University of Tübingen
Faculty of Science
Department of Computer Science

Master Thesis Bioinformatics

Title of thesis

First Name and Surname

Date

Reviewers

Name First Reviewer
(Bioinformatics)
Institute for Bioinformatics and
Medical Informatics
University of Tübingen

Name Second Reviewer
(Field of Study)
Institute of second reviewer
University of Tübingen

Surname, First Name:
Title of thesis
Master Thesis Bioinformatics
University of Tübingen
Thesis period: dd.mm.yyyy – dd.mm.yyyy

Abstract

Write here your abstract.

Acknowledgements

Write here your acknowledgements.

Contents

List of Figures

List of Tables

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
...	...

Chapter 1

Introduction

Chapter 2

Material and Methods

2.1 Libraries used for Implementation

The generation of phylogenetic outlines using FracMinHash was implemented using `java-17`, `jloda3` 1.0.0, `splitstree6` 1.0.0 [5] and `openapi-generator-maven-plugin` 6.3.0. For the hash functionality, `zero-allocation-hashing` 0.16 is used.

Check: Do I need a full list or just the main contributers? Do I need a reference for each?

2.2 Benchmarking the Hash Functions

To get an understanding of which hash function performs best in terms of runtime, the `jmh` benchmarking framework 1.37 was used to benchmark the following hash functions:

- MurMur3
- XX3
- XX64
- XX128
- City 1.1
- Farm 1.0
- Farm 1.1
- Wy 3
- Metro

add citation/links to all hash functions

While other hash function implementations were considered, the implementation of all hash functions is given by `zero-allocation-hashing` 0.16. The benchmark was executed in throughput mode using a fork count of 2. Besides that, default parameters were used.

2.3 Datasets used

To evaluate properties of FracMinHash, analysis was performed using different data sets.

The first dataset (**A**) consists of 128 different *Phytophthora* genomes. The list is taken from [10] without further modifications. Genomes were downloaded from NCBI uding the `datasets` utility [17] using the accession codes listed in that study.

The second dataset (**B**) consists of 72 mtDNA sequences of *Phytophthora* and other Oomycetes of the Peronosporaceae family. The list is taken without further modifications from Supplementary Table 1 from [20]. Genomes were downloaded from NCBI using the Entrez interface [17] using the accession codes listed in that study, appended by the identifier of the most recent version for that accession code.

The third dataset (**C**) consists of all 64 *Phytophthora* reference sequences in the NCBI databasase ("reference") as well as five different query sequences that are typically found in soil samples of avocado orchards that are infected with *Phytophthora cinnamomi* [18]. Those query sequences are divided into two bacterial genomes, two fungal and one *Phytophthora cinnamomi* genome:

- fungal: *Mortierella claussenii* - GCA_022750515.1
- bacterial: *Thermogemmatispora aurantia* - GCA_008974285.1
- fungal: *Venturia carpophila* - GCA_014858625.1
- bacterial: *Pseudomonas syringae* - GCA_018394375.1
- fungal: *Phytophthora cinnamomi* - GCA_001314365.1

ref

The full list of reference sequences can be found in the appendix . The reference sequences were downloaded using the web interface (Filter: reference sequences), the query sequences were downloaded using the `datasets` utility.

2.4 Comparison with published Phylogenies

[10] displays a rooted phylogenetic tree in Figure 4 that was generated using the `mashtree` [6, 12]. Unfortunately, the corresponding distance matrix or a

serialized version of the tree are not available. Thus, the data needs to be recomputed. For this, the `mashtree_bootstrap.pl` 1.4.6 was applied to dataset A as described by [10], additionally the distance matrix was saved using the `--outmatrix` parameter.

Distance matrices for dataset A using the FracMinHash method were calculated using different combinations of the scaling parameter s and k-mer size k:

- $k = 19$, $s = 2000$
- $k = 20$, $s = 2000$
- $k = 21$, $s = 2000$
- $k = 25$, $s = 2000$
- $k = 30$, $s = 2000$
- $k = 21$, $s = 500$
- $k = 21$, $s = 1000$
- $k = 21$, $s = 40000$

To ensure comparability with the `mashtree` results, the same hashing function (MurMur) and random seed (42) were used to calculate the sketches [6, 12].

For all distance matrices, SplitsTree 6.0.0-alpha [5] was used to obtain trees, splits and outlines. Trees were calculated using the Neighbor Joining method [16]. Splits were obtained using the Neighbor Net method [2, 3] using the default parameters. Based on this, a phylogenetic outline was calculated [1].

