Reducing haystacks to needles

Comparative Genomics based on viral clusters

Kevin Lamkiewicz

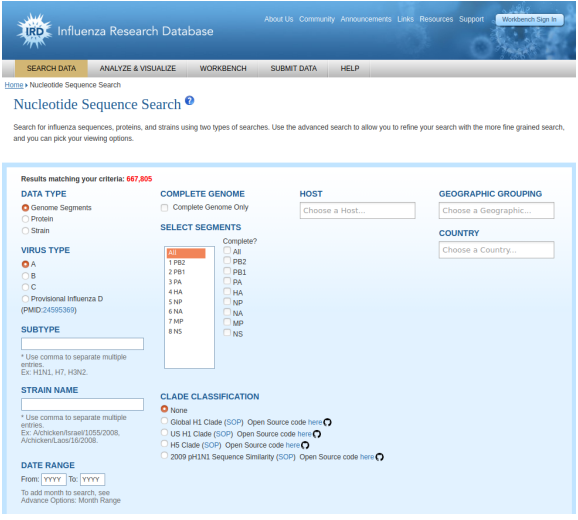
12.02.2020

RNA Bioinformatics and High-Throughput Analysis

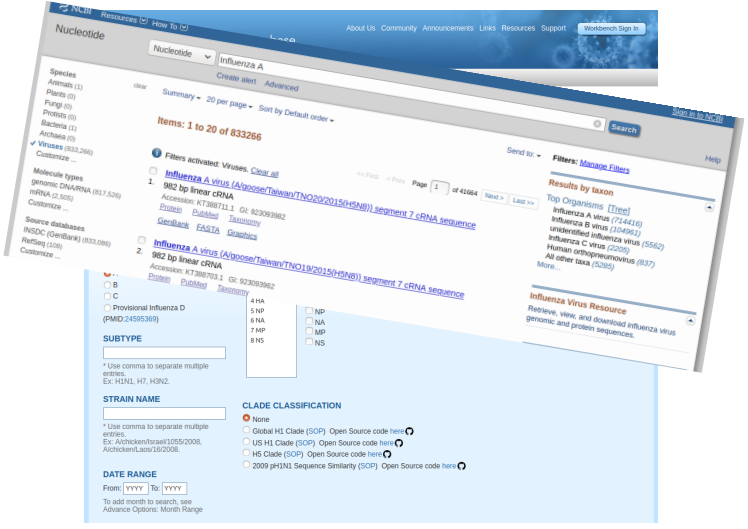
Friedrich Schiller University Jena

Why cluster in the first place?

