Scaled MinHash Containment enables alignment-free distance estimation across the tree of life

This manuscript was automatically generated on March 30, 2021.

Authors

- **N. Tessa Pierce-Ward**
  0000-0002-2942-5331 · bluegenes · saltyscientist
  Department of Population Health and Reproduction, University of California, Davis · Funded by NSF 1711984, NSF 2018911

- **C. Titus Brown**
  0000-0001-6001-2677 · ctb · ctitusbrown
  Department of Population Health and Reproduction, University of California, Davis · Funded by Moore Foundation GBMF4551

Abstract

Background Sequence similarity estimation is critical for genome analyses ranging from taxonomic classification to phylogenetic reconstruction. Current practices are still insufficient for environmental samples. Even best-practices samples can only assign a small fraction of metagenomic sequencing to known species... Given the scale of sequencing data now available, there is a great need for tools that can accurately estimate pairwise genome relatedness in an assembly-free and alignment-free manner.

Results Here, we introduce an alignment-free k-mer based method for quickly and accurately estimating pairwise sequence similarity, including Average Nucleotide Identity (ANI) and Average Amino Acid Identity (AAI). Because this method is assembly-independent and sequence-agnostic, it can be applied to both DNA and protein sequence data from across the tree of life. We demonstrate the utility of this technique with two primary examples: phylogenetic reconstruction of the genome-based prokaryotic taxonomy (GTDB) and fast and accurate microbial taxonomic classification. To demonstrate utility in other domains, we demonstrate proof-of-concept analysis on two published eukaryotic and viral datasets.

Conclusions Containment-based pairwise distance estimation via Scaled Minhash enables accurate assembly-free and alignment-free phylogenomic reconstruction and taxonomic classification across a wide range of evolutionary distances.

Keywords (3-10)

Overall Genome Relatedness Index (OGRI), Average Nucleotide Identity (ANI), Average Amino Acid Identity (AAI), pairwise evolutionary distance

Background

As the scale of genomic sequencing continues to grow, alignment-free methods for estimating sequence similarity have become critical for conducting tasks ranging from taxonomic classification to phylogenetic analysis on large-scale datasets [1,2]. The majority of alignment-free methods rely upon exact matching of k-mers: subsequences of length k, that can be counted and compared across datasets, with or without use of subsampling methods such as MinHash. As k-mer based methods rely on exact sequence matches, they can suffer from limited sensitivity when comparing highly polymorphic sequences or classifying organisms from groups that are not well represented in reference databases.

Current best practices methods can still only categorize a fraction of the metagenomic and metatranscriptomic data, especially for understudied and/or diverse habitats (xx% recovery for soil, xx% recovery ocean metagenomes, etc). Even well-studied environments such as human gut can produce significant uncharacterized metagenome content. “For example, a reference-based approach failed to map 35% of reads in the iHMP study on inflammatory bowel disease (Supp. Data. of (Franzosa et al., 2019)), omitting them from any further analysis. These reads may belong to unknown microbes, phage or viruses, plasmids, or accessory elements of known microbes, all of which can play a role in disease.[from RO1]”. This phenomenon is not restricted to metagenome samples. Alignment-based estimates can fail at larger evolutionary distances and even rRNA amplicon surveys may underestimate bacterial diversity [3].

To increase sensitivity of alignment-free methods, modified k-mer approaches have been introduced, including spaced seeds /split k-mers, which accommodate polymorphic sites in highly similar genomes (CITE). For larger evolutionary distances, protein-based comparisons have long been the gold-standard approach for taxonomic and functional annotation, as protein sequence is more conserved than the underlying DNA sequence [4,5]. As microbial and viral genomes are gene-dense, [MinHash-based] alignment-free comparisons of translated protein sequence have been shown to increase sensitivity for taxonomic classification and genome discovery [6,7]. Here, we demonstrate the utility of protein k-mer comparisons for phylogenomic reconstruction and taxonomic classification at larger evolutionary distances and across both gene-rich and [gene-sparse] sequences. We use Scaled Minhash subsampling to facilitate conducting these comparisons at scale [8].

