The GRAViTy 1.1 User Guide

Pakorn Aiewsakun[1] and Peter Simmonds[2]
[1] Department of Microbiology, Faculty of Science, Mahidol University, Bangkok, 10400 Thailand
[2] Nuffield Department of Medicine, University of Oxford, South Parks Road, Oxford, OX1 3SY, UK

Emails:
Pakorn Aiewsakun: Pakorn.Aiewsakun@gmail.com
Peter Simmonds: Peter.Simmonds@ndm.ox.ac.uk

Contents

Overview

"**G**enome **R**elationships **A**pplied to **Vi**rus **T**axonom**y**" or "**GRAViTy**" is a virus identification and classification framework based on the analysis of whole virus genomes (1, 2). GRAViTy assigns a virus to a reference taxonomic group, typically a family, using a range of virally encoded genomic features, including its genomic organisation (gene locations, order and orientations) and metrics of sequence similarity of genes to those of classified viruses.

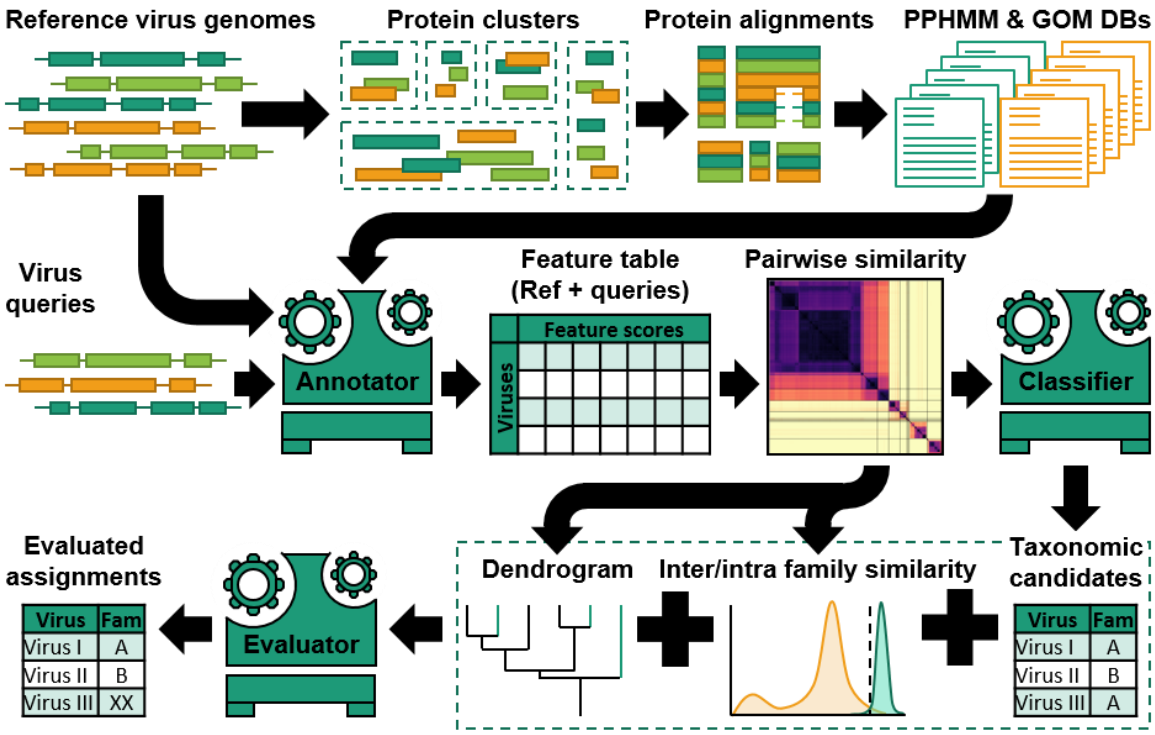

Figure 1 Overview of virus taxonomy prediction by GRAViTy

*Schematic diagram of the processing steps used to construct classifiers based on viruses with assigned taxonomic status (refence virus genomes) and the pipeline used to classify viruses of interest (virus queries). In summary, protein sequences were extracted from reference virus genomes, and were clustered based on pairwise BLASTp bit scores. Sequences in each cluster were then aligned to make an alignment, and turned into a protein profile Hidden Markov model (PPHMM). Reference genomes were subsequently scanned against the database of PPHMMs to determine the locations of their genes, and genomic organisation models (GOMs) for each virus family were constructed from this data. PPHMM and GOM databases are the main machinery of our genome annotator (Annotator). To classify viruses of interest, they, together with the reference viruses, will first be annotated by the annotator to produce a summary table containing information regarding the presence of genes and the degree of similarity of their genomic organisation to various reference families (Feature table). Pairwise similarity scores (composite generalised Jaccard similarity) will then be estimated, and this will be passed to the classifier to identify a taxonomic candidate for each query using the 1-nearest neighbour algorithm. A UPGMA dendrogram and a similarity acceptance cut-off for each virus family will also be estimated from the pairwise similarity scores, and these will be used by the evaluator to evaluate the taxonomic candidates. The analysis is performed separately for the 6 virus Baltimore groups in parallel, and those showing best matches are the finalised taxonomic assignments. Figure reproduced from **Aiewsakun P**, **Simmonds P**. 2018 (1).*