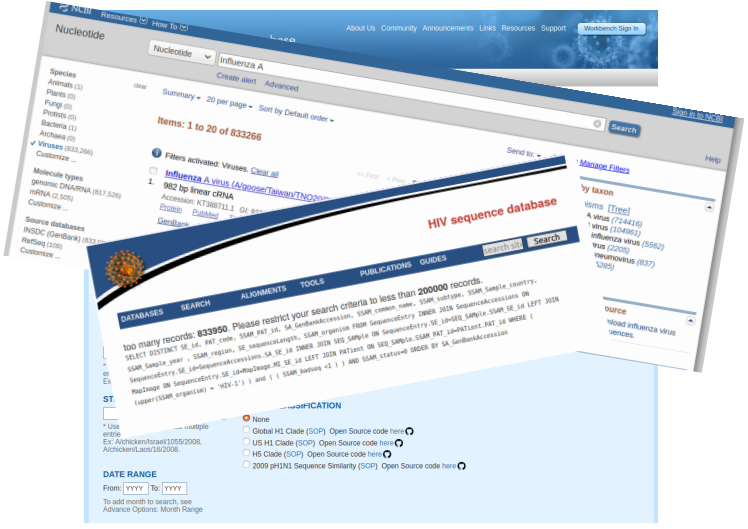
MASSIVE AMOUNTS OF DATA



MASSIVE AMOUNTS OF DATA



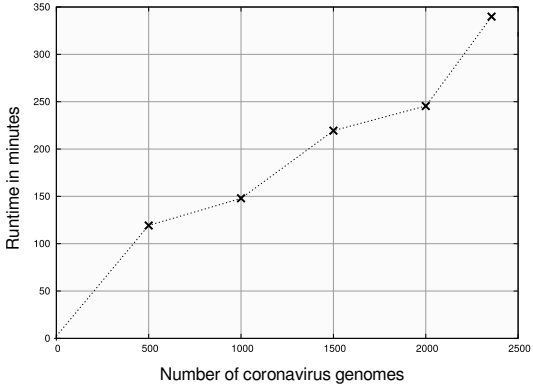
MASSIVE AMOUNTS OF DATA



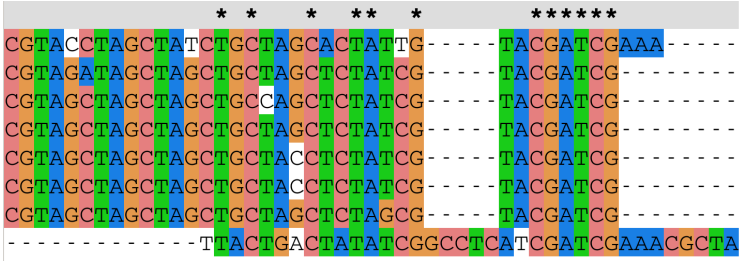
MASSIVE AMOUNTS OF DATA

SAVING TIME AND MEMORY

- `MAFFT` alignment with CoV
- just a fraction of all available genomes

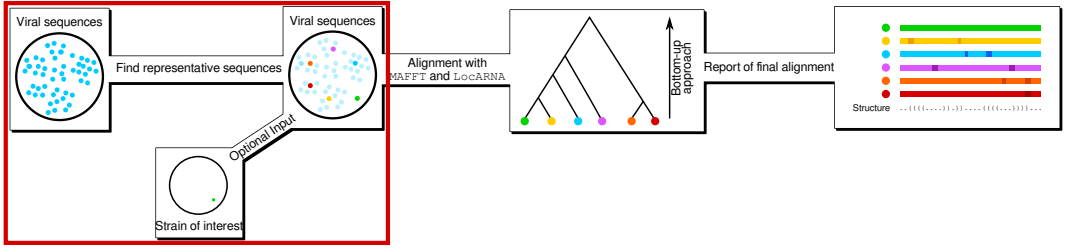


INTRODUCED BIAS...



We need something smarter

GENERAL WORKFLOW



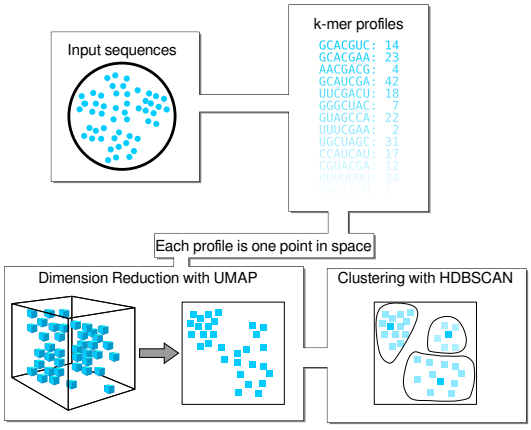
UMAP AND HDBSCAN

UMAP[1]

- connect data in space using simplices (based on fuzzy open cover)
- find low-dimensional representation with similar topological representation

HDBSCAN[2]

- minimum spanning tree on transformed distances
- convert tree to hierarchy of connected components, extract cluster from these



[1]: https://umap-learn.readthedocs.io/en/latest/index.html

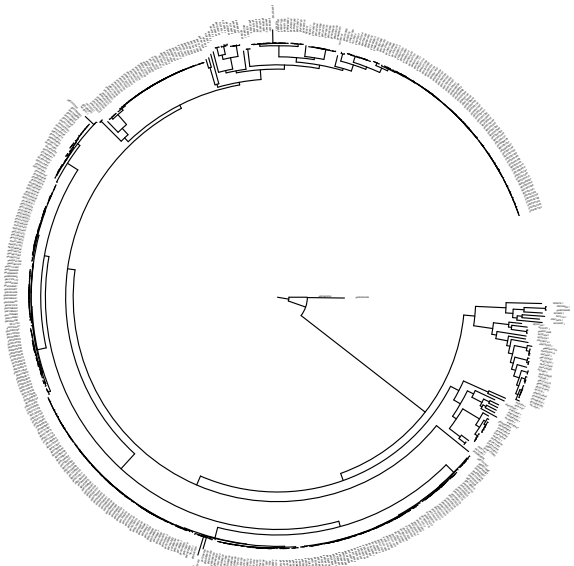
[2]: https://hdbscan.readthedocs.io/en/latest/index.html