Scaled Minhash is a MinHash variant for selecting and hashing a set of representative k-mers from a sequence dataset [8]. Unlike traditional MinHash, Scaled MinHash sketches scale with the size of the dataset, meaning each sketch is comprised of the chosen proportion of k-mers in the input dataset, rather than a chosen number of k-mers. Downsampling sequencing datasets in this way enables estimation of containment, which has been shown to permit more accurate estimation of genomic distance, particularly for genomes of very different lengths [9,10]. Streaming containment estimates have been shown to facilitate genome discovery and correlate with Mash Distance, a proxy for Average Nucleotide Identity (ANI) [7,11].

Standardized genomic measures of relatedness such as ANI and its protein counterpart, Average Amino Acid Identity (AAI) have shown lasting utility for genome relatedness and phylogenomic analysis. Traditional ANI and AAI describe the sequence similarity of all orthologous genes, either in nucleotide or protein space, respectively. Both been shown to be robust measure of overall pairwise genome relatedness even for highly incomplete datasets, such as those comprised of only ~4% of the genome or 100 genes [12,13]. ANI has emerged as the most widely-accepted method for estimating pairwise similarity of microbial genomes and delimiting species boundaries [14]. Recent research appears to confirm 95% ANI species threshold for prokaryotic species, although there is some debate as to the universality of this threshold [15,16,17]. AAI thresholds have been proposed for higher taxonomic ranks, <45%, 45-65% and 65-95% for family, genus, and species [13,18]. While traditional alignment-based estimation of ANI and AAI are computationally intensive, sketching-based estimates and sketching-facilitated estimates have permitted ANI calculations at the scale of whole-databases [1,7,15].

[Pierce-Ward et al., 2021 (tbd technical paper)] showed that Scaled MinHash containment estimates can be used to approximate both ANI (nucleotide k-mers) and Average Amino Acid Identity (AAI; protein k-mers), while accounting for the non-independence of mutated k-mers [19]. Furthermore, Scaled MinHash containment estimates work well for genome pairs of varying lengths and for compositional analysis of metagenome samples. Taken together, these properties enable robust assembly and alignment-free pairwise relatedness estimation that can be used on sequences separated by a wide range of evolutionary distances. Here, we demonstrate that the utility of Scaled MinHash protein containment, both used directly and a an approximation of ANI and AAI, for taxonomic classification and phylogenomic reconstruction for species across the tree of life.

Notes

- AAI::phylogeny https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1236649/

Results

Protein k-mer containment enables similarity searches at increased evolutionary distances

(DNA vs Protein) - *(just containment, no ANI/AAI) - include dayhoff or just protein?*

Protein sequences are more conserved than their underlying DNA sequences. Whole-proteome MinHash sketches are more similar than whole-genome DNA sketches, enabling us to find protein-level similarity across divergent genomes.

For , e.g. Pseudomonas, XX% of k-mers are shared within the chosen/published genomes within species. For all published genomes within the genus, a median of xx% of k-mers are shared between genomes of one species and genomes of the a different species in the same genus.

rankinfo ... at ksize of 10... -xx% of DNA k-mers are shared within-species -yy% of protein k-mers are shared within-species - zz% of DNA k-mers are shared within-genus ... etc == median or mean containment at rank? containent = % of a genome's k-mers that are shared – do using ALL of gtdb, BUT, start with just a single set of genomes.. e.g. Pseudomonas? == similar to "shared k-mers" paper [20]