In more traditional approaches to virus classification, the degree of virus similarity – i.e. genetic relatedness – is typically estimated by using multiple sequence alignments of a few selected homologous proteins or genes. These are however limited by how 'alignable' or 'conserved' they are and often arbitrary decisions about which sites are homologous. Under the GRAViTy framework (**Figure 1**), the degree of similarity among a set of viruses is computed by comparing their sets of genes, and their gene locations, orders, and orientations through computation of a composite generalised Jaccard (CGJ) distance. These can be calculated between any pair of viruses, enabling GRAViTy to infer virus relatedness between very divergent, or indeed, assign two viruses as being evolutionarily independent. This capability enables a combined taxonomy of all viruses.

In addition, GRAViTy can propose assignment to a novel virus groups (typically families) with a high predictive value, where viruses display CGJ distances above the typical threshold for within-family distances. We have demonstrated GRAViTy's ability to currently differentiate "known viruses" from the "unknown ones", and to classify known viruses to their "correct" reference taxonomic groups at the family level. We found a very high level of concordance between GRAViTy assignments based purely on genomic features and the current taxonomy of eukaryotic viruses (1). However, the current family assignments of tailed phages to *Siphoviridae, Myoviride* and *Podoviridae* is discordant with their genomic relationships and other virological features (and with analysis by GRAViTy (2)). Virus relationships recovered by GRAViTy' support the ongoing complete re-classification of these viruses by the Bacterial and Archaeal virus sub-committee of the International Committee on Taxonomy of Viruses (ICTV).

Under the GRAViTy framework, viruses are annotated by using two databases, namely the "PPHMM" database and the "GOM" database. The PPHMM database contains protein profile hidden Markov models (PPHMMs) of various genes, which capture their molecular diversity and allow for highly divergent genes to be detected. The GOM database contains genomic organisation models (GOMs) of various virus groups. Just like a PPHMM, a GOM of a virus group captures the diversity of genomic organisations that differ among its members. These two databases are built based on reference virus genomes, e.g. those whose taxonomic assignments are officially recognised by the ICTV.

To annotate a virus, GRAViTy first 6-frame translates a virus genome and scans it against the PPHMM database. This process generates two signatures: 1) a PPHMM signature, which contains information about which genes are present or absent in its genome as well as the degrees to which the genome exhibits similarity to each of the PPHMMs, and 2) a PPHMM location signature, which contains information about where the genes are located and in what orientations. GRAViTy then scans the PPHMM location signature against the GOM database, and this process generates a GOM signature. A GOM signature contains information about the degree of similarity between the genomic organisation of a test virus genome sequence and those of various reference virus groups. Unlike many of traditional approaches of virus identification and classification, which typically represent a virus by its molecular sequence, a virus under the GRAViTy framework is represented by a PPHMM signature and a GOM signature. This process is repeated for all viruses submitted to GRAViTy.

GRAViTy then computes pairwise distances among the reference and test viruses based on their PPHMM and GOM signatures to estimate virus relatedness, depicted by using a dendrogram and more impressionistically as a heatmap. Subsequently, GRAViTy proposes candidate taxonomic groups for test viruses based on their closest reference viruses. The candidate assignments will then be evaluated according to the observed degrees of similarities to the closest references and their placements in the estimated dendrogram. Each virus group has its own similarity cut-off, depending on the distribution of intra/inter group similarity scores, and this is determined automatically by

GRAViTy. Some test viruses might be assigned as “Unclassified viruses” if they do not show close similarity to any reference viruses. Finally, based on the estimated dendrogram, an overall dissimilarity cut-off that best separates reference viruses into reference taxonomic groupings can be computed. This can then be used to suggest novel virus groupings for virus sequences that are submitted.

GRAViTy primarily acts as a guide towards classification using sequence-based metrics that recapitulate the current ICTV taxonomy. While it can propose family membership or assignments into a new virus group for test virus sequences, these are purely advisory. Similarly, it can provide genomics-based evidence in support of a classification proposal but its output should be regarded as advisory and other, often more pragmatic, factors may need to be taken into account. The analysis of genome relationships by other methods is advised, particularly in the assignments of novel virus families and orders.

Running GRAViTy

We have provided a web interface (at http://gravity.cvr.gla.ac.uk) for analysing test sequences by GRAViTy (Pipeline II). See below for help with the interpretation of outputs provided by the analysis.

The program can also be run through a program download at GitHub: Paiewsakun/GRAViTy. Two main programs are implemented in GRAViTy: GRAViTy_Pipeline_I and GRAViTy_Pipeline_II. In summary, GRAViTy_Pipeline_I is used construct reference PPHMM and GOM databases, and GRAViTy_Pipeline_II is used to identify and classify your viruses.

Installation

Execute the command "sudo pip install ." in the GRAViTy directory that contains the "setup.py" file. All dependencies should be installed for you. Note that this is an ALPHA version of the program, meaning that this collection of scripts likely contains a lot of bugs, and it is still under development… and hence the following disclaimer.