First Results

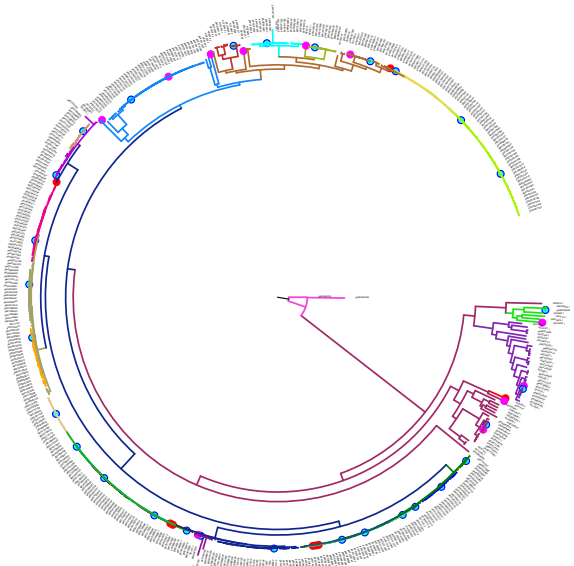
Viral Clade	Input Sequences	Cluster	Runtime[1] in sec.
Denguevirus	5,470	302	5,695
Ebolavirus	634	33	43
Filoviridae	728	39	56
Zikavirus	789	38	36
Alphacoronavirus	927	57	141
Betacoronavirus	1,146	67	223
Poxviridae	688	45	467
Herpesviridae	1,758	89	2,023

[1]: 8 cores, 3.6 GHz

My approach versus intuition



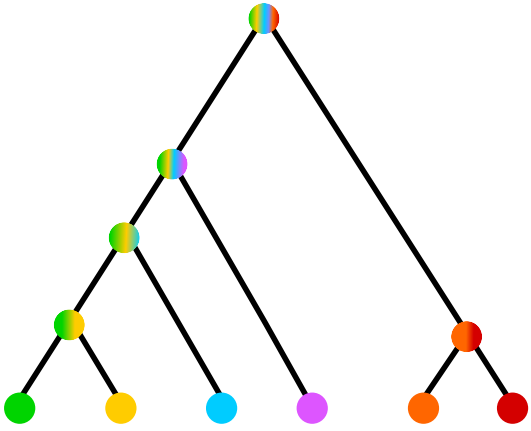
My approach versus intuition



Some examples for downstream analyses

BASICALLY ANYTHING

- Phylogeny

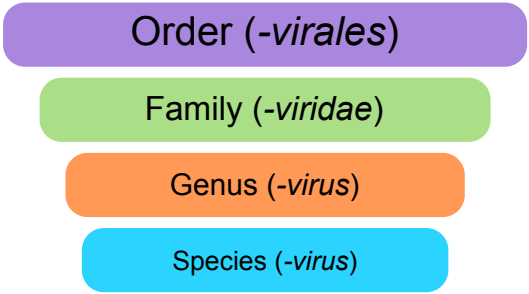


BASICALLY ANYTHING

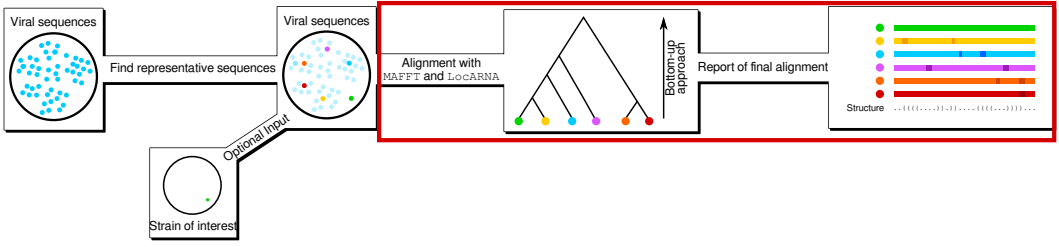
- Phylogeny
- Functional estimations of genes/protein