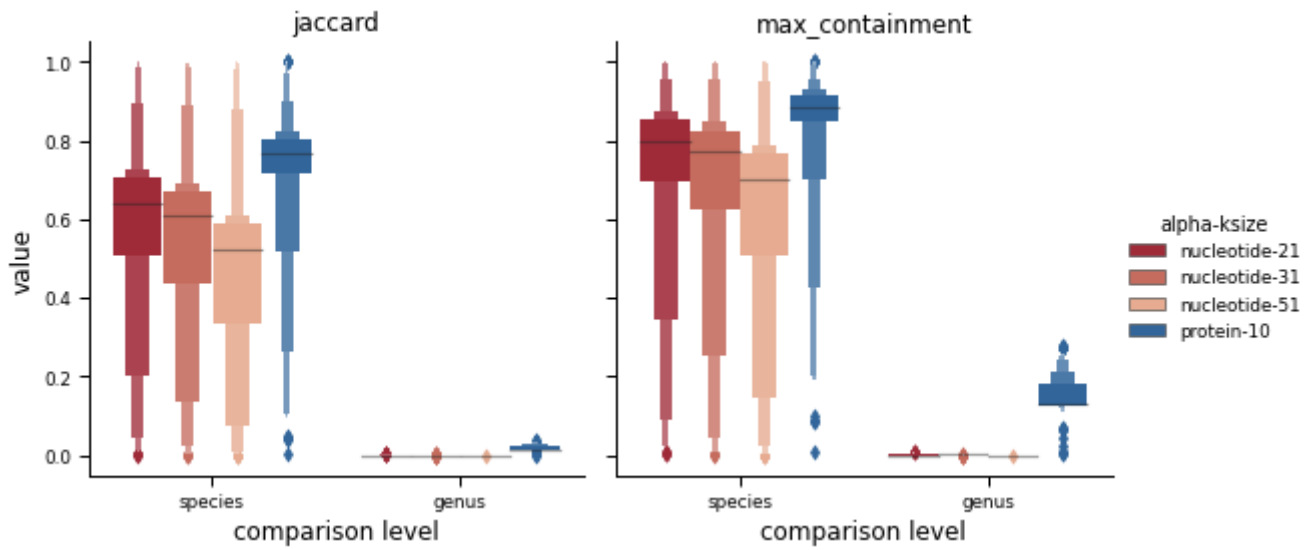

Figure 1: Protein k-mer containment facilitates genus-level comparisons 10k pseudomonas genome sequences, median containment at each alphabet

k-size selection for optimal comparisons / distance estimation

- num shared k-mers at different ksizes
- e.g. k=7 much more common – share far more k-mers. I assumed this would hurt, rather than help classification. Check!
- do rankinfo on each database??

because kmer size matters –> conversion to AAI is useful!? conversion to AAI does two things: accounts for k-mer length, ...