Disclaimer

The material embodied in this software is provided to you "as-is", “with all faults”, and without warranty of any kind, express, implied or otherwise, including without limitation, any warranty of fitness for a particular purpose, warranty of non-infringement, or warranties of any kind concerning the safety, suitability, lack of viruses, inaccuracies, or other harmful components of this software. There are inherent dangers in the use of any software, and you are solely responsible for determining whether this software is compatible with your equipment and other software installed on your equipment. You are also solely responsible for the protection of your equipment and backup of your data, and the developers/providers will not be liable for any damages you may suffer in connection with using, modifying, or distributing this software. Without limiting the foregoing, the developers/providers make no warranty that:

- the software will meet your requirements
- the software will be uninterrupted, timely, secure, or error-free
- the results that may be obtained from the use of the software will be effective, accurate, or reliable
- the quality of the software will meet your expectations
- any errors in the software will be corrected.

Software and its documentation made available here:

- could include technical or other mistakes, inaccuracies, or typographical errors. The developers/providers may make changes to the software or documentation made available here
- may be out of date, and the developers/providers make no commitment to update such materials.

The developers/providers assume no responsibility for errors or omissions in the software or documentation available from here.

In no event shall the developers/providers be liable to you or anyone else for any direct, special, incidental, indirect, or consequential damages of any kind, or any damages whatsoever, including without limitation, loss of data, loss of profit, loss of use, savings or revenue, or the claims of third parties, whether or not the developers/providers have been advised of the possibility of such

damages and loss, however caused, and on any theory of liability, arising out of or in connection with the possession, use, or performance of this software.

The use of this software is done at your own discretion and risk and with agreement that you will be solely responsible for any damage to your computer system or loss of data that results from such activities. No advice or information, whether oral or written, obtained by you from the developers/providers shall create any warranty for the software.

GRAViTy_Pipeline_I

Description

```
GRAViTy_Pipeline_I
      |_ReadGenomeDescTable
      |_PPHMMDBConstruction
      |_RefVirusAnnotator
      |_GRAViTyDendrogramAndHeatmapConstruction (optional)
      |_MutualInformationCalculator (optional)
```

The main purpose of `GRAViTy_Pipeline_I` is to:

- construct a PPHMM database and a GOM database from reference virus genomes
- annotate reference viruses using the PPHMM and GOM databases, generating PPHMM signatures, PPHMM location signatures, and GOM signatures for reference viruses
- (optional) analyse reference viruses and produce a heatmap and dendrogram that depict the degree of (dis)similarity among them
- (optional) determine which genes support the current virus taxonomy and which genes do not.

You need will need to provide a table of descriptions of your reference viruses to `GRAViTy`. `GRAViTy` first reads the table and extract viruses' taxonomic assignments and accession numbers from this table. This step is done by `ReadGenomeDescTable`. There is no fixed format for the table (see Basic usage, and Option descriptions for more information), but we recommend you using viruses whose taxonomic assignments are officially recognised by the ICTV as your reference viruses. The list of such viruses can be found in the Virus Metadata Resource (VMR) file, provided by the ICTV from https://talk.ictvonline.org/taxonomy/vmr/.

The next step is to extract protein sequences from the virus genomes, cluster the sequences based on all-versus-all BLASTp bit scores by using Markov Clustering algorithm (3), align protein sequences within each of the clusters by using MUSCLE (4), and turn them into PPHMMs by using HMMER (http://hmmer.org/). If a (concatenated) GenBank file of reference virus genomes are not present in your computer, `GRAViTy` will download one for you from the NCBI database by using their accession numbers. This is done by `PPHMMDBConstruction`.

Subsequently, `GRAViTy` scans reference viruses against the PPHMM database to determine what genes they have, and where the genes are. The presence and absence of genes are not recorded in binaries but weighted by the HMM scores. We called each of these records a “PPHMM signature”. The data of the gene locations, or “PPHMM location signatures”, are subsequently used to build genomic organisation models (GOMs) for each reference virus taxonomic group. A GOM is simply a matrix with each row being a PPHMM location signature. PPHMM location signatures of reference viruses are then scanned against the GOM database to estimate the degrees of their genomic organisation similarity to various taxonomic groups. We call these “GOM signatures”. This is done by `RefVirusAnnotator`.

GRAViTyDendrogramAndHeatmapConstruction (optional) calculates similarities between each pair of viruses. A similarity between two viruses are measured by using the CGJ similarity index. Under the default settings, two generalised Jaccard scores are computed for each pair of viruses, one for their PPHMM signatures, and the other one for their GOM signatures. Their CGJ similarity is simply a geometric mean of the two scores, which ranges in value between 0 (no detectable similarity) and 1 (sequence identity). The degree of dissimilarity between the two viruses is 1 – CGJ. A heat map is then constructed from a pair-wise similarity matrix for the purpose of visualisation. By default, a UPGMA dendrogram (more specially, a phenogram) is also constructed to depict the degree of overall similarity among viruses. This function can also estimate the distance cutoff that best separates the (reference) taxonomic groupings overall, and report virus groupings as suggested by the estimated cutoff. Theil's uncertainty correlation for the reference taxonomic grouping given the predicted grouping, and vice versa, are reported. Symmetrical Theil's uncertainty correlation between the reference and predicted taxonomic grouping are also reported. These statistics can be used to evaluate the consistency between the reference virus groupings and the groupings suggested by GRAViTy. The robustness of the results (dendrogram, and virus grouping) can be evaluated by using the bootstrap analysis. Furthermore, note that GRAViTy also offers a variety of options for the calculation of pair-wise similarity and phenogram construction. (see GRAViTy_Pipeline_I --help for more information.)

