Phylogenetics and virus classification

Pakorn Aiewsakun
SCMI 512 Sequence analysis
20/03/19

Different schools of taxonomy
Schools of taxonomy

Cladistics

- Organisms are grouped based on their most recent common ancestor

Phenetics/numerical taxonomy

- Organisms are grouped based on overall (dis)similarity

Evolutionary systematics

- Organisms are grouped based on a combination of shared descent, serial descent, and the degree of (dis)similarity
- “Synthetics” school of taxonomy

Schools of taxonomy

Cladistics

- Their shared descent is inferred based on (homologous) synapomorphies (which sometimes can be controversial and subjective)

Phenetics/numerical taxonomy

- Overall (dis)similarity is calculated using all available characters / data (often at the dispense of homology)

Evolutionary systematics

- All characters are used (synthesis of all information), but, at the same time, the method tries to account for homology and homoplasy
Schools of taxonomy

**Cladistics**
- Groupings are shown using a cladogram
- Only recognises “clades” as true groupings
- Paraphyletic groups are “grades”

**Phenetics/numerical taxonomy**
- Groupings are shown using a phenogram
- Ignores the concept of evolution entirely
- Only generates an (empirical) identification scheme

**Evolutionary systematics**
- Groupings are shown using a Besseyan cactus/commagram
- Recognise paraphyletic groupings (i.e. accommodating series descent) as official (higher taxa = convenience classes)

---

Besseyan cactus

[Diagram of Besseyan cactus showing groupings and species]
Morphology-based virus classification

Family level classification is typically morphological based

- Baltimore classification system is based on genomic material and replication strategy

Temin, 1985, Mol. Biol. Evol. – *Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts*

Baltimore D, 1971, Bacteriol. Rev. – *Expression of animal virus genomes*
Family level classification is typically morphological based

- **Reverse transcribing viruses:**
  - **Baltimore classification group VI:** members have ssRNA genomes and frequently integrate into the host genomes as part of their replication cycles
    - *Belpaoviridae* (Bel/Pao retrotransposons)
    - *Metaviridae* (Ty3/Gypsy retrotransposons)
    - *Pseudoviridae* (Ty1/Copia retrotransposons)
    - *Retroviridae*
  - **Baltimore classification group VII:** members have circular dsDNA genomes and do not actively integrate into host chromosomes
    - *Caulimoviridae*
    - *Hepadnaviridae*

The 3 families of tailed bacteriophages (*Caudovirales*)

**Siphoviridae**
- *Ssbetavirus*

**Myoviridae**
- *T4virus*

**Podoviridae**
- *Luz24virus*

ICTV website: https://talk.ictvonline.org/ictv-reports/ictv_9th_report/dsdna_viruses-2011/73/siphoviridae-figures
ICTV website: https://talk.ictvonline.org/ictv-reports/ictv_9th_report/dsdna_viruses-2011/70/myoviridae-figures
Family level classification is typically morphological based

- **Mononegavirales; Rhabdoviridae**
- Its members have non-segmented, linear, (−)ssRNA genomes, and helical nucleocapsids with bullet-shaped and bacilliform geometries (*rhabdos* = rod), and are enveloped.

https://talk.ictvonline.org/ictv-reports/ictv_9th_report/negative-sense-rna-viruses-2011/w/negrna_viruses/201/rhabdoviridae

Family level classification is typically morphological based

- **Mononegavirales; Rhabdoviridae; Novirhabdovirus**
- “Almost every aspect of their virion structure, genome organisation, expression strategy, protein sequence alignment and protein structure where available tell us they are rhabdoviruses”
  - Peter J Walker

Electron micrograph showing budding of the fish novirhabdovirus (Graznow et al., 1997, Journal of Fish Diseases)
Family level classification is typically morphological based

- **Mononegavirales; Rhabdoviridae; Novirhabdovirus**

- Characteristic rhabdovirus bullet-shaped virion structure

- Their M proteins have the late budding domain (PPPH/Y)

- There is a set of 12 conserved cysteine residues that tie the G protein structure together, just like other rhabdoviruses.

