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

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

June 7, 2024

Overview

A short story

The setting

Figure 1: A potato farm in Sweden[1].

[1] Magnusson, `https://flic.kr/p/85qcc9`, CC BY-ND 2.0

The trace

Figure 2: Infected potato plant[2].

Figure 3: Infected potato tuber[3].

[3] Cooke, `https://flic.kr/p/nPy5Qa`, CC BY-SA 2.0
[3] Millet, `https://flic.kr/p/qRXVt`, CC BY-NC-ND 2.0

The villain: *Phytophthora*

Phytophthora infestans

Phytophthora cinnamomi

Phytophthora ramorum

The hero?

Research on Phylogeny[4][5]

Research on effector proteins[6] and genome architecture[7]

Research on impact on microbial communities[8]

[8] Yang et al., “An Expanded Phylogeny for the Genus Phytophthora”

[8] Abad et al., “Phytophthora”

[8] Raffaele et al., “Analyses of Genome Architecture and Gene Expression Reveal Novel Candidate Virulence Factors in the Secretome of Phytophthora Infestans”

[8] Dong et al., “The Two-Speed Genomes of Filamentous Pathogens”

[8] Solís-García et al., “Phytophthora Root Rot Modifies the Composition of the Avocado Rhizosphere Microbiome and Increases the Abundance of Opportunistic Fungal Pathogens”

Background

Phylogenetic Trees[9][10][11]

- X: set of n taxa
- D: distance matrix for X
- $S = A|B$: a bipartition of X (split) based on D
- $\omega(S)$: the weight of S
- Σ: the set of all splits

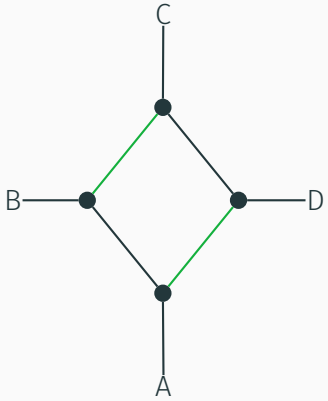
[9] Scornavacca et al., "Splits and Unrooted Phylogenetic Networks".

[10] Bryant and Huson, "NeighborNet".

[11] Bagci et al., "Microbial Phylogenetic Context Using Phylogenetic Outlines".

Phylogenetic Outlines[12][13][14]

- X: set of n taxa
- D: distance matrix for X
- $S = A|B$: a bipartition of X (split) based on D
- $\omega(S)$: the weight of S
- Σ: the set of all splits

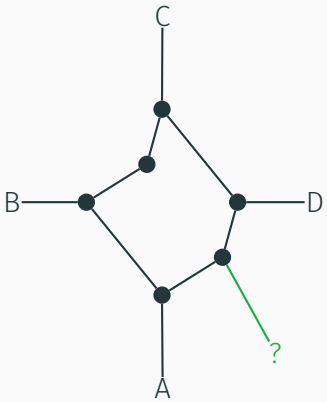


[12] Scornavacca et al., "Splits and Unrooted Phylogenetic Networks".

[13] Bryant and Huson, "NeighborNet".

[14] Bagci et al., "Microbial Phylogenetic Context Using Phylogenetic Outlines".

Phylogenetic Context[15]



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

FracMinHash[16] (1)

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

- Sequence: *ATGCATGATG*

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

FracMinHash[16] (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*

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

FracMinHash[16] (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: $1, 8, 3, 4, 1, 10, 6, 1$

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

FracMinHash[16] (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

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

FracMinHash[16] (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: $\{1, 3\}$

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

FracMinHash[17] (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

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

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

From sketches to distances[18]

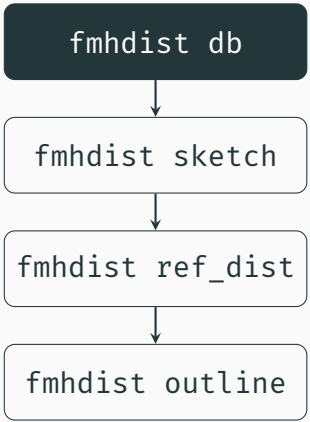
$$J_{frac}(A,B) = \frac{1}{1-(1-\frac{1}{s})^{|A\cup B|}} \frac{|\mathbf{FRAC}_s(A) \cap \mathbf{FRAC}_s(B)|}{|\mathbf{FRAC}_s(A) \cup \mathbf{FRAC}_s(B)|} \tag{2}$$