In addition, GRAViTy_Pipeline_I has an option to calculate mutual information between taxonomic grouping scheme(s) and PPHMM scores (by MutualInformationCalculator) to determine which PPHMMs are highly (or weakly) correlated with the taxonomic grouping scheme(s). The results can be useful for determining what genes are predictive of virus taxonomy.

Basic usage

```
GRAViTy_Pipeline_I \
--GenomeDescTableFile "/PATH/TO/virus_description_table" \
--ShelveDir "/PATH/TO/OUTPUT_DIR" \
--Database "DATABASE" \
--Database_Header "DATABASE_HEADER" \
--TaxoGrouping_Header "TaxoGrouping_Header" \
--GenomeSeqFile "/PATH/TO/SEQ" \
--N_Bootstrap "INT"
```

Option descriptions

--GenomeDescTableFile = Path to your virus description table. It should be a tab delimited file (.txt), with headers. We recommend using the VMR file by the ICTV as a template. An excel version of VMR can be downloaded from https://talk.ictvonline.org/taxonomy/vmr/. The file should contain at least all of the following columns: "Baltimore Group", "Order", "Family", "Subfamily", "Genus", "Virus name (s)", "Virus GENBANK accession", "Virus sequence complete", and "Genetic code table". See "/Test/Data/Ref/VMR_Test_Ref.txt" for an example.

--ShelveDir = Path to a directory that stores all GRAViTy outputs. This is where the PPHMM and GOM databases are stored, together with other outputs.

--Database = GRAViTy will analyse only those that are labelled with DATABASE in the database column in the virus description table. The database column can be specified by using the "--Database_Header" option. If 'none', all entries are analysed. [default: none]

--Database_Header = The header of the database column. Cannot be none if DATABASE is specified. [default: none]

--TaxoGrouping_Header = The header of Taxonomic grouping column. Since GRAViTy mainly focuses on the family taxonomic assignment, the default value is "Family".

--TaxoGroupingFile = It is possible that the user might want to associate different viruses with different taxonomic assignment levels – family assignments for some, but subfamily or genus assignments for others, for example. To accommodate this, the user can either add a taxonomic grouping column in the virus description table, and use --TaxoGrouping_Header option to specify the column (see --TaxoGrouping_Header). Alternatively, the user can provide a file (with no header) that contains a single column of taxonomic groupings for all viruses in the order that appears in the description table. The user can specify the path to the file using this option. If this option is used, it will override the one specified by --TaxoGrouping_Header. [default: none]

--GenomeSeqFile = Path to the genome sequence file in the GenBank format (*.gb). If the file doesn't exist, GRAViTy will download one for you from the NCBI database using the accession numbers specified in the "Virus GENBANK accession" column in the description table.

--N_Bootstrap = "INT" is the number of bootstrap resampling [default: 10].

For more options, use GRAViTy_Pipeline_I --help.

Example

```
GRAViTy_Pipeline_I \
--GenomeDescTableFile "/PATH/TO/virus_description_table.Ref.txt" \
--ShelveDir "/PATH/TO/GRAViTyAnalyses/RefViruses/VI" \
--Database "VI" \
--Database_Header "Baltimore Group" \
--TaxoGrouping_Header "Taxonomic grouping" \
--GenomeSeqFile "/PATH/TO/GenomeSeqs.VI.gb" \
--N_Bootstrap 10
```

This command analyses reference viruses, whose descriptions are in "/PATH/TO/virus_description_table.Ref.txt". GRAViTy will only perform analysis on viruses labelled "VI" in the "Baltimore Group" column in the virus description table. The assigned taxonomic grouping is provided in the "Taxonomic grouping" column. The associated GenBank file is automatically downloaded by GRAViTy, if not present in the computer, stored at "/PATH/TO/GenomeSeqs.VI.gb". Bootstrapping analysis is to be performed with N = 10. The results will be stored at "/PATH/TO/GRAViTyAnalyses/RefViruses/VI". **Figure 2** shows how (some of) the results might look like.

Output descriptions

Outputs are organised into three directories.

- BLAST directory contains files generated during the all-versus-all BLASTp analyses and protein multiple sequence alignments.
- HMMER directory contains the PPHMM database.
- Shelves directory contains several key outputs.
    - "*.shelve" are files that keep python objects generated by GRAViTy, so don't worry about them.
    - PPHMMandGOMsignatures.txt contains the PPHMM and GOM signatures.