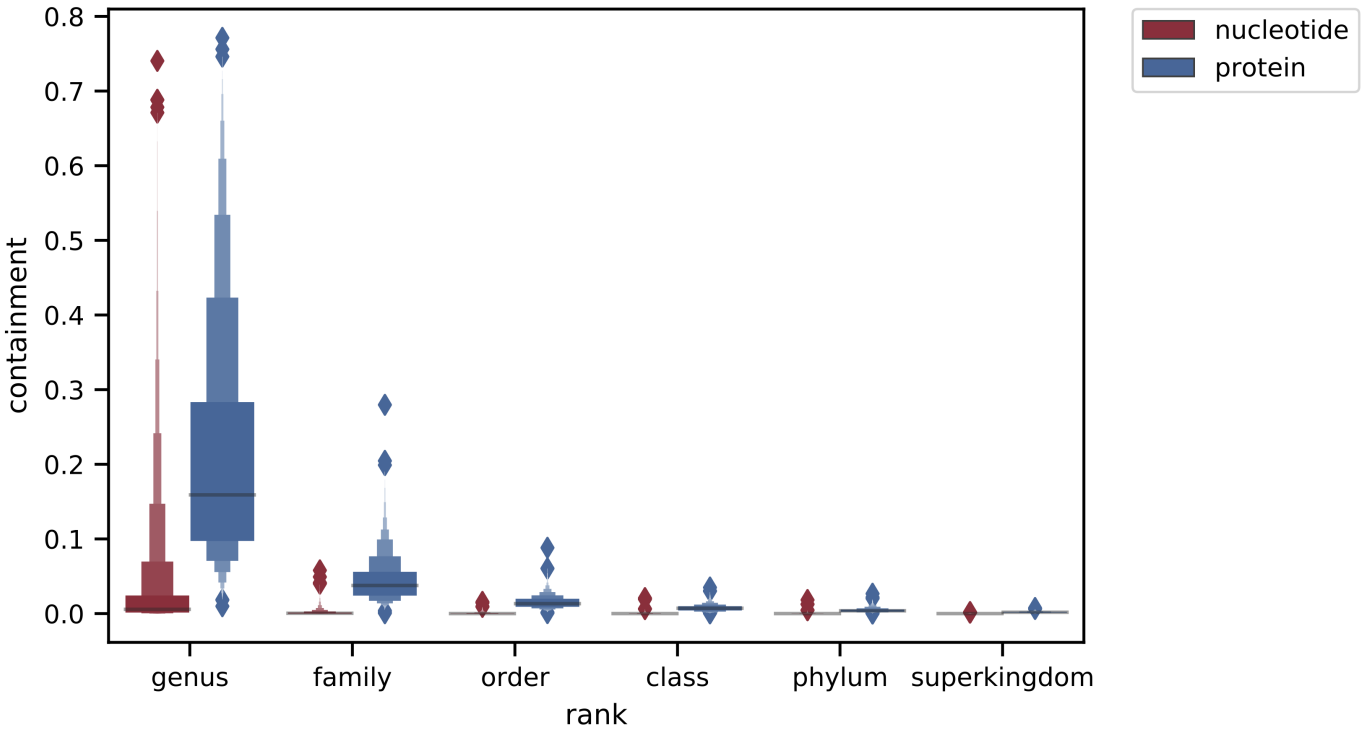

Figure 2: Protein k-mers facilitate comparisons across species This currently uses the evolutionary paths dataset. Perhaps better to demonstrate with a different test set – say, just the species, genus family level, using something like Pseudomonas that has a lot of published genomes. Also show jaccard to emphasize how it gets progressively worse when you start comparing genomes that are different sizes? Or separate figure for this...?

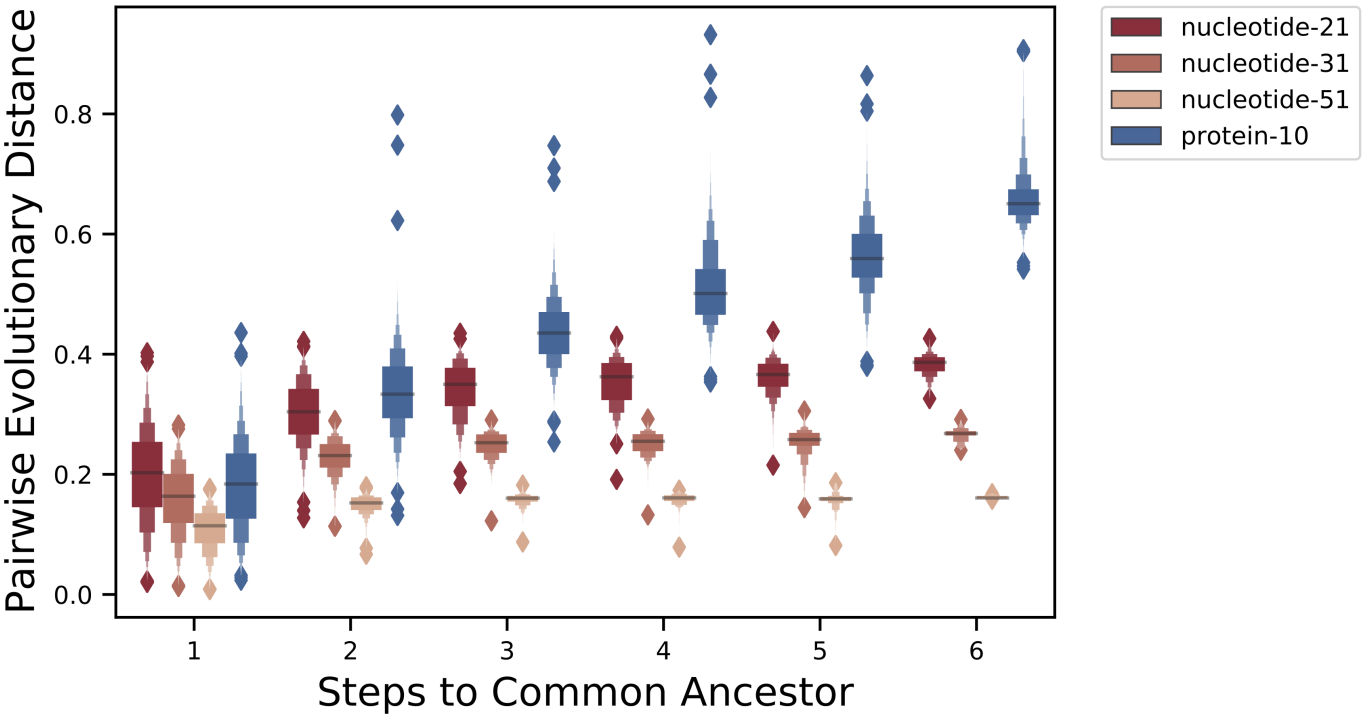

Figure 3: Containment-based ANI, AAI estimates, evolpaths

```
** core vs accessory distances **
ANI/AAI == really getting at _core_
```

