Application of FracMinHash to analyse the phylogenetic context of *Phytophthora*

Felix Seidel
felix.seidel@student.uni-tuebingen.de

June 6, 2024

Overview

A short story

The setting

Picture of idyllic farms, potato, maybe avocado?

The villain: *Phytophthora*

Picture of affected plants, maybe three bullet points (yearly damage)

The hero?

”Unfortunately, we cannot just walk to some mountain and throw jewelrey in it to defeat the villain” Research on the modes of operation, effector genes, and **phylogeny** of that species

Background

Phylogenetic Outline[1]

[1] Bagci et al., “Microbial Phylogenetic Context Using Phylogenetic Outlines”; Bryant and Huson, “NeighborNet”.

Phylogenetic Context[2]

[2] Bagci et al., “Microbial Phylogenetic Context Using Phylogenetic Outlines”.

FracMinHash[3] (1)

Idea: Hash k-mers and keep values smaller or equal to a threshold

- Sequence: *ATGCATGATG*

[3] Irber et al., *Lightweight Compositional Analysis of Metagenomes with FracMinHash and Minimum Metagenome Covers.*

FracMinHash[3] (1)

Idea: Hash k-mers and keep values smaller or equal to a threshold

- Sequence: *ATGCATGATG*
- k-mers: *ATG*, *TGC*, *GCA*, *CAT*, *ATG*, *TGA*, *GAT*, *ATG*

[3] Irber et al., *Lightweight Compositional Analysis of Metagenomes with FracMinHash and Minimum Metagenome Covers.*

FracMinHash[3] (1)

Idea: Hash k-mers and keep values smaller or equal to a threshold

- Sequence: *ATGCATGATG*
- k-mers: *ATG*, *TGC*, *GCA*, *CAT*, *ATG*, *TGA*, *GAT*, *ATG*
- hash values: $1, 8, 3, 4, 1, 10, 6, 1$

[3] Irber et al., *Lightweight Compositional Analysis of Metagenomes with FracMinHash and Minimum Metagenome Covers.*

FracMinHash[3] (1)

Idea: Hash k-mers and keep values smaller or equal to a threshold

- Sequence: *ATGCATGATG*
- k-mers: *ATG*, *TGC*, *GCA*, *CAT*, *ATG*, *TGA*, *GAT*, *ATG*
- hash values: $1, 8, 3, 4, 1, 10, 6, 1$
- define threshold: 3

[3] Irber et al., *Lightweight Compositional Analysis of Metagenomes with FracMinHash and Minimum Metagenome Covers.*

FracMinHash[3] (1)

Idea: Hash k-mers and keep values smaller or equal to a threshold

- Sequence: $ATGCATGATG$
- k-mers: $ATG, TGC, GCA, CAT, ATG, TGA, GAT, ATG$
- hash values: $1, 8, 3, 4, 1, 10, 6, 1$
- define threshold: 3
- FracMinHash sketch: $\{1, 3\}$

[3] Irber et al., *Lightweight Compositional Analysis of Metagenomes with FracMinHash and Minimum Metagenome Covers.*

FracMinHash[4] (2)

- W: input sequence
- h: hash function producing values in $[0, H]$
- s: scaling parameter $0 < s \leq H$
- $k(W)$: set of k-mers of W

$$\mathsf{FRAC}_s(W) = \{h(w) \leq \frac{H}{s} \mid \forall w \in k(W)\} \tag{1}$$

[4] Irber et al., *Lightweight Compositional Analysis of Metagenomes with FracMinHash and Minimum Metagenome Covers.*

From sketches to distances[5]

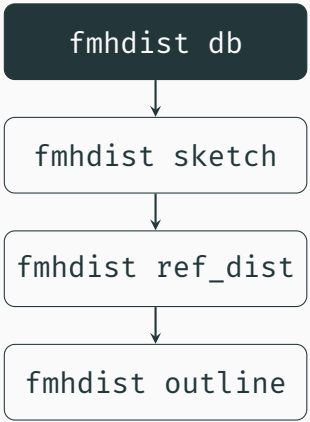
$$J_{frac}(A,B) = \frac{1}{1-(1-\frac{1}{s})^{|A\cup B|}} \frac{|\mathbf{FRAC}_s(A) \cap \mathbf{FRAC}_s(B)|}{|\mathbf{FRAC}_s(A) \cup \mathbf{FRAC}_s(B)|} \tag{2}$$

$$D_{frac}(A,B) = 1 - \left(\frac{2J_{frac}(A,B)}{1+J_{frac}(A,B)}\right)^{\frac{1}{k}} \tag{3}$$

[5] Hera et al., "Deriving Confidence Intervals for Mutation Rates across a Wide Range of Evolutionary Distances Using FracMinHash".

