

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



[About Us](#) [Community](#) [Announcements](#) [Links](#) [Resources](#) [Support](#) [Workbench Sign In](#)

SEARCH DATA
ANALYZE & VISUALIZE
WORKBENCH
SUBMIT DATA
HELP

[Home](#) » [Nucleotide Sequence Search](#)

Nucleotide Sequence Search ?

Search for influenza sequences, proteins, and strains using two types of searches. Use the advanced search to allow you to refine your search with the more fine grained search, and you can pick your viewing options.

Results matching your criteria: 667,805

DATA TYPE

Genome Segments

Protein

Strain

COMPLETE GENOME

Complete Genome Only

HOST

GEOGRAPHIC GROUPING

VIRUS TYPE

A

B

C

Provisional Influenza D
(PMID:24595369)

SELECT SEGMENTS

All

1 PB2

2 PB1

3 PA

4 HA

5 NP

6 NA

7 MP

8 NS

Complete?

All

PB2

PB1

PA

HA

NP

NA

MP

NS

COUNTRY

SUBTYPE

* Use comma to separate multiple entries.
Ex: H1N1, H7, H3N2.

STRAIN NAME

* Use comma to separate multiple entries.
Ex: A/chicken/Israel/1055/2008, A/chicken/Laos/16/2008.

DATE RANGE

From: To:

To add month to search, see Advance Options: Month Range

CLADE CLASSIFICATION

None

Global H1 Clade (SOP) Open Source code [here](#) ?

US H1 Clade (SOP) Open Source code [here](#) ?

H5 Clade (SOP) Open Source code [here](#) ?

2009 pH1N1 Sequence Similarity (SOP) Open Source code [here](#) ?

MASSIVE AMOUNTS OF DATA

NCBI Resources | How To | Log out

Nucleotide Influenza A

Summary - 20 per page - Sort by Default order

Items: 1 to 20 of 833266

Filters activated: Viruses [Clear all](#)

1. [Influenza A virus \(A/opose/Taiwan/TNO20/2015\(H5N8\)\) segment 7 cRNA sequence](#)
 982 bp linear cRNA
 Accession: KT388711.1 GI: 923093962
[Protein](#) [PubMed](#) [Taxonomy](#)
[GenBank](#) [FASTA](#) [Genpep](#)

2. [Influenza A virus \(A/opose/Taiwan/TNO19/2015\(H5N8\)\) segment 7 cRNA sequence](#)
 982 bp linear cRNA
 Accession: KT388703.1 GI: 923093962
[Protein](#) [PubMed](#) [Taxonomy](#)

Species: Animals (1), Plants (0), Fungi (0), Protists (0), Bacteria (1), Archaea (0), **Viruses (833,266)**, Customize...

Molecule types: genomic DNA/RNA (817,526), mRNA (2,505), Customize...

Source databases: INSDC (GenBank) (833,094), RefSeq (109), Customize...

Send to: [Manage Filters](#)

Results by taxon

Top Organisms [Tree](#)

- Influenza A virus (7,4416)
- Influenza B virus (10,4982)
- unclassified influenza virus (5562)
- Influenza C virus (2205)
- Human orthopneumovirus (837)
- All other taxa (5285)
- More...

Influenza Virus Resource
 Retrieve, view, and download influenza virus genomic and protein sequences.

B
 C
 Provisional influenza D (PMD:2459369)

SUBTYPE

* Use comma to separate multiple entries.
 Ex: H1N1, H7, H3N2.

STRAIN NAME

* Use comma to separate multiple entries.
 Ex: A/chicken/Israel/1055/2008, A/chicken/Laos/16/2008.