- `HeatmapWithDendrogram.*.pdf` is the heatmap depicted together with the dendrogram generated by `GRAViTy`.
- `Dendrogram.*.nwk` is the dendrogram generated by `GRAViTy` in the newick format, estimated based on complete pairwise CGJ distances.
- `DendrogramDist.*.nwk` contains the distribution of the bootstrapped resampled dendrograms.
- `BootstrappedDendrogram.*.nwk` is the dendrogram but with bootstrap clade support values.
- `VirusGrouping.*.txt` provide virus groupings that are based on the CGJ distance cutoff that best separates the reference taxonomic groupings overall. Various Theil's uncertainty correlations are reported. These statistics can be used to evaluate the similarity between the reference virus groupings and the groupings suggested by `GRAViTy`.
- `MutualInformationScore` directory contains mutual information scores between (various schemes of) taxonomic groupings and values of PPHMM scores to determine which PPHMMs are highly (or weakly) correlated with the virus taxonomic scheme(s). The default grouping scheme (the Overall scheme) is the one as specified in the `Taxonomic grouping` column in the description table. If you want to examine other schemes, see `--VirusGroupingSchemesFile` option using `--help`.

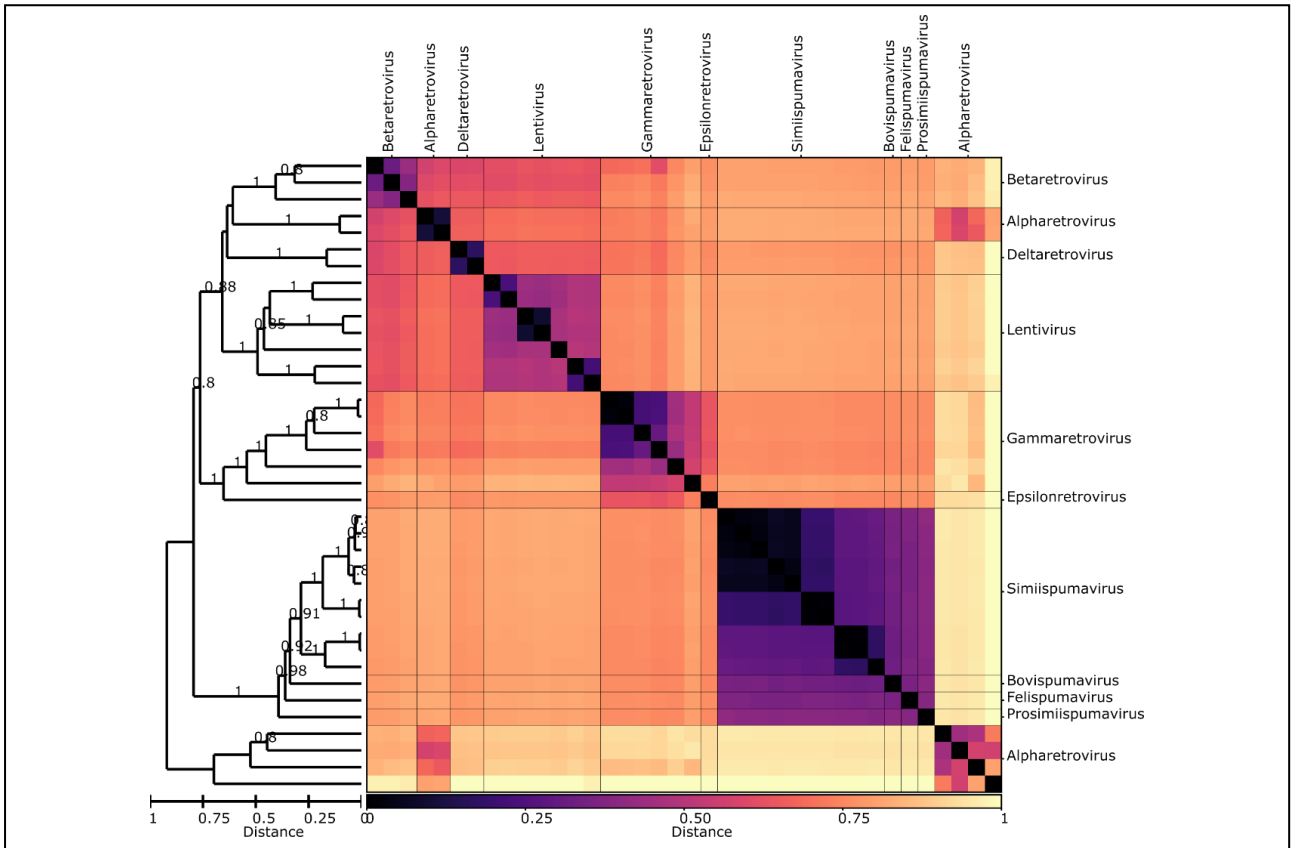

Figure 2. Heatmap of CGJ distances between virus taxonomic groups depicted together with a dendrogram.

Pairwise CGJ distances between each sequence pairs of viruses are plotted on the heatmap as colour coded points (see scale at the bottom of the figure). The solid lines in the heatmap indicate boundaries between each virus taxonomic group, and the data is organised such that groups with high similarities are closer to one another. An UPGMA dendrogram is shown on the left, constructed from the pairwise distance matric. The scale bar for the CGJ distance is shown at the bottom. Bootstrap clade support values are shown on branches.