2.5 Split difference analysis

To analyse the differences between the phylogenetic outline based on the Mash distances and the outlines based on FracMinHash further, the splits were analysed in details. For this, the splits were exported from SplitsTree 6.0.0-alpha [5] using the plain text format.

Those files were processed with the script `compare_splity.py`. This is based on `python` 3.12 and `pandas` 2.1.4 [11, 19]. All sets of splits are compared pairwise, that is for two sets Σ_A and Σ_B, the following properties are calculated:

- the total sum of weight for all splits in Σ_A and Σ_B, respectively: $\sum_{s \in \Sigma_A} \omega(s)$ and $\sum_{s \in \Sigma_B} \omega(s)$

- the number of splits in Σ_A/Σ_B and the number of splits in Σ_B/Σ_A
- the total weight difference associated with the remaining splits, i.e. $\sum_{s\in\Sigma_A/\Sigma_B}\omega(s)$ and $\sum_{s\in\Sigma_B/\Sigma_A}\omega(s)$
- the Robinson-Foulds distance [14] of the two sets of splits, i.e. $D_{RF} = \frac{1}{2}|(\Sigma_A/\Sigma_B)\cup(\Sigma_B/\Sigma_A)|$

The script outputs a list of the diverging splits sorted by the weight of the split. Each split is formatted as a list of taxa names on the smaller side of the split sorted alphabetically and joined using the "|" symbol such that the split can be searched in SplitsTree 6 using the regular expression search.

2.6 Reproducing phylogenies for shorter sequences

To get an idea of the practical lower boundarys of FracMinHash in terms of input genome size, dataset B was sketched. As this dataset consists of just a single long FASTA file, it was split into its records using `split-fasta` 1.0.0 . The files were then sketched with FracMinHash using $k = 21$, $s \in \{1, 10, 50, 100, 1000\}$ and the MurMur hash function with random seed 42.

check citation

redo with most recent implementation, check sketch sizes. Redo because I don't know with which version the sketches were computed, thus I don't know how to calculate the sizes.

Given the calculated distances, SplitsTree 6.0.0-alpha [5] was used to obtain trees, splits and outlines. Trees were calculated using the Neighbor Joining method [16]. Splits were obtained using the Neighbor Net method [2, 3] using the default parameters. Based on this, a phylogenetic outline was calculated [1].

2.7 Calculating distances of distantly related genomes and genomes with different sizes

To analyse the claim that FracMinHash works better for genomes of different sizes , both reference and query sequences of dataset C were sketched, distances calculated and the results compared.

Sketching using FracMinHash

ensure this claim is cited at least once!

Reference and query sequences were sketched with $k = 21$, $s = 2000$ using FarmHash with random seeds $rs \in \{10, 20, 30, 40, 50\}$. As the default sketch size for Mash is 10000 [12] and to ensure that it is not the fact that the relevant sketches are just larger, additional sketching with $s = 500$ (which aims to

bring the sketch size of the bacterial query sequences to 10000) and $s = 3500$ (which aims to bring the sketch size of the fungal query sequences to 10000) was performed. Using the sketches, the distance matrix was calculated. For this, the implementation was changed such that an additional log is created that lists all empty intersections when calculating the intersection of the two sketches.

Sketching using MashTree

To obtain a base value to compare agains, `mashtree` 1.4.6 [6, 12] was applied with hash seed $rs \in \{10, 20, 30, 40, 50\}$, `--outmatrix` and the `--save-sketches` parameter. The resulting sketches were then also converted into JSON format using `mash info -d`.

Visual comparison and split differences

Given all calculated distances, SplitsTree 6.0.0-alpha [5] was used to obtain trees, splits and outlines. Trees were calculated using the Neighbor Joining method [16]. Splits were obtained using the Neighbor Net method [2, 3] using the default parameters. Based on this, a phylogenetic outline was calculated [1].

The splits were then also analysed using the script outlined in Section 2.5.

Obtaining reference values

As there are no published phylogenies for dataset C that enable a comparison, I need a different ground truth to compare the results of above distance calculation against. As the aim for this experiment is to analyze the influcence of different genome sizes of distantly related organisms, I have limited this analysis to the five query sequences of dataset C and the reference genome for *Phytophthora infestans* (GCF_000142945.1) as this is one of the largest genomes in the set of reference sequences. For those sequences, I have prepared two different values to compare against.

The first ist the Average Nucleotide Identity (ANI) [7] calculated with `OrthoANI` [8] using the default parameters with `BLAST+` 2.14.1 [4] as the aligner.

