includes 17 473 genomes (or genome fragments) of viruses infecting a broad range of prokaryotic hosts (410 prokaryotic genera). The great majority of the 938 864 proteins encoded within these genomes could be assigned to a PHROG, and only 7.5% of them remained as singletons.

Due to the ancient origin of viruses, their gene families have a long evolutionary history and often encompass distant homologs not detected by standard sequence comparison tools such as BLAST. To be able to identify these distant homologs, new clustering strategy involved the use of HMM comparison tools was considered.

The clusters built here proved to be larger while remaining cohesive, leading to an increase in annotated proteins.

Annotations of PHROG families were performed based on Refseq and several other databases, and manually curated by experts, adding a real value to the PHROG database. These annotations will be updated in the future, as new phage functions are discovered.

A singleton is a read with a sequence that is present exactly once, i.e. is unique among the reads.