BASICALLY ANYTHING

- Phylogeny
- Functional estimations of genes/protein
- Classification



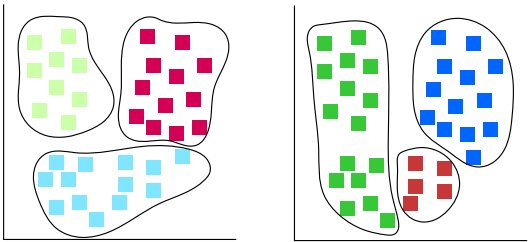
MY PROJECT: SECONDARY STRUCTURES



Let’s talk about the future

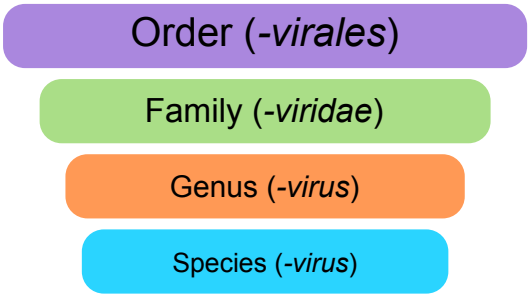
WHAT'S NEXT ON THE (HAY)STACK

- Comparison with other cluster pipelines



WHAT'S NEXT ON THE (HAY)STACK

- Comparison with other cluster pipelines
- Evaluation on more viral genera and families



WHAT'S NEXT ON THE (HAY)STACK

- Comparison with other cluster pipelines
- Evaluation on more viral genera and families
- Optimizing parameter for local and global structures

```
cluster_seq.py -l ? -s ?? -u ??? --help? <INPUT>
```

What's next on the (hay)stack

- Comparison with other cluster pipelines
- Evaluation on more viral genera and families
- Optimizing parameter for local and global structures
- conda and docker environments

What's next on the (hay)stack

- Comparison with other cluster pipelines
- Evaluation on more viral genera and families
- Optimizing parameter for local and global structures
- conda and docker environments

- get those nasty whole-genome structure alignments done...

Thank you for your attention!

Acknowledgements:


RNA Bioinformatics Group Jena

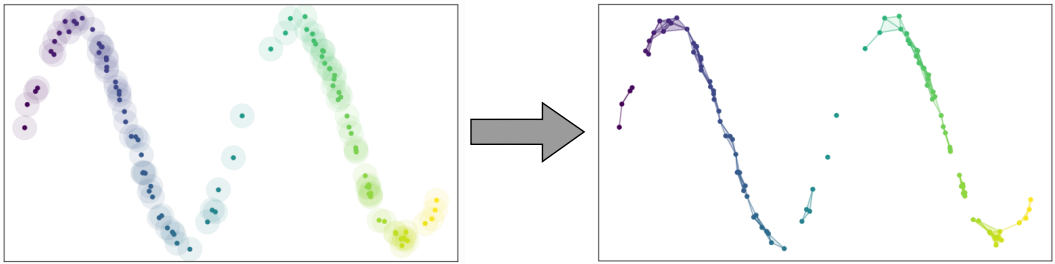
Federal Ministry of Education and Research

Carl Zeiss Stiftung

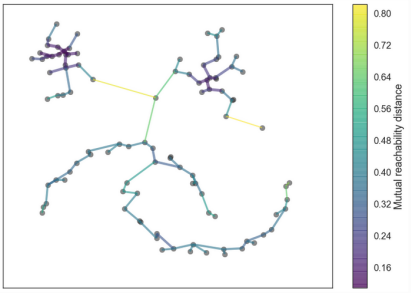
European VIRUS BIOINFORMATICS Center

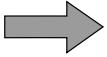
MichaelStifelCenterJena for Data-Driven and Simulation Science

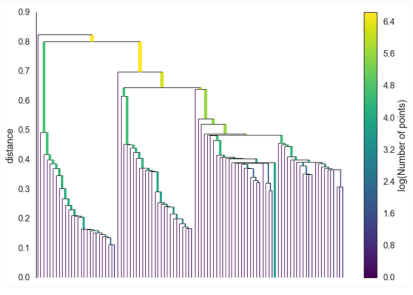
UMAP: Fuzzy open cover and simplices



HDBSCAN: Spanning Tree to Hierarchy







COSINE DISTANCE

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Ranges from -1 to 1, where -1 means exact opposite, 1 being exactly the same and 0 indicate orthogonality