The second is the Average Amino Acid Identity (AAI) calculated with the `Enveomics Collection` online resource [15] using default parameters. As this requires amino acid sequences as input and those are not available for the *Venturia carpophila* and *Phytophthora cinnamomi* sequences, they were predicted using the `gmes_petap.pl` script provided in the `GeneMark ES` 4.72 package [9] using the default parameters and extracted using the `gffread` utility 0.12.7 [13].

Calculating empty intersections

When the sketch intersections are empty, the estiamted Jaccard similarity is always 0. Mash does not only calculate the sketch intersections, but also intersects that result with the sketch of the union of the two input sketches [12]. This increases the chances of getting a Jaccard estimation of 0.

To test this, I have prepared a the script `get_empty_intersections.py` which takes a list of Mash sketches in JSON format and outputs all pairs for which the numerator of the Jaccard estimation is empty.

The same is possible using the added log output generated by the distance calculation described in Section 2.7.

Using the five different random seeds, we can convert this output into a table that lists the number of empty intersections for each pairwise calculation.

2.8 Hash Density analysis

Chapter 3

Results

In this chapter which also could be more than one chapter, depending on the nature of the thesis, the results of the thesis are presented. Make sure you illustrate your results with appropriate figures and tables, but do not discuss the results here. This should be done in a separate discussion chapter.

Chapter 4

Discussion

Of course very important! You need to discuss the informatics as well as bio part of your thesis topic.

This section should include the following:

- Short summary of the aim of the study
- Discussion of the results
- Contextualisation of the results/ Putting the results into context (that was set in the introduction)
- You can also discuss limitations of your thesis and formulate an outlook for future/ follow-up research (Outlook can become an extra chapter.)
- Finally, you can add a short conclusion of the main findings/ take aways of the thesis

Take your time for writing the discussion, besides the introduction chapter it is the most important chapter of your thesis.

Also do not subsection the discussion too heavily.

At least 5 pages.

Appendix A

Further Tables and Figures

A.1 List of reference sequences for dataset C

- GCA_000439335.1
- GCA_000443045.1
- GCA_000468175.2
- GCA_000500205.2
- GCA_000500225.2
- GCA_000687305.2
- GCA_001314345.1
- GCA_001314375.1
- GCA_001314425.1
- GCA_002215365.1
- GCA_002812785.1
- GCA_007655245.1
- GCA_008079305.1
- GCA_008080845.1
- GCA_009729435.1
- GCA_011320135.1
- GCA_011947325.1

- GCA_011947335.1
- GCA_011947345.1
- GCA_011947355.1
- GCA_012295415.1
- GCA_012295475.1
- GCA_012656075.1
- GCA_012656105.1
- GCA_014706105.1
- GCA_014706115.1
- GCA_014706125.1
- GCA_014706135.1
- GCA_014706145.1
- GCA_016169955.1
- GCA_016864655.1
- GCA_016880985.1
- GCA_018691715.1
- GCA_018806915.1
- GCA_018873745.1
- GCA_019155715.1
- GCA_020800215.1
- GCA_023611945.1
- GCA_024211575.1
- GCA_024679045.1
- GCA_024679075.1
- GCA_024679115.1
- GCA_024679135.1

- GCA_024679175.1
- GCA_024679225.1
- GCA_024679275.1
- GCA_024679295.1
- GCA_025722995.1
- GCA_030027945.1
- GCA_030267725.1
- GCA_030267785.1
- GCA_030324255.1
- GCA_030463285.1
- GCA_031305395.1
- GCA_032158285.1
- GCA_032432875.1
- GCA_033557915.1
- GCA_033557925.1
- GCA_033557995.1
- GCA_033558005.1
- GCA_033558025.1
- GCF_000142945.1
- GCF_000149755.1
- GCF_000247585.1

Bibliography

[1] Caner Bagci, David Bryant, Banu Cetinkaya, and Daniel H Huson. "Microbial Phylogenetic Context Using Phylogenetic Outlines". In: *Genome Biology and Evolution* 13.9 (Sept. 1, 2021), evab213. DOI: 10.1093/gbe/evab213.

[2] David Bryant and Daniel H. Huson. "NeighborNet: Improved Algorithms and Implementation". In: *Frontiers in Bioinformatics* 3 (Sept. 20, 2023), p. 1178600. DOI: 10.3389/fbinf.2023.1178600. pmid: 37799982.

[3] David Bryant and Vincent Moulton. "Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks". In: *Molecular Biology and Evolution* 21.2 (Feb. 1, 2004), pp. 255–265. DOI: 10.1093/molbev/msh018.

[4] Christiam Camacho et al. "BLAST+: Architecture and Applications". In: *BMC Bioinformatics* 10.1 (Dec. 15, 2009), p. 421. DOI: 10.1186/1471-2105-10-421.