Scaled Minhash distance estimation is robust to completeness

(unlike standard minhash https://drep.readthedocs.io/en/latest/choosing_parameters.html#importance-of-genome-completeness)

Protein containment searches enable Sensitive/fast/accurate taxonomic classification

(just containment, no ANI/AAI)

to do, classification: 0. fix thumper (refactor branch) –> working + tests 1. implement “leave one xx clade out classification check” –> instead of just ignoring exact matches, ignore any matches in same species/genus/family 2. prelim figure for tara classification vs GTDB-Tk vs BAT - classification of incomplete genomes - for “contaminated” genomes .. can we randomly add contigs from diff species, see the impact? Like classification still works until xx% contaminated with something present in our database? 3. ksize diffs for classification? k=7 vs k=10 vs k=11? - time, sensitivity, specificity 4. virus development! (could be separate paper)

While more protein k-mers are shared across genomes within the same genus (and different species), min-set-cov + LCA allows us to find/report the most similar genome.

benchmarking :: Leave one out classification

leave one clade out version? see CAT/BAT paper)

[protein vs dna]

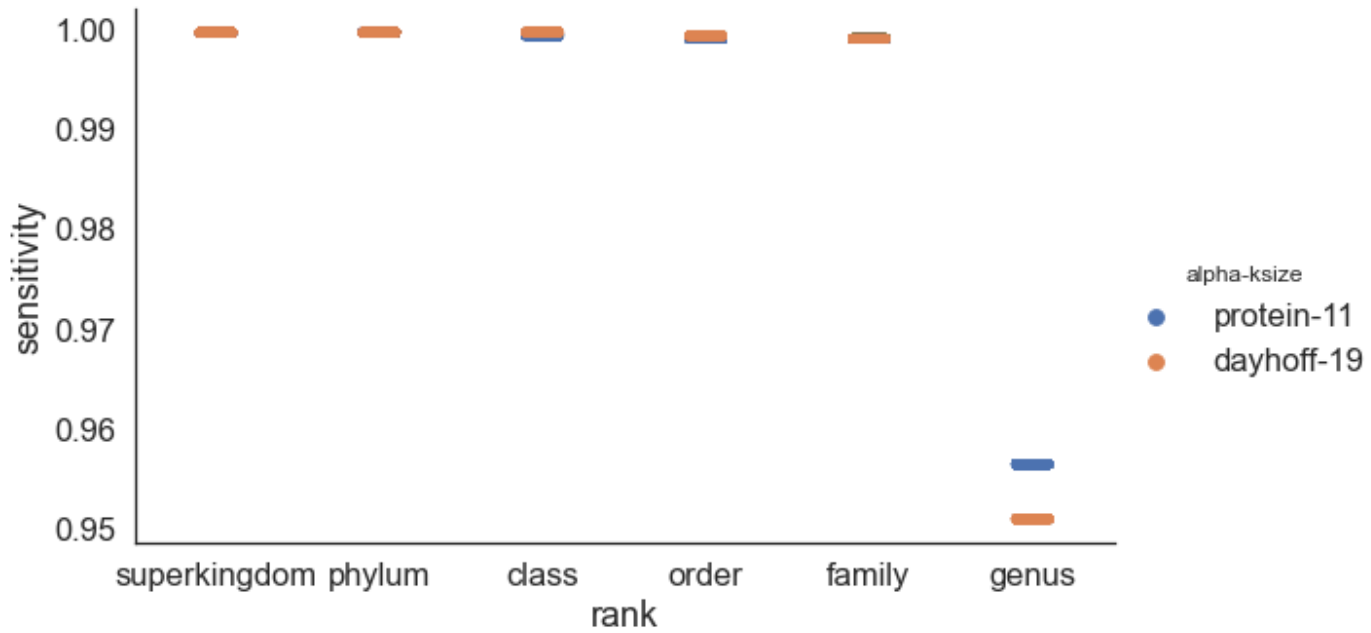

Figure 4: Protein classification sensitivity

include 6-frame translation works well for database search (sensitivity/specificity of Prodigal-translated vs 6-frame translated)