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

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

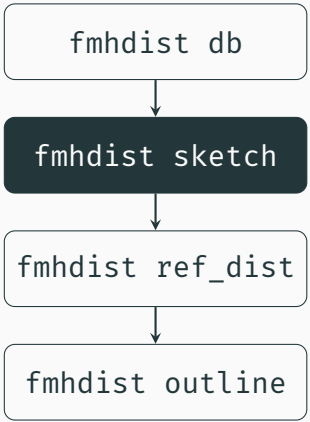
Phylogenetic Context of *Phytophthora*

The implementation with `fmhdist` (1)



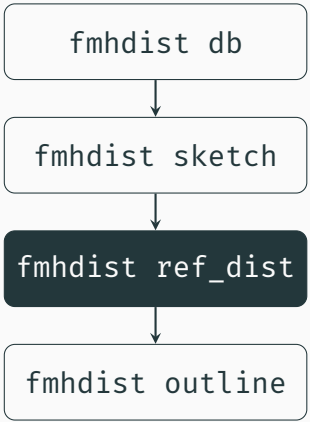
- Input: List of NCBI accession codes
- Sketching parameters: s, k, h and random seed for h
- Output: Reference database

The implementation with `fmhdist` (2)



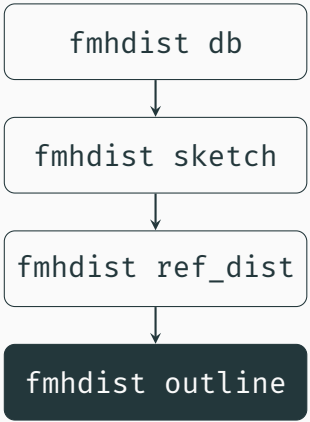
- Input: List of paths to FASTA files
- Sketching parameters: s, k, h and random seed for h
- Output: sketches and their coordinates

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`

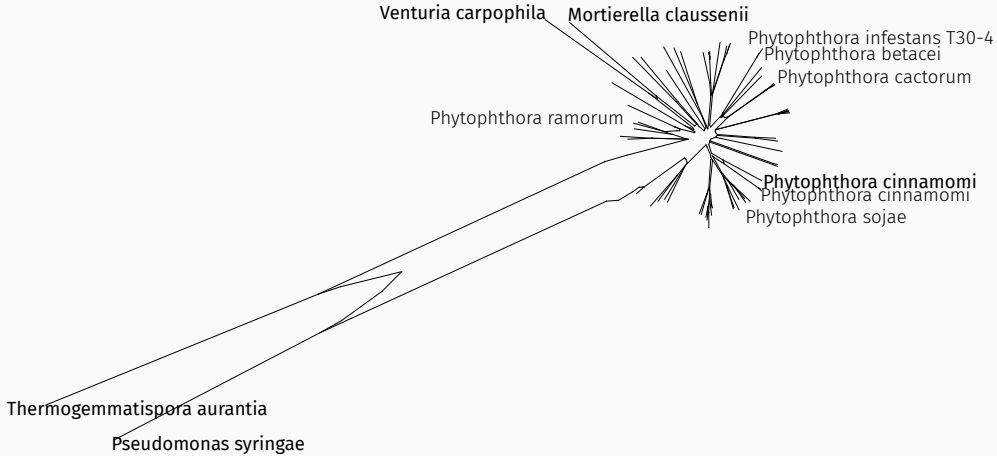


Figure 4: Outline based on FracMinHash distances of *Phytophthora* reference sequences and fungal and bacterial query sequences (bold). Only some labels shown.