[5] Daniel H. Huson and David Bryant. "Application of Phylogenetic Networks in Evolutionary Studies". In: *Molecular Biology and Evolution* 23.2 (Feb. 1, 2006), pp. 254–267. DOI: 10.1093/molbev/msj030.

[6] Lee S. Katz et al. "Mashtree: A Rapid Comparison of Whole Genome Sequence Files". In: *Journal of Open Source Software* 4.44 (Dec. 10, 2019), p. 1762. DOI: 10.21105/joss.01762.

[7] Konstantinos T. Konstantinidis and James M. Tiedje. "Towards a Genome-Based Taxonomy for Prokaryotes". In: *Journal of Bacteriology* 187.18 (Sept. 15, 2005), pp. 6258–6264. DOI: 10.1128/jb.187.18.6258-6264.2005.

[8] Imchang Lee, Yeong Ouk Kim, Sang-Cheol Park, and Jongsik Chun. "OrthoANI: An Improved Algorithm and Software for Calculating Average Nucleotide Identity". In: *International Journal of Systematic and Evolutionary Microbiology* 66.2 (2016), pp. 1100–1103. DOI: 10.1099/ijsem.0.000760.

[9] Alexandre Lomsadze, Vardges Ter-Hovhannisyan, Yury O. Chernoff, and Mark Borodovsky. "Gene Identification in Novel Eukaryotic Genomes by Self-Training Algorithm". In: *Nucleic Acids Research* 33.20 (Nov. 1, 2005), pp. 6494–6506. DOI: 10.1093/nar/gki937.

[10] Kajal Mandal, Subhajeet Dutta, Aditya Upadhyay, Arijit Panda, and Sucheta Tripathy. "Comparative Genome Analysis Across 128 Phytophthora Isolates Reveal Species-Specific Microsatellite Distribution and Localized Evolution of Compartmentalized Genomes". In: *Frontiers in Microbiology* 13 (Mar. 16, 2022), p. 806398. DOI: 10.3389/fmicb.2022.806398. pmid: 35369471.

[11] Wes McKinney. "Data Structures for Statistical Computing in Python". In: Python in Science Conference. Austin, Texas, 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[12] Brian D. Ondov et al. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash". In: *Genome Biology* 17.1 (June 20, 2016), p. 132. DOI: 10.1186/s13059-016-0997-x.

[13] Geo Pertea and Mihaela Pertea. *GFF Utilities: GffRead and GffCompare*. Sept. 9, 2020. DOI: 10.12688/f1000research.23297.2. F1000Research: 9:304. URL: https://f1000research.com/articles/9-304 (visited on 05/12/2024). preprint.

[14] D. F. Robinson and L. R. Foulds. "Comparison of Phylogenetic Trees". In: *Mathematical Biosciences* 53.1 (Feb. 1, 1981), pp. 131–147. DOI: 10.1016/0025-5564(81)90043-2.

[15] Luis M. Rodriguez-R and Konstantinos T. Konstantinidis. *The Enveomics Collection: A Toolbox for Specialized Analyses of Microbial Genomes and Metagenomes*. e1900v1. PeerJ Inc., Mar. 27, 2016. DOI: 10.7287/peerj.preprints.1900v1.

[16] N Saitou and M Nei. "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees." In: *Molecular Biology and Evolution* 4.4 (July 1, 1987), pp. 406–425. DOI: 10.1093/oxfordjournals.molbev.a040454.

[17] Eric W. Sayers et al. "Database Resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 50.D1 (Jan. 7, 2022), pp. D20–D26. DOI: 10.1093/nar/gkab1112. pmid: 34850941.

[18] Itzel A. Solís-García et al. "Phytophthora Root Rot Modifies the Composition of the Avocado Rhizosphere Microbiome and Increases the Abundance of Opportunistic Fungal Pathogens". In: *Frontiers in Microbiology* 11 (2020), p. 574110. DOI: 10.3389/fmicb.2020.574110. pmid: 33510714.

[19] The pandas development team. *Pandas-Dev/Pandas: Pandas*. Version v2.2.2. Zenodo, Apr. 10, 2024. DOI: 10.5281/zenodo.10957263.

[20] Richard C. Winkworth et al. "Comparative Analyses of Complete Peronosporaceae (Oomycota) Mitogenome Sequences—Insights into Structural Evolution and Phylogeny". In: *Genome Biology and Evolution* 14.4 (Apr. 10, 2022). Ed. by Liliana Milani, evac049. DOI: 10.1093/gbe/evac049.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Masterarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum Unterschrift