CAT/BAT paper [21] ("cat" = contig annotation, "bat" = bin annotation)

main point: more k-mers are shared = more k-mers available for matching

Classification of incomplete and contaminated genomes

virus classification

euk classification?? Too much.

Containment-AAI enable alignment-free phylogenomic reconstruction

evolpaths analysis

Containment searches enable similiarity estimation, especially between genomes of different lengths.

Max containment normalizes the shared content by the smaller of the two genomes

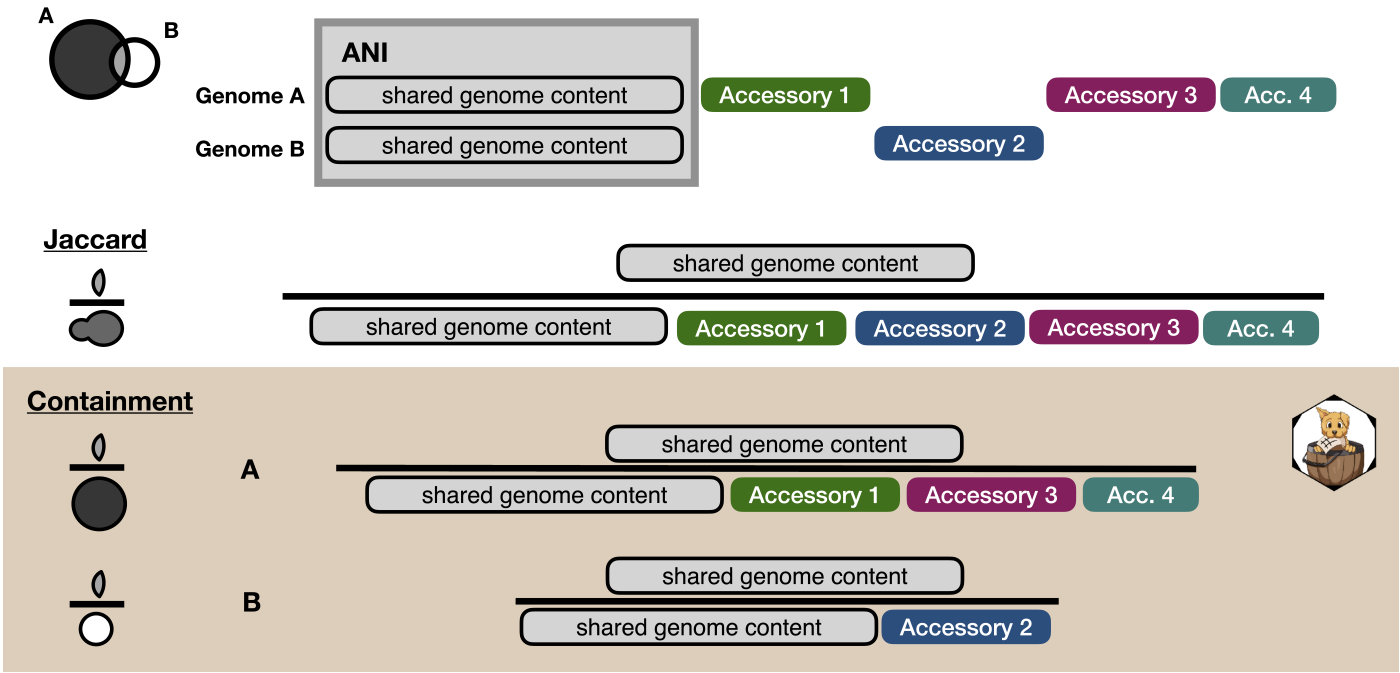

Figure 5: Max Containment to ANI and AAI. Containment calculation is guaranteed to be more similar to traditional calculation of Average Nucleotide Identity and Average Amino Acid Identity, which compared only the sections of genome that align. The shared k-mer content (containment numerator) can be thought of as the alignable sections of the genomes. The denominator of the Jaccard index is the alignable sections + the unalignable sections. The lower bound of the containment denominator will be the exact same as the numerator at 100% containment, where all k-mers are found within the comparison dataset. The upper bound will be the same as the Jaccard denominator, where all k-mers of the comparison dataset are found within the query dataset, and it is the query that contains any additional nonshared k-mers/unalignable sequence.