DATE RANGE
 From: To:
 To add month to search, see Advance Options: Month Range

4 HA NP
 5 NP NA
 6 NA MP
 7 MP NS
 8 NS

CLADE CLASSIFICATION

- None
- Global H1 Clade (SOP) Open Source code [here](#)
- US H1 Clade (SOP) Open Source code [here](#)
- H5 Clade (SOP) Open Source code [here](#)
- 2009 pH1N1 Sequence Similarity (SOP) Open Source code [here](#)

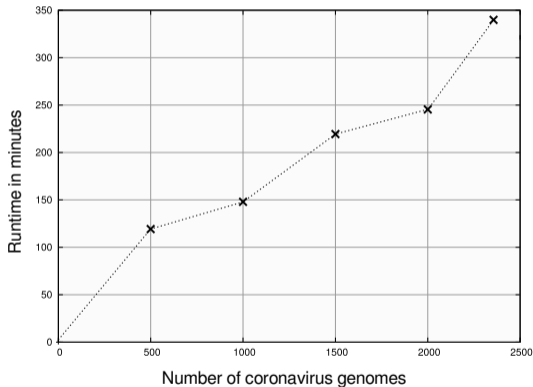
MASSIVE AMOUNTS OF DATA

The image illustrates the challenge of handling massive amounts of data in a biological database. It features three overlapping screenshots from the NCBI GenBank interface:

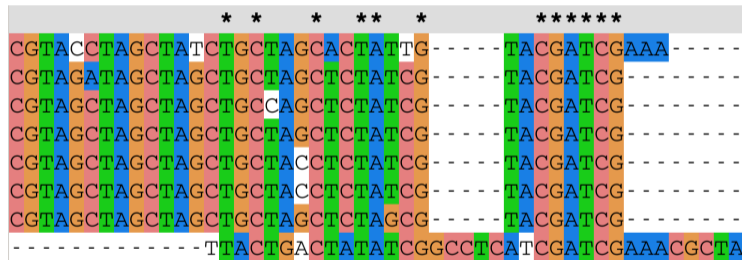
- Top Screenshot:** A search for "Influenza A" in the "Nucleotide" database. The results show "Items: 1 to 20 of 833266". A sidebar on the left lists various taxonomic groups, with "Viruses (833,266)" selected.
- Middle Screenshot:** A search for "HIV sequence database". The results indicate "too many records: 833950. Please restrict your search criteria to less than 200000 records." Below this, a complex SQL query is displayed, filtering for HIV-1 sequences. A "HIV sequence database" banner is overlaid on the search results.
- Bottom Screenshot:** A search filter for "HIV-1" sequences. It includes a "DATE RANGE" section with "From: YYYY To: YYYY" and a "CLASSIFICATION" section with radio buttons for "None", "Global H1 Clade (SOP)", "US H1 Clade (SOP)", "H5 Clade (SOP)", and "2009 pH1N1 Sequence Similarity (SOP)".

SAVING TIME AND MEMORY

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

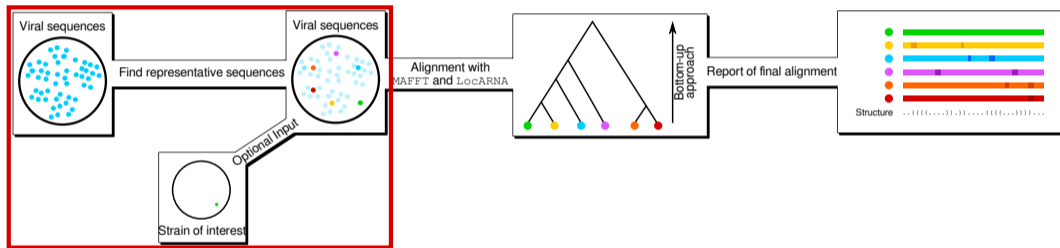


INTRODUCED BIAS...



