Lab for Data Intensive Biology

UCDAVIS VETERINARY MEDICINE

A new method for taxonomic classification using MinHash and Sequence Bloom Trees

L.C. Irber Jr., P.T. Brooks, T.E. Reiter, C. Titus Brown

Lab for Data Intensive Biology, School of Veterinary Medicine, University of California Davis

@luizirber @brooksph @reitertaylor @ctb

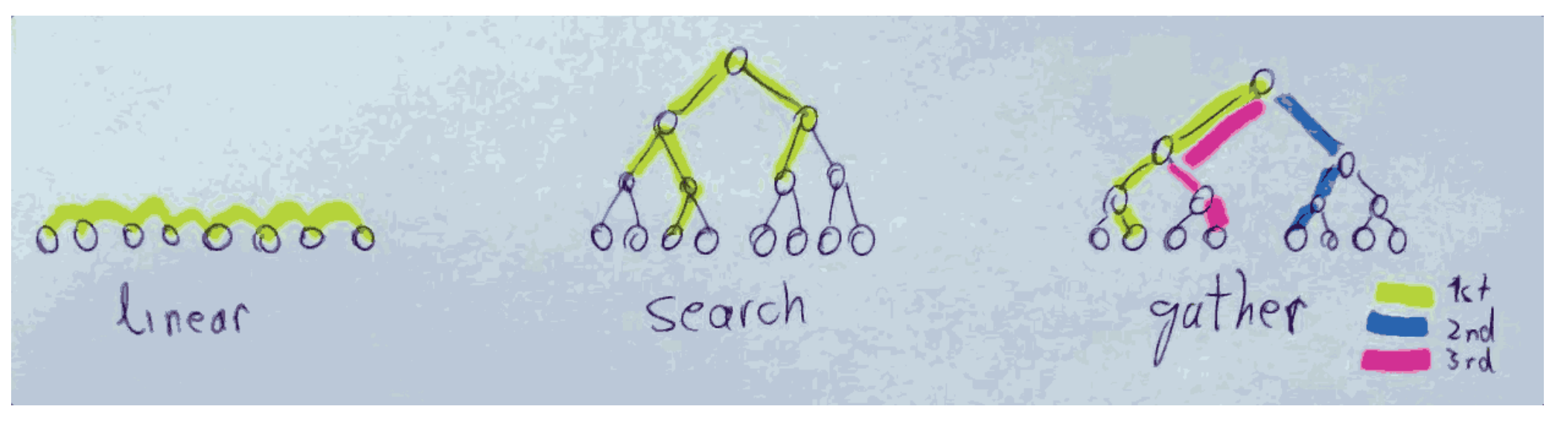
luizirber/2018-sms

GORDON AND BETTY MOORE FOUNDATION

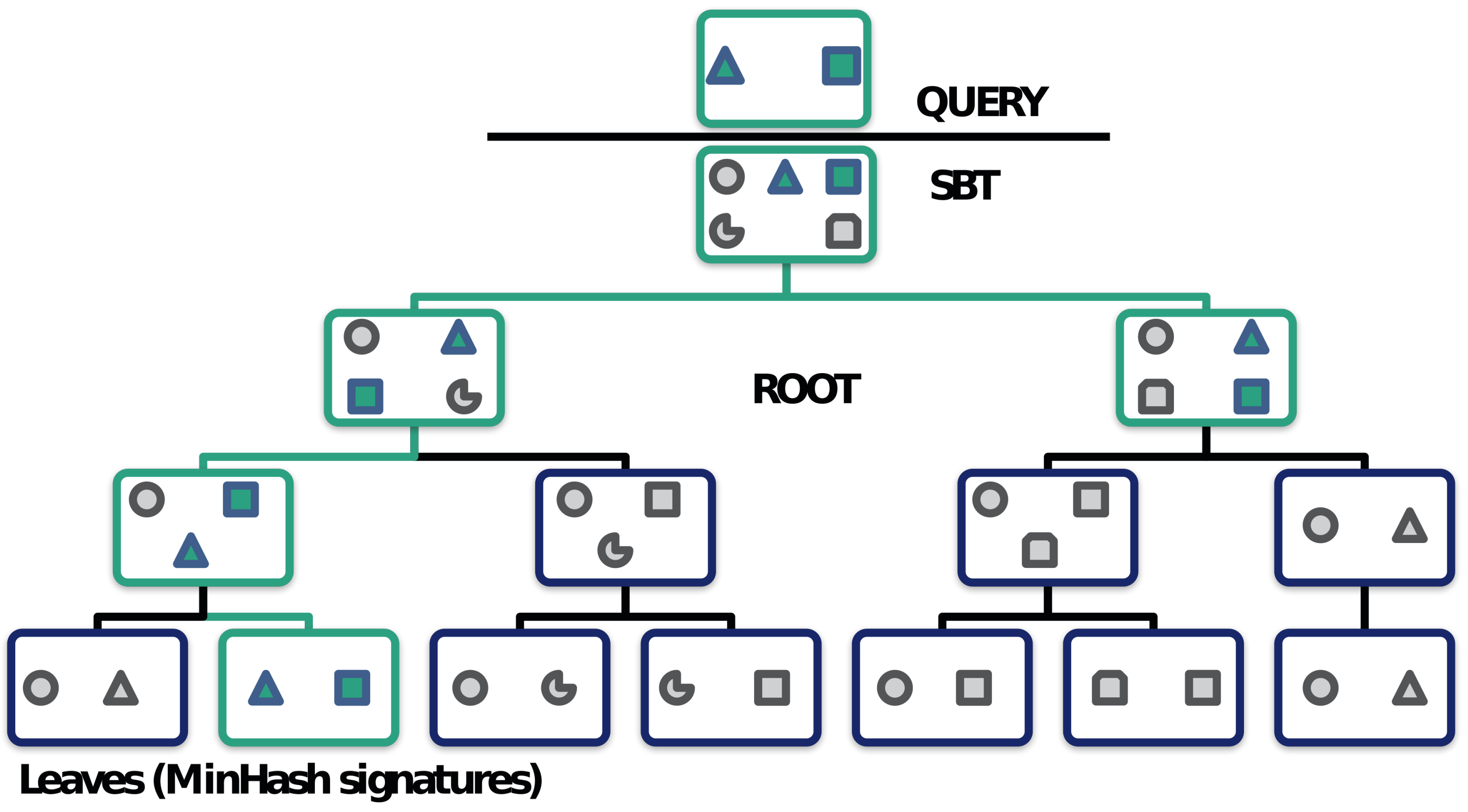
Introduction

MinHash [Broder, 1997] is a technique for **estimating the similarity** of two or more datasets. Expanding on the work pioneered by **Mash** [Ondov et al, 2016] and extended in our library **sourmash** [Brown and Irber, 2016], we calculated signatures for **microbial GenBank** and **RefSeq** and prepared **search indexes** using **Sequence Bloom Trees** [Solomon and Kingsford, 2016] adapted for searching MinHash signatures.

sourmash **gather** is a new method for **taxonomic classification** using the **same search indexes** we already use for searching similar datasets in public databases but with a **different search strategy**: instead of **looking for all** datasets above a similarity threshold, gather does a **greedy search for the best match**, report it and then **remove the match** from the **original query**. This process is repeated while there are enough items in the query to find matches **above a defined threshold**.