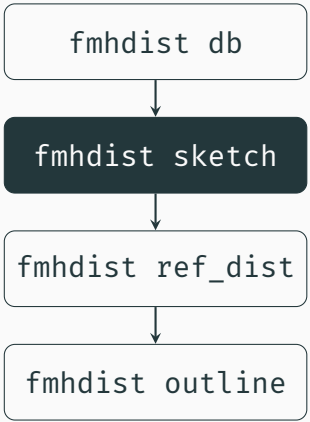
Phylogenetic Context of *Phytophthora*

The implementation with `fmhdist` (1)



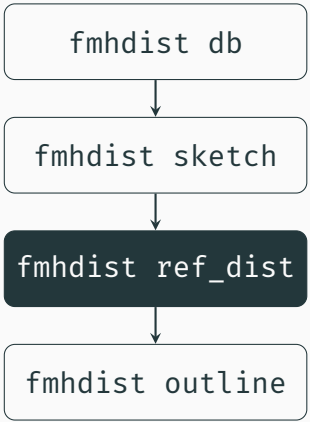
- Input: List of NCBI accession codes
- Sketching parameters: s, k, h and random seed for h
- Output: Reference database

The implementation with `fmhdist` (2)



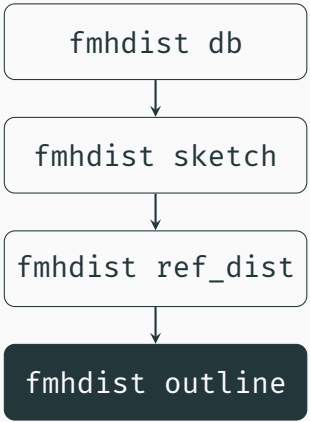
- Input: List of paths to FASTA files
- Sketching parameters: s, k, h and random seed for h
- Output: sketches and their coordinates

The implementation with `fmhdist` (3)



- Input: reference databse (`fmhdist db` or `fmhdist sketch`), query sketches (`fmhdist sketch`)
- Distance threshold
- Output: distance matrix

The implementation with `fmhdist` (4)



- Input: distance matrix
- Image parameters (width, height, scaling, offset)
- Output: phylogenetic outline as SVG

An outline produced by `fmhdist`

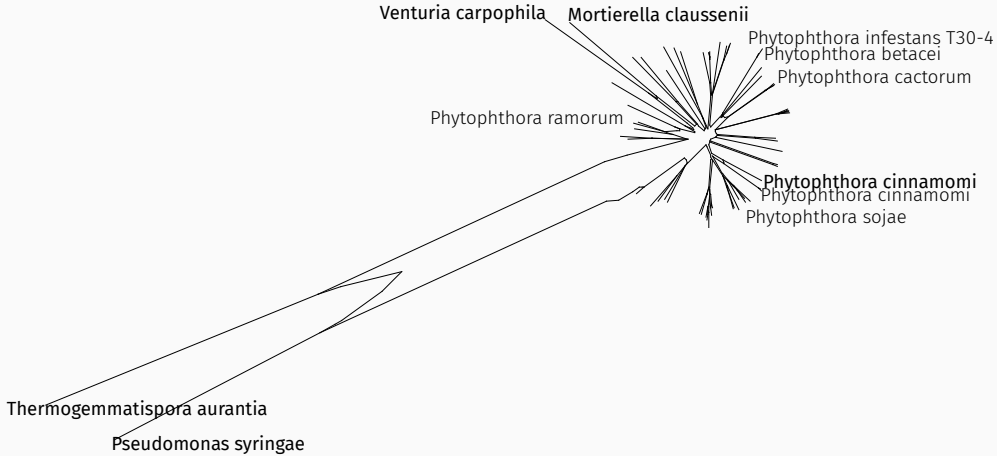


Figure 1: Outline based on FracMinHash distances of *Phytophthora* reference sequences and fungal and bacterial query sequences (bold). Only some labels shown.

Compare this to Mash

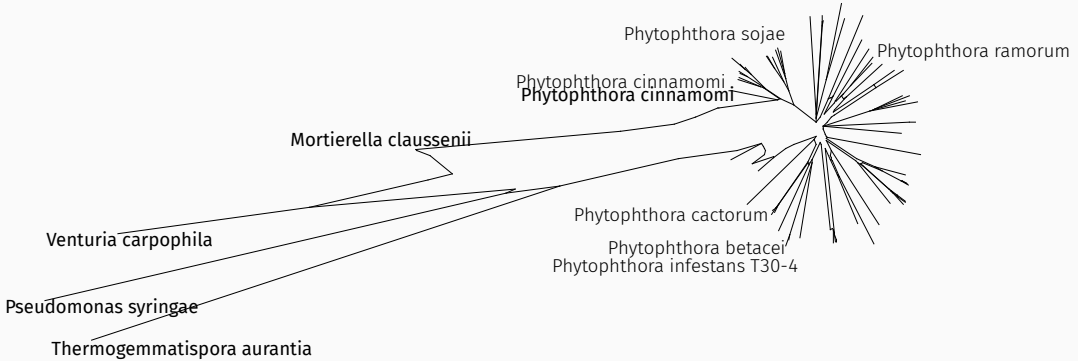


Figure 2: Outline based on Mash distances of *Phytophthora* reference sequences and fungal and bacterial query sequences (bold). Only some labels shown.