GRAViTy_Pipeline_II

Description

```
GRAViTy_Pipeline_II
      |_ReadGenomeDescTable
      |_PPHMMDBConstruction
      |_UcfVirusAnnotator
      |_VirusClassificationAndEvaluation
```

The main purpose of GRAViTy_Pipeline_II is to identify and classify your viruses of interest.

GRAViTy first reads the description table of your viruses to extract their sequence identifiers by using ReadGenomeDescTable. In this case, the taxonomic assignments for the viruses may be blank. The file should contain at least all of the following columns: "Baltimore Group", "Order", "Family", "Subfamily", "Genus", "Virus name (s)", "Virus GENBANK accession", "Virus sequence complete", and "Genetic code table". The sequence identifiers should be in the "Virus GENBANK accession" column. See "/Test/Data/Ucf/VMR_Test_Ucf.txt" for an example.

GRAViTy then annotates your viruses by using the reference PPHMM and GOM databases, producing their PPHMM signatures, PPHMM location signatures, and GOM signatures. If their sequence files are not present in the computer, GRAViTy will attempt to download one for you from the NCBI database by using their "accession numbers" in the "Virus GENBANK accession" column of your virus description table. This is done by PPHMMDBConstruction. If the sequence identifiers are not GENBANK accession numbers and the sequence file is not provided by you, this will result in an error.

As an option, you may tell GRAViTy to construct a PPHMM database from your viruses, and this will also be performed by PPHMMDBConstruction. If so, this database will be used to improve both reference and unclassified virus annotations (i.e. the signatures).

Finally, GRAViTy will propose taxonomic candidates for your viruses. This is done by VirusClassificationAndEvaluation function. In summary, for each virus, GRAViTy proposes a candidate taxonomic group by identifying the best match reference virus(es). To validate the candidate assignment, GRAViTy employs a two-step evaluation protocol.

In the first step, GRAViTy checks whether or not the unclassified virus is 'similar enough' to the proposed candidate group, of which the CGJ similarity threshold is group specific. The cut-off is the CGJ score that best separates the distributions of inter- and intra-group CGJ similarity scores. If the observed CGJ similarity is less than the threshold, the candidate assignment is rejected, and the sample is relabelled as "Unclassified"; otherwise, the second step of the evaluation will be employed to further evaluate the candidate assignment.

In the second step, a phenogram containing both reference and unclassified virus is used, and the evaluator will look at its neighbourhood. The taxonomic proposal will be accepted if any of the following conditions are met:

i) the sister clade is composed entirely of the members of the proposed candidate taxonomic group
ii) the immediate out group is composed entirely of the members of the proposed candidate taxonomic group

iii) one of the two basal branches of its sister clade leading to a clade that is composed entirely of the members of the proposed candidate taxonomic group

To best estimate the placement of viruses, if multiple viruses are to be analysed at the same time, a phenogram containing all of your viruses will be used.

Furthermore, since there can be multiple reference databases being used at the same time (see Basic usage), there are possibilities that a virus might be assigned to multiple taxonomic groups belonging to different databases. In such cases, the finalised taxonomic assignment is the one associated with the highest CGJ similarity score. `GRAViTy` can also perform bootstrapping analysis to evaluate the uncertainty of the proposed taxonomic group. In addition, `GRAViTy` can produce a heat map to help visualising the pair-wise (dis)similarity among reference and unclassified viruses.

Basic usage

```
GRAViTy_Pipeline_II \
--GenomeDescTableFile_UcfVirus "/PATH/TO/virus_description_table" \
--ShelveDir_UcfVirus "/PATH/TO/OUTPUT_DIR" \
--ShelveDirs_RefVirus "/PATH/TO/REF_DIR_I, /PATH/TO/REF_DIR_II, …" \
--GenomeSeqFile_UcfVirus "/PATH/TO/SEQ" \
--UseUcfVirusPPHMMs "BOOLEAN" \
--GenomeSeqFiles_RefVirus "/PATH/TO/REF_SEQ_I, /PATH/TO/REF_SEQ_II,
…" \
--N_Bootstrap "INT"
```

Option descriptions

`--GenomeDescTableFile_UcfVirus` = Path to the description table of your viruses. It should be a tab delimited file (`.txt`), with headers. The file should contain at least all of the following columns: : "`Baltimore Group`", "`Order`", "`Family`", "`Subfamily`", "`Genus`", "`Virus name (s)`", "`Virus GENBANK accession`", "`Virus sequence complete`", and "`Genetic code table`". See "`/Test/Data/Ucf/VMR_Test_Ucf.txt`" for an example.

`--ShelveDir_UcfVirus` = Path to a directory that stores all `GRAViTy` outputs.

`--ShelveDirs_RefVirus` = Path(s) to the shelve director(y/ies) of reference virus(es).

`--GenomeSeqFile_UcfVirus` = Path to the genome sequences of your viruses in the GenBank format (`*.gb`). Their sequence identifiers should match those in the "Virus GENBANK accession" column in the description table.