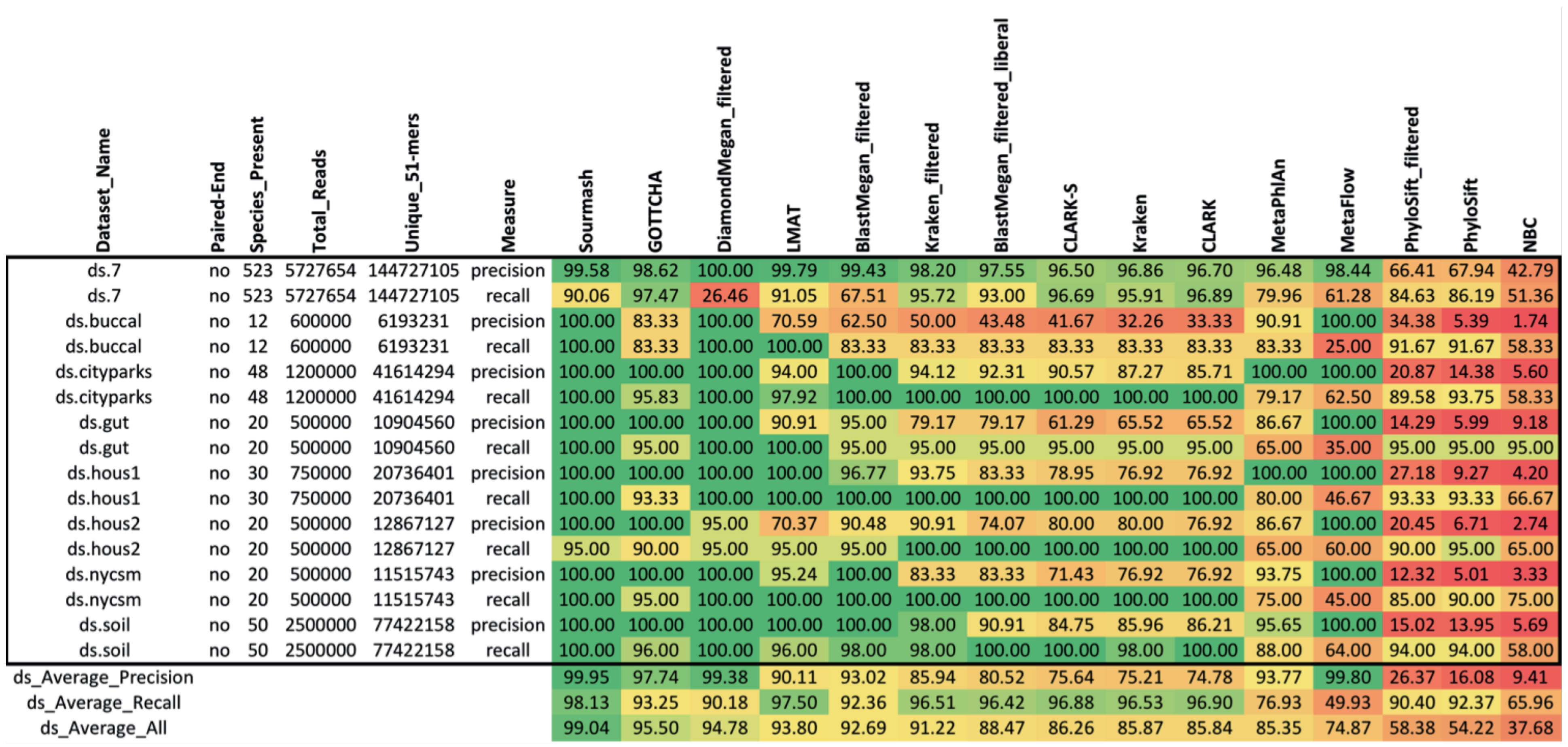


Sequence Bloom Trees



Using Sequence Bloom Trees and MinHash to find matches. SBTs were constructed using MinHash signatures generated with sourmash compute. Nodes are bloom filters containing the union of signatures and leaves are signatures. All MinHash signatures contained in the tree are present in the root. Matches are determined by comparing the query MinHash signatures to the nodes until the best match is found.