median AAI across GTDB?

alphabet and k-size selection for optimal distance estimation

- num shared k-mers at different ksizes
- e.g. k=7 much more common – do rankinfo on each database!

Comparison with other alignment-free methods (advantages, disadvantages, etc)

Alignment-based metrics are looking at the specific sequence variation of aligned regions, while k-mer based comparisons are comparing shared k-mers vs distinct k-mers. Since each nucleotide polymorphims generates mutated k-mers with an expected frequency, k-mer containment estimates can be used to accurately estimate both the Average Nucleotide Identity and Average Amino Acid Identity [1,19]

Discussion

Comparison with other alignment-free methods (advantages, disadvantages, etc)

Conclusions

Methods

Availability of data and materials

Competing Interests

The authors declare that they have no competing interests.

Funding

Authors' Contributions

Author	Contributions
NTP	Conceptualization; Methodology; Writing - Original Draft; Writing - Review and Editing; Visualization; Supervision; Funding Acquisition
...	...
...	...
...	...
CTB	Conceptualization; Methodology; Writing - Original Draft; Writing - Review and Editing; Visualization; Supervision; Funding Acquisition

Acknowledgements

Thank you to all the members and affiliates of the Lab for Data-Intensive Biology at UC Davis for providing valuable feedback on this manuscript. This manuscript was written using manubot [22] and is available in a GitHub repository [23].

References

1. **Mash: fast genome and metagenome distance estimation using MinHash**
   Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, Adam M. Phillippy
   Genome Biology (2016-06-20) https://doi.org/gfx74q
   DOI: 10.1186/s13059-016-0997-x · PMID: 27323842 · PMCID: PMC4915045

2. **Kraken: ultrafast metagenomic sequence classification using exact alignments**
   Derrick E Wood, Steven L Salzberg
   Genome Biology (2014) https://doi.org/gfkndk
   DOI: 10.1186/gb-2014-15-3-r46 · PMID: 24580807 · PMCID: PMC4053813

3. **How Much Do rRNA Gene Surveys Underestimate Extant Bacterial Diversity?**
   Luis M. Rodriguez-R, Juan C. Castro, Nikos C. Kyrpides, James R. Cole, James M. Tiedje, Konstantinos T. Konstantinidis
   Applied and Environmental Microbiology (2018-03-01) https://doi.org/ghtrdq
   DOI: 10.1128/aem.00014-18 · PMID: 29305502 · PMCID: PMC5835724

4. **Basic local alignment search tool.**
   SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman
   Journal of molecular biology (1990-10-05) https://www.ncbi.nlm.nih.gov/pubmed/2231712
   DOI: 10.1016/s0022-2836(05)80360-2 · PMID: 2231712

5. **Fast and sensitive protein alignment using DIAMOND**
   Benjamin Buchfink, Chao Xie, Daniel H Huson
   Nature Methods (2014-11-17) https://doi.org/gftzcs
   DOI: 10.1038/nmeth.3176 · PMID: 25402007

6. **Fast and sensitive taxonomic classification for metagenomics with Kaiju**
   Peter Menzel, Kim Lee Ng, Anders Krogh
   Nature Communications (2016-04-13) https://doi.org/f8h4b6
   DOI: 10.1038/ncomms11257 · PMID: 27071849 · PMCID: PMC4833860

7. **Mash Screen: high-throughput sequence containment estimation for genome discovery**
   Brian D. Ondov, Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, Adam M. Phillippy
   Genome Biology (2019-11-05) https://doi.org/ghtqmb
   DOI: 10.1186/s13059-019-1841-x · PMID: 31690338 · PMCID: PMC6833257