Origin of the hashes in the sketch of *Phytophthora sojae*

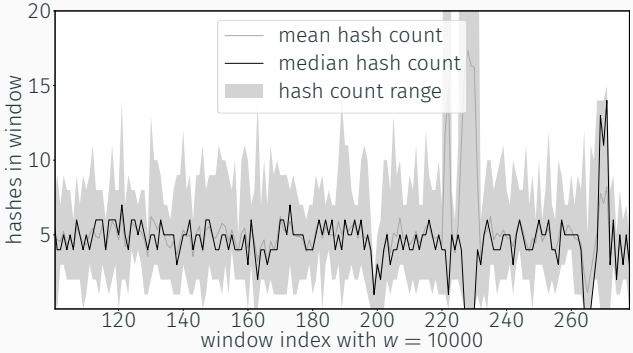


Figure 3: Hash counts in windows of the NW_009258123.1 sequence of the *Phytophthora sojae* reference genome.

...and the corresponding sequence complexity[6]

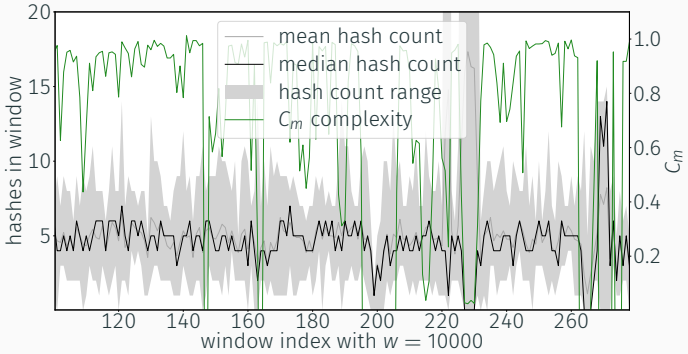


Figure 4: Hash counts and sequence complexity in windows of the NW_009258123.1 sequence of the *Phytophthora sojae* reference genome.

[6] Pirogov et al., "High-Complexity Regions in Mammalian Genomes Are Enriched for Developmental Genes".

How about other genomes?

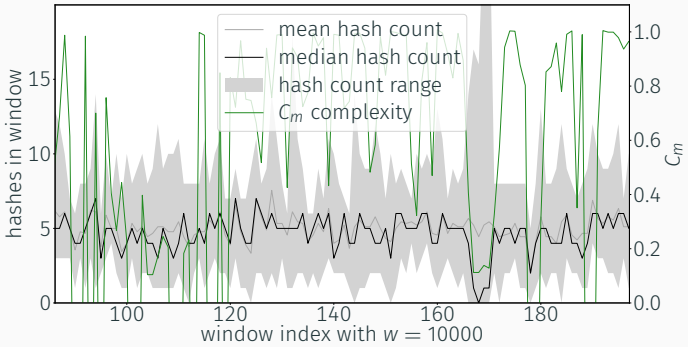


Figure 5: Hash counts and sequence complexity in windows of the NW_003303751.1 sequence of the *Phytophthora infestans* reference genome.

Statistical analysis

	n	u	r	p
P. cambivora	727	0	-0.0272	
P. betacei	26688	437	0.2446	4.5575e-265
P. cinnamomi	10799	28	-0.0933	4.3767e-19
P. infestans	11173	37	0.1830	2.8352e-24
P. sojae	7339	74	0.1570	6.5134e-48
P. nicotianae	2529	0	-0.0042	

Table 1: Excerpt of the statistical analysis of window ($w = 10000$) hash counts and sequence complexity of *Phytophthora* genomes.

Benchmarks

	`mash`[7] (1)	`mash` (6)	`sourmash`[8] (1)	`fmhdist` (1)	`fmhdist` (6)
min (s)	135	44	171	201	75
max (s)	140	58	178	215	91
avg (s)	137	51	174	208	84

Table 2: Runtime comparison of three tools calculating sketches for sequences totalling 5.477Gb. The number of threads is in parantheses.

[7] Ondov et al., "Mash".

[8] Irber et al., *Lightweight Compositional Analysis of Metagenomes with FracMinHash and Minimum Metagenome Covers*.

Conclusion

Method works in general

Potential next steps

Thanks!