`--UseUcfVirusPPHMMs` = Annotate reference and unclassified viruses using the PPHMM database derived from unclassified viruses if True. [default: True]

`--GenomeSeqFiles_RefVirus` = Path(s) to the genome sequence GenBank file(s) of reference viruses. This cannot be 'None' if `--UseUcfVirusPPHMMs = True`.

`--N_Bootstrap` = "`INT`" is the number of bootstrap resampling [default: 10].

For more options, use `GRAViTy_Pipeline_I --help`.

Example

```
GRAViTy_Pipeline_II \
```

```
--GenomeDescTableFile_UcfVirus
"/PATH/TO/virus_description_table.Ucf.txt" \
--ShelveDir_UcfVirus "/PATH/TO/GRAViTyAnalyses/UcfViruses" \
--ShelveDirs_RefVirus "/PATH/TO/GRAViTyAnalyses/RefViruses/VI,
/PATH/TO/GRAViTyAnalyses/RefViruses/VII" \
--GenomeSeqFile_UcfVirus "/PATH/TO/GenomeSeqs.Ucf.gb" \
--UseUcfVirusPPHMMs True \
--GenomeSeqFiles_RefVirus "/PATH/TO/GenomeSeqs.VI.gb,
/PATH/TO/GenomeSeqs.VII.gb" \
--N_Bootstrap 10
```

This command will analyse your (unclassified) viruses, whose descriptions are in "`/PATH/TO/virus_description_table.Ucf.txt`", and keeps the results at "`/PATH/TO/GRAViTyAnalyses/UcfViruses`". `GRAViTy` will find the genomes of your viruses at "`/PATH/TO/GenomeSeqs.Ucf.gb`". Two reference `GRAViTy` databases are used, one at "`/PATH/TO/GRAViTyAnalyses/RefViruses/VI`" and the other at "`/PATH/TO/GRAViTyAnalyses/RefViruses/VII`". Since `UseUcfVirusPPHMMs` is `True`, `GRAViTy` will update the virus annotations (i.e. the PPHMM and GOM signatures) of both the reference and your viruses by using the PPHMM database derived from your viruses. The genomes of reference viruses can be found at "`/PATH/TO/GenomeSeqs.VI.gb`", and at "`/PATH/TO/GenomeSeqs.VII.gb`". Bootstrapping analysis is to be performed with N = 10. **Figure 3** shows how (some of) the results might look like.

Output descriptions

Outputs are organised into three directories.

- `BLAST` directory contains files generated during the all-versus-all BLASTp analyses and protein multiple sequence alignments. This folder will be generated only when `UseUcfVirusPPHMMs` is `True`.
- `HMMER` directory contains the PPHMM database. This folder will be generated only when `UseUcfVirusPPHMMs` is `True`.
- `Shelves` directory contains several key outputs.
    - "`*.shelve`" are files that keep python objects generated by `GRAViTy`, so don't worry about them.
    - `HeatmapWithDendrogram.*.pdf` is the heatmap depicted together with the dendrogram generated by `GRAViTy`. If multiple reference databases are used, multiple `HeatmapWithDendrogram.*.pdf` files will be generated.
    - `Dendrogram.*.nwk` is the dendrogram generated by `GRAViTy` in the newick format, estimated based on complete pairwise CGJ distances. If multiple reference databases are used, multiple `Dendrogram.*.nwk` files will be generated.
    - `DendrogramDist.*.nwk` contains the distribution of the bootstrapped resampled dendrograms. If multiple reference databases are used, multiple `DendrogramDist.*.nwk` files will be generated.
    - `BootstrappedDendrogram.*.nwk` is the dendrogram but with bootstrap clade support values. If multiple reference databases are used, multiple `BootstrappedDendrogram.*.nwk` files will be generated.
    - `VirusGrouping.*.txt` provide virus groupings that are based on the CGJ distance cutoff that best separates the reference taxonomic groupings overall. Various Theil's uncertainty correlations are reported. These statistics can be used to evaluate the similarity between the reference virus groupings and the

groupings suggested by GRAViTy. If multiple reference databases are used, multiple VirusGrouping.*.nwk files will be generated.

- ClassificationResults.txt provides the results of virus identification and classification. This file contains lots of information. Here are brief explanations.
  - Candidate class (class of the best match reference virus): This column shows candidate taxonomic assignment, transferred from the most similar reference virus.
  - Similarity score: This column shows the CGJ similarity score to the best match reference virus. The similarity score cut-offs for each of the reference virus taxonomic groups are shown below the table.
  - Support from dendrogram: This column summaries how each of your viruses is related to the proposed taxonomic group.
    - NA: the sequence is not similar enough to any of the reference sequences
    - 1: the sequence is embedded within the clade of the candidate taxonomic group
    - 2: the sequence has a sister relationship with the candidate taxonomic group and they are similar enough
    - 3: the sequence is 'sandwished' between 2 branches of the candidate taxonomic group
    - 4: the sequence has a paraphyletic relationship with the candidate taxonomic group (just inside)
    - 5: the sequence has a paraphyletic relationship with the candidate taxonomic group (just outside)
    - 6: the candidate taxonomic group is not supported by the dendrogram
  - Evaluated taxonomic assignment: This column tells you if the candidate taxonomic assignment passes the evaluation criteria or not. If not, it will be labelled “Unclassified”.
  - Best taxonomic assignment: This column tells you the best taxonomic assignment. This is particularly relevant when you use multiple reference GRAViTy databases to analyse your viruses, since there are possibilities that a virus might be assigned to multiple taxonomic groups belonging to different databases. In such cases, the finalised taxonomic assignment is the one associated with the highest CGJ similarity score. In the case of “Unclassified virus”, this column tells you if your virus exhibits similarity to any viruses at all or not. If so, GRAViTy will attempt to tell which database it might belong to even though it cannot be assigned to any specific virus group.
  - Provisional virus taxonomy: This column tells you the final virus taxonomic groupings. For viruses that can be identified, the provisional virus taxonomic assignment will be the same as the best taxonomic assignment. For unclassified viruses, GRAViTy will attempt to group them together based on the CGJ distance cutoff that best separates the reference taxonomic groupings overall.