8. **Lightweight compositional analysis of metagenomes with sourmash gather**
   Luiz Irber, C. Titus Brown
   Manubot (2021-01-11) https://dib-lab.github.io/2020-paper-sourmash-gather/

9. **Improving MinHash via the containment index with applications to metagenomic analysis**
   David Koslicki, Hooman Zabeti
   Applied Mathematics and Computation (2019-08) https://doi.org/ghtqrv
   DOI: 10.1016/j.amc.2019.02.018

10. **Dashing: fast and accurate genomic distances with HyperLogLog**
Daniel N. Baker, Ben Langmead
Genome Biology (2019-12-04) https://doi.org/ggkmjc
DOI: 10.1186/s13059-019-1875-0 · PMID: 31801633 · PMCID: PMC6892282

11. **Metalign: efficient alignment-based metagenomic profiling via containment min hash**
Nathan LaPierre, Mohammed Alser, Eleazar Eskin, David Koslicki, Serghei Mangul
Genome Biology (2020-09-10) https://doi.org/ghtqrz
DOI: 10.1186/s13059-020-02159-0 · PMID: 32912225 · PMCID: PMC7488264

12. **Toward a More Robust Assessment of Intraspecies Diversity, Using Fewer Genetic Markers**
Konstantinos T. Konstantinidis, Alban Ramette, James M. Tiedje
Applied and Environmental Microbiology (2006-11) https://doi.org/dcmw9q
DOI: 10.1128/aem.01398-06 · PMID: 16980418 · PMCID: PMC1636164

13. **Uncultivated microbes in need of their own taxonomy**
Konstantinos T Konstantinidis, Ramon Rosselló-Móra, Rudolf Amann
The ISME Journal (2017-07-21) https://doi.org/gbprgw
DOI: 10.1038/ismej.2017.113 · PMID: 28731467 · PMCID: PMC5649169

14. **Shifting the genomic gold standard for the prokaryotic species definition**
Michael Richter, Ramon Rosselló-Móra
Proceedings of the National Academy of Sciences (2009-11-10) https://doi.org/dvchzz
DOI: 10.1073/pnas.0906412106 · PMID: 19855009 · PMCID: PMC2776425

15. **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries**
Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, Srinivas Aluru
Nature Communications (2018-11-30) https://doi.org/gfknmg
DOI: 10.1038/s41467-018-07641-9 · PMID: 30504855 · PMCID: PMC6269478

16. **Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries**
Matthew R. Olm, Alexander Crits-Christoph, Spencer Diamond, Adi Lavy, Paula B. Matheus Carnevali, Jillian F. Banfield
mSystems (2020-01-14) https://doi.org/ggwqh6
DOI: 10.1128/msystems.00731-19 · PMID: 31937678 · PMCID: PMC6967389

17. **There is no evidence of a universal genetic boundary among microbial species**
Connor S. Murray, Yingnan Gao, Martin Wu
Cold Spring Harbor Laboratory (2020-08-01) https://doi.org/ghtrdw
DOI: 10.1101/2020.07.27.223511

18. **Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead**
Konstantinos T Konstantinidis, James M Tiedje
Current Opinion in Microbiology (2007-10) https://doi.org/b2q3jd
DOI: 10.1016/j.mib.2007.08.006 · PMID: 17923431

19. **The statistics of *k*-mers from a sequence undergoing a simple mutation process without spurious matches**
Antonio Blanca, Robert S. Harris, David Koslicki, Paul Medvedev

Cold Spring Harbor Laboratory (2021-01-17) https://doi.org/fq3g
DOI: 10.1101/2021.01.15.426881

20. **A quick alternative method for resolving bacterial taxonomy using short identical DNA sequences in genomes or metagenomes**
Jesse Shapiro
Peer Community In Genomics (2020-09-24) https://doi.org/ghvgd3
DOI: 10.24072/pci.genomics.100001

21. **Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT**
F. A. Bastiaan von Meijenfeldt, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, Bas E. Dutilh
Genome Biology (2019-10-22) https://doi.org/ggfm6r
DOI: 10.1186/s13059-019-1817-x · PMID: 31640809 · PMCID: PMC6805573

22. **Open collaborative writing with Manubot**
Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653

23. https://github.com/bluegenes/2021-ani-paper