We need something smarter

GENERAL WORKFLOW



UMAP AND HDBSCAN

UMAP¹

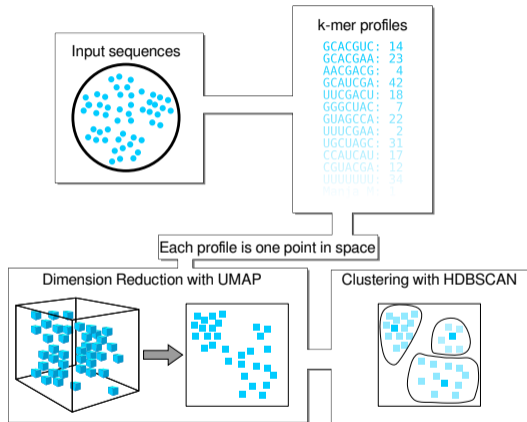
- ▶ connect data in space using simplices (based on fuzzy open cover)
- ▶ find low-dimensional representation with similar topological representation

HDBSCAN²

- ▶ 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>

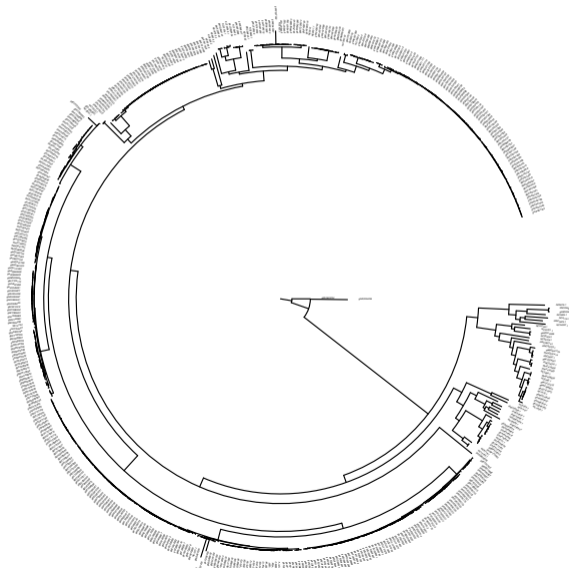


FIRST RESULTS

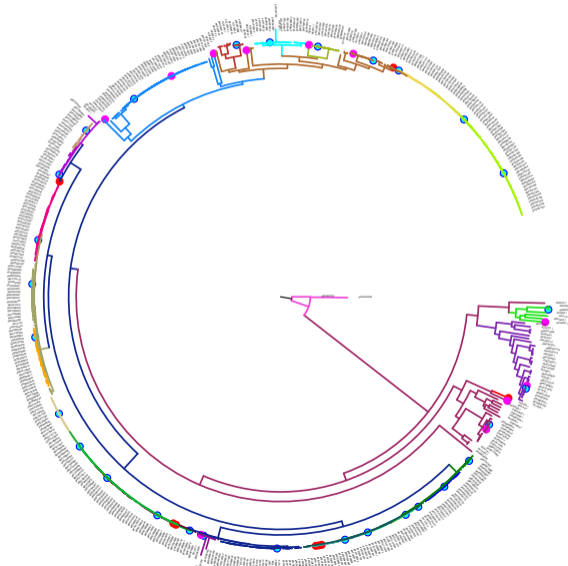
Viral Clade	Input Sequences	Cluster	Runtime ¹ in sec.
Denguevirus	5,470	302	5,695
Ebolavirus	634	33	43
Filoviridae	728	39	56
Zikavirus	789	38	36
Alphacoronavirus	927	57	141
Betacoronavirus	1,146	67	223
Poxviridae	688	45	467
Herpesviridae	1,758	89	2,023