Electron micrograph showing budding of the fish novirhabdovirus (Graznow et al., 1997, Journal of Fish Diseases)

Family level classification is typically morphological based

- This is all good, but... historical criteria might not be effective for classification of newly discovered viruses

![Bpp1virus](image1)

![Epsilon15virus](image2)

![Phi29virus](image3)

![Luz24virus](image4)

![Phieco32virus](image5)

![T7virus](image6)

![N4virus](image7)

![P22virus](image8)

Family level classification is typically morphological based

• This is all good, but... with metagenomically-derived viruses, we have no morphological information

>NC_033289.1 Shahe yuevirus-like virus 1 strain SHWC0209c11789 RNA-dependent RNA polymerase gene, complete cds

CAAAATTTCAAGCATAATATTTAAAAATAAACTTTAACACCTTTAAATTTTAAATATTTAAAAGAGAATTAAATATAACAGAAATGTCAGATAGACTTGTTACTACTGATTATACTATGCATTTATTAGCTAGCTCAATATGGAGACCCTTTTCTCTTTAGAAGAAATCTACTCAAGCATATATAGAAAGAGATACATGCTTCCAAGG

Family level classification is typically morphological based

• This is all good, but... with the explosion of novel viruses...

![Morphology-Based Classification](image_url)
Gene-based classification of viruses

RDRP phylogeny of RNA viruses

Rooted phylogeny of palm domain-containing RNA dependent RNA polymerases and reverse transcriptases of +ssRNA, -ssRNA, and dsRNA viruses

Wolf et al., 2018, mBio
RDRP phylogeny of RNA viruses

Ortervirales: New virus order unifying five families of reverse-transcribing viruses

Krupovic et al., 2018, J Virol
**Ortervirales**: New virus order unifying five families of reverse-transcribing viruses

Maximum-likelihood phylogeny of viral reverse transcriptases. The tree includes sequences of 290 viruses belonging to all ICTV-recognized genera of reverse-transcribing viruses

Krupovic et al., 2018, J Virol

“This phylogenetic position suggests that among pararetroviruses, encapsidation of a DNA genome is a homoplasious trait and, therefore, is not a reliable criterion for classification.”

Krupovic et al., 2018, J Virol

The hepadnaviruses, which typically (i) branch out at the base of the viral RT clade (Fig. 1), (ii) possess a unique capsid protein, and (iii) employ a distinct replication mechanism, appear to be more distantly related to all these virus families

In recognition of these relationships, the ICTV has recently regrouped the families Belpaoviridae, Caulimoviridae, Metaviridae, Pseudoviridae, and Retroviridae into an order, Ortervirales (Orter, an inversion of “retro,” which was derived from reverse transcription; -virales, suffix for an order)

Krupovic et al., 2018, J Virol
**Caudovirales:** tailed bacteriophages

Phylogenetic relationships of Caudovirales infecting Pseudomonas species: **Major capsid proteins** and the neighbor-joining method were used to construct the phylogenetic tree.

Shen et al., 2016, Scientific Report

---

**Caudovirales:** tailed bacteriophages

Neighbor-joining trees of TerL (A) and portal proteins (B; φPLPE gp19 homologs). The eight dwarf φPLPE-like myoviruses are highlighted with red arrows. Values at the nodes are the results of 100 bootstrap replicates. The scale bar indicates 0.1 substitutions per site.

Comeau et al., 2012, PLOS ONE
Phylogeny of rhabdoviruses

Phylogenetic relationships of members of the family Rhabdoviridae, inferred using L protein sequences, under the ML framework.

ML phylogeny of the Rhabdoviridae. (A) shows the basal fish-infecting novirhabdoviruses, an unassigned group of arthropod associated viruses, the plant infecting cyto- and nucleo-rhabdoviruses, as well as the vertebrate specific lyssaviruses. (B) shows the dimarhabdovirus supergroup, which is predominantly composed of arthropod-vectored vertebrate viruses, along with the arthropod-specific sigma virus clade.

Longdon et al., 2015, Virus Evolution
Phylogeny of rhabdoviruses

BS phylogeny of the order Mononegavirales, based on a multiple alignment of the L proteins. Only informative blocks representing most conserved alignment regions were included in the analysis (428 positions in total).