Compare this to Mash

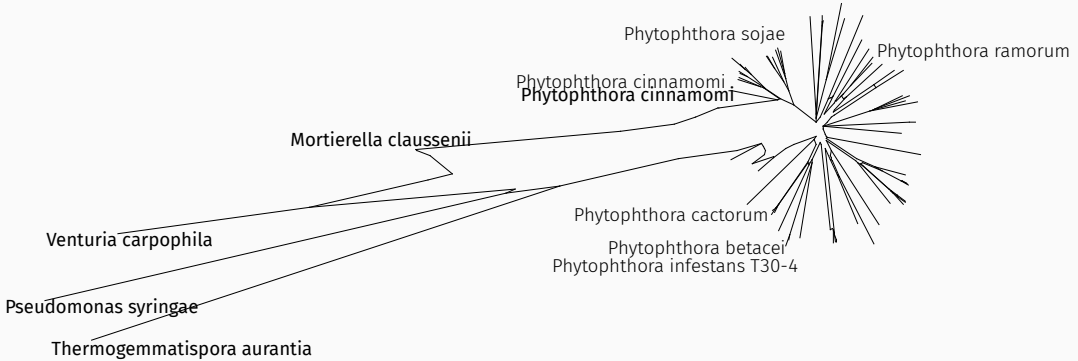


Figure 5: Outline based on Mash distances of *Phytophthora* reference sequences and fungal and bacterial query sequences (bold). Only some labels shown.

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

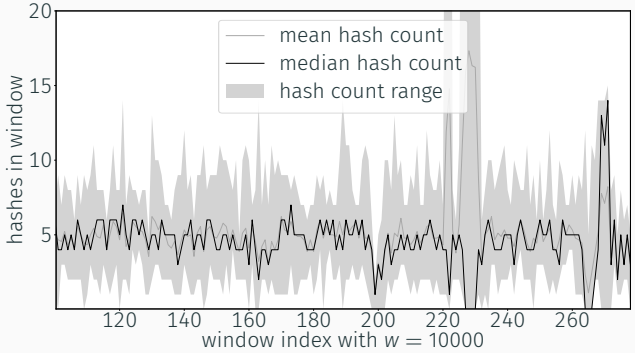


Figure 6: Hash counts in windows of the NW_009258123.1 sequence of the *Phytophthora sojae* reference genome.

...and the corresponding sequence complexity[19]

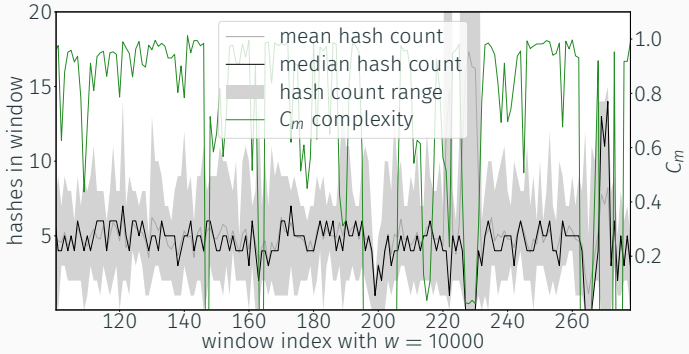


Figure 7: Hash counts and sequence complexity in windows of the NW_009258123.1 sequence of the *Phytophthora sojae* reference genome.

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

How about other genomes?

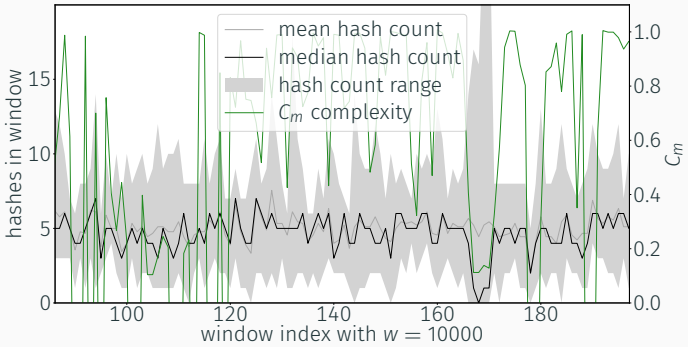


Figure 8: Hash counts and sequence complexity in windows of the NW_003303751.1 sequence of the *Phytophthora infestans* reference genome.

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`[20] (1)	`mash` (6)	`sourmash`[21] (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.

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

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

Conclusion

Conclusion

Summary

- FracMinHash is a solid foundation for Phylogenetic outlines.
- Distantly related genomes are better represented.
- Some *Phytophthora* genomes have windows with unusual densities.

Open questions

- What happens to the left and the right of windows with unusual densities?
- Can we use this method to assign a label to a draft genome?

Thanks!