  - Note that if multiple reference databases are used, there will be results, one from each reference database. If bootstrapping technique is used to evaluate the uncertainties of the assignments, the distributions of the scores will also be shown.

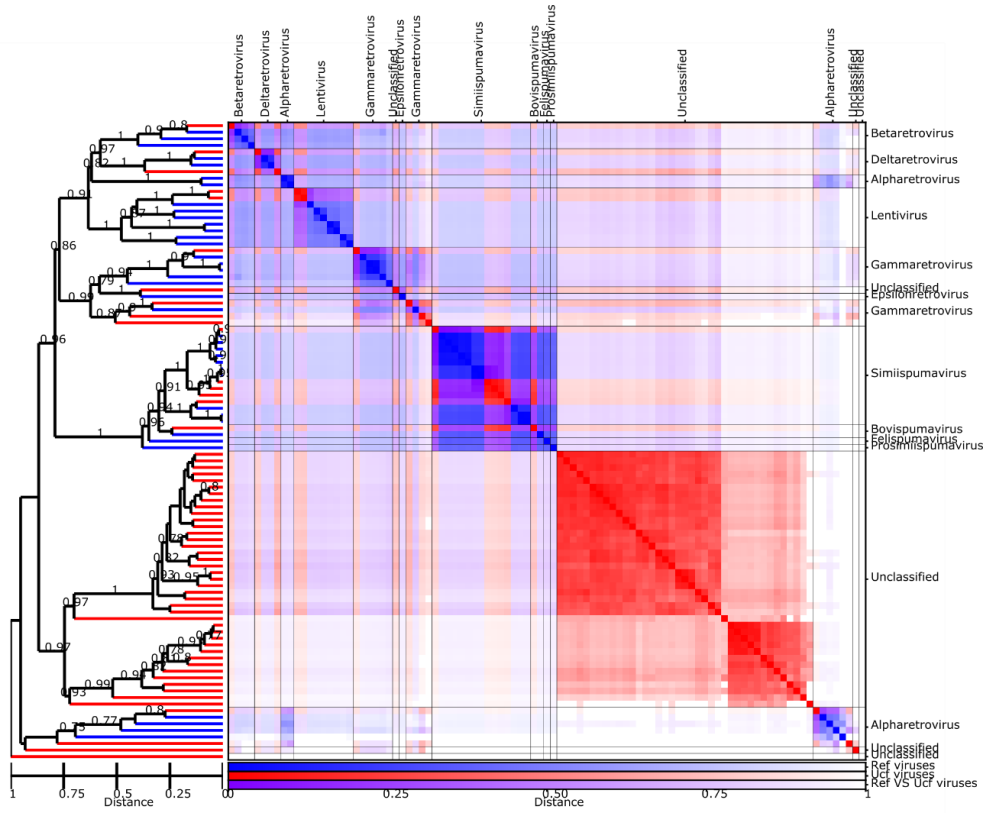

Figure 3. Heatmap of CGJ distances between reference and test viruses depicted together with a dendrogram.

Pairwise CGJ distances between each sequence pairs of viruses are plotted on the heatmap as colour coded points (blue: between pairs of reference viruses; red: between pairs of test viruses; and purple: between pairs of reference and test viruses, see scale at the bottom of the figure). The solid lines in the heatmap indicate boundaries between each virus taxonomic group, and the data is organised such that groups with high similarities are closer to one another. An UPGMA dendrogram is shown on the left, constructed from the pairwise distance matric. The scale bar for the CGJ distance is shown at the bottom. Terminal branches leading to reference viruses are in blue, and those leading to test viruses are in red. Bootstrap clade support values are shown on branches.

References

1. **Aiewsakun P**, **Simmonds P**. 2018. The genomic underpinnings of eukaryotic virus taxonomy: Creating a sequence-based framework for family-level virus classification. Microbiome **6**:38.
2. **Aiewsakun P**, **Adriaenssens EM**, **Lavigne R**, **Kropinski AM**, **Simmonds P**. 2018. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: Steps towards a unified taxonomy. J Gen Virol **99**:1331–1343.
3. **Van Dongen S**. 2008. Graph clustering via a discrete uncoupling process. SIAM J Matrix Anal Appl **30**:121–141.
4. **Edgar RC**. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113.