Problems with gene-based classification

• Throws away lots of data, and the choice of genes are subjective, and subjected to how “alignable” they are

• Different genes might tell different stories

• Requires multiple sequence alignments

• Viruses don’t have universal genes
Phenetic approach to virus classification

PASC – PAirwise Sequence Comparison

• Calculates pairwise identities of virus sequences within families, using the BLAST-based alignment method

• Determines cut-off for viruses of the same strain, species, genus and subfamily from the distribution

• Assigns new virus to a taxonomic group based on pairwise identities

• The choice of thresholds is subjective and same cut off is applied for all virus taxonomic groups of the same level

Bao et al., 2014, Arch Virol
PASC – PAirwise Sequence Comparison

Bao et al., 2014, Arch Virol

PASC – PAirwise Sequence Comparison

Bao et al., 2014, Arch Virol
PASC – PAirwise Sequence Comparison

Bao et al., 2014, Arch Virol

PASC – PAirwise Sequence Comparison

Bao et al., 2014, Arch Virol
GRAViTy: Genome Relationships Applied to Virus Taxonomy

- Calculates pairwise identities for an arbitrary set of viruses, based on
  - Similarity in gene sets
  - Similarity in genomic organisation (gene order & orientation)

Aiewsakun & Simmonds, 2018, Microbiome

---

GRAViTy: Genome Relationships Applied to Virus Taxonomy

- Calculates pairwise identities for an arbitrary set of viruses, based on
  - Similarity in gene sets
  - Similarity in genomic organisation (gene order & orientation)

Krupovic et al., 2018, J Virol

<table>
<thead>
<tr>
<th>Virus</th>
<th>Gene model 1</th>
<th>Gene model 2</th>
<th>Gene model 3</th>
<th>...</th>
<th>GOM1</th>
<th>GOM2</th>
<th>GOM3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virus 1</td>
<td>1000</td>
<td>500</td>
<td>0</td>
<td></td>
<td>0.25</td>
<td>0.90</td>
<td>0.30</td>
<td>...</td>
</tr>
<tr>
<td>Virus 2</td>
<td>890</td>
<td>600</td>
<td>50</td>
<td></td>
<td>0.10</td>
<td>0.95</td>
<td>0.25</td>
<td>...</td>
</tr>
<tr>
<td>Virus 3</td>
<td>10</td>
<td>1500</td>
<td>700</td>
<td></td>
<td>0.40</td>
<td>0.12</td>
<td>0.20</td>
<td>...</td>
</tr>
<tr>
<td>Virus 4</td>
<td>0</td>
<td>1800</td>
<td>300</td>
<td></td>
<td>0.45</td>
<td>0.20</td>
<td>0.35</td>
<td>...</td>
</tr>
</tbody>
</table>

Aiewsakun & Simmonds, 2018, Microbiome
GRAViTy: Genome Relationships Applied to Virus Taxonomy

- Calculates pairwise identities for an arbitrary set of viruses, based on
  - Similarity in gene sets
  - Similarity in genomic organisation (gene order & orientation)

Aiewsakun & Simmonds, 2018, Microbiome

GRAViTy: Genome Relationships Applied to Virus Taxonomy

- Generates a phenogram using a clustering method (UPGMA)
- Examines virus groupings, and determine cut-off distances for each reference virus group

Aiewsakun & Simmonds, 2018, Microbiome
GRAViTy: Genome Relationships Applied to Virus Taxonomy

- To classify your favourite virus, GRAViTy:
  - Computes its pairwise distances against all reference viruses
  - Finds the ‘most similar’ one
  - Asks if it is ‘similar enough’
  - If so, finds out if your virus sits in the right place

Aiewsakun & Simmonds, 2018, Microbiome

GRAViTy: Genome Relationships Applied to Virus Taxonomy

- GRAViTy can also determine what genes are correlated with the current established virus taxonomy

Aiewsakun & Simmonds, 2018, Microbiome
GRAViTy: Genome Relationships Applied to Virus Taxonomy

- Challenges
  - Horizontal gene transfers
  - Fuzzy taxonomic boundary
  - No one threshold to rule them all
  - Novel viruses, and naming
  - Convergence evolution

Summary

- Traditional virus classification framework cannot accommodate the explosion of viral genomic sequence data

- ICTV supports the shift from the traditional morphology-based virus classification framework to a genomic-based one

- Phylogenetics plays a crucial role in this development. Although a genome-based taxonomy might be different from those based on phenotypic features, this approach may be the only solution