¹: 8 cores, 3.6 GHz

MY APPROACH VERSUS INTUITION



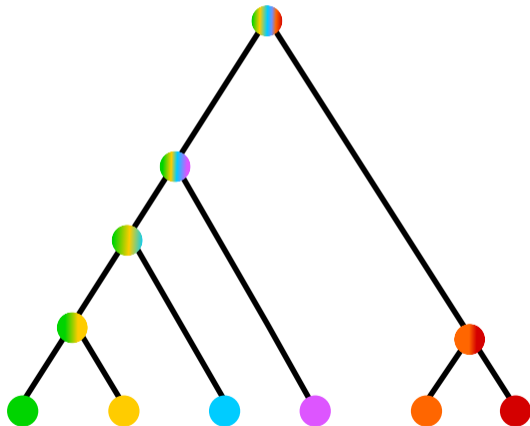
MY APPROACH VERSUS INTUITION



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

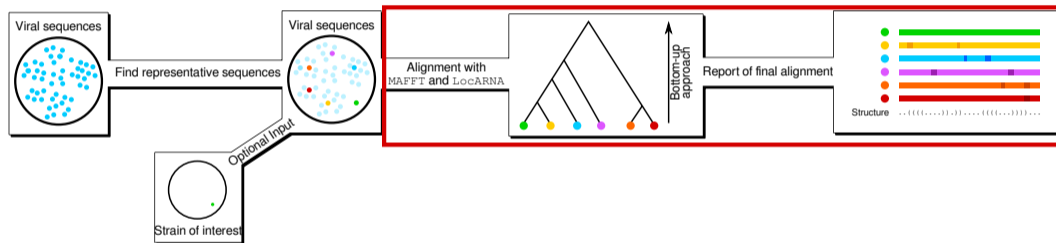
Order (*-virales*)

Family (*-viridae*)

Genus (*-virus*)

Species (*-virus*)

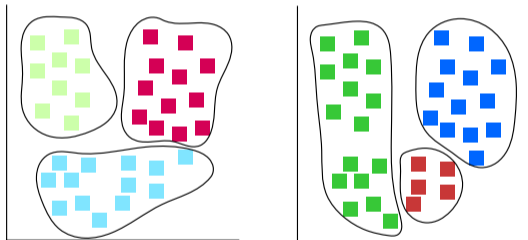
MY PROJECT: SECONDARY STRUCTURES



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

Order (*-virales*)

Family (*-viridae*)

Genus (*-virus*)

Species (*-virus*)

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!



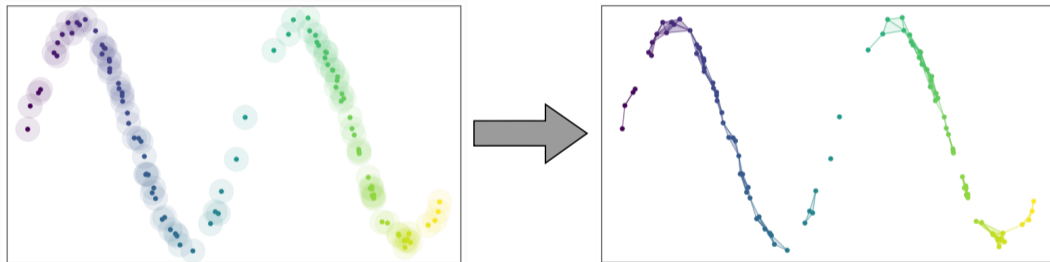
Acknowledgements:



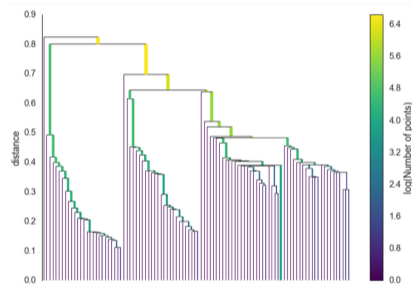
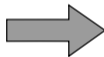
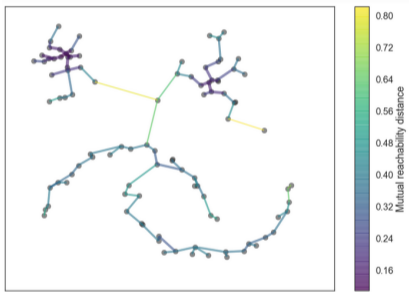
RNA Bioinformatics Group Jena



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