The NIST-IMMSA benchmark



Dataset_Name	Paired-End	Species_Present	Total_Reads	Unique_51-mers	Measure	Sourmash	GOTTCHA	DiamondMegan_filtered	LMAT	BlastMegan_filtered	Kraken_filtered	BlastMegan_filtered_liberal	CLARK-S	Kraken	CLARK	MetaPhlAn	MetaFlow	PhyloSift_filtered	PhyloSift	NBC
ds.7	no	523	5727654	144727105	precision	99.58	98.62	100.00	99.79	99.43	98.20	97.55	96.50	96.86	96.70	96.48	98.44	66.41	67.94	42.79
ds.7	no	523	5727654	144727105	recall	90.06	97.47	26.46	91.05	67.51	95.72	93.00	96.69	95.91	96.89	79.96	61.28	84.63	86.19	51.36
ds.buccal	no	12	600000	6193231	precision	100.00	83.33	100.00	70.59	62.50	50.00	43.48	41.67	32.26	33.33	90.91	100.00	34.38	5.39	1.74
ds.buccal	no	12	600000	6193231	recall	100.00	83.33	100.00	100.00	83.33	83.33	83.33	83.33	83.33	83.33	83.33	25.00	91.67	91.67	58.33
ds.cityparks	no	48	1200000	41614294	precision	100.00	100.00	100.00	94.00	100.00	94.12	92.31	90.57	87.27	85.71	100.00	100.00	20.87	14.38	5.60
ds.cityparks	no	48	1200000	41614294	recall	100.00	95.83	100.00	97.92	100.00	100.00	100.00	100.00	100.00	100.00	79.17	62.50	89.58	93.75	58.33
ds.gut	no	20	500000	10904560	precision	100.00	100.00	100.00	90.91	95.00	79.17	79.17	61.29	65.52	65.52	86.67	100.00	14.29	5.99	9.18
ds.gut	no	20	500000	10904560	recall	100.00	95.00	100.00	100.00	95.00	95.00	95.00	95.00	95.00	95.00	65.00	35.00	95.00	95.00	95.00
ds.hous1	no	30	750000	20736401	precision	100.00	100.00	100.00	100.00	96.77	93.75	83.33	78.95	76.92	76.92	100.00	100.00	27.18	9.27	4.20
ds.hous1	no	30	750000	20736401	recall	100.00	93.33	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	80.00	46.67	93.33	93.33	66.67
ds.hous2	no	20	500000	12867127	precision	100.00	100.00	95.00	70.37	90.48	90.91	74.07	80.00	80.00	76.92	86.67	100.00	20.45	6.71	2.74
ds.hous2	no	20	500000	12867127	recall	95.00	90.00	95.00	95.00	95.00	100.00	100.00	100.00	100.00	100.00	65.00	60.00	90.00	95.00	65.00
ds.nycsm	no	20	500000	11515743	precision	100.00	100.00	100.00	95.24	100.00	83.33	83.33	71.43	76.92	76.92	93.75	100.00	12.32	5.01	3.33
ds.nycsm	no	20	500000	11515743	recall	100.00	95.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	75.00	45.00	85.00	90.00	75.00
ds.soil	no	50	2500000	77422158	precision	100.00	100.00	100.00	100.00	100.00	98.00	90.91	84.75	85.96	86.21	95.65	100.00	15.02	13.95	5.69
ds.soil	no	50	2500000	77422158	recall	100.00	96.00	100.00	96.00	98.00	98.00	100.00	100.00	98.00	100.00	88.00	64.00	94.00	94.00	58.00
ds_Average_Precision						99.95	97.74	99.38	90.11	93.02	85.94	80.52	75.64	75.21	74.78	93.77	99.80	26.37	16.08	9.41
ds_Average_Recall						98.13	93.25	90.18	97.50	92.36	96.51	96.42	96.88	96.53	96.90	76.93	49.93	90.40	92.37	65.96
ds_Average_All						99.04	95.50	94.78	93.80	92.69	91.22	88.47	86.26	85.87	85.84	85.35	74.87	58.38	54.22	37.68

McIntyre et al. 2017 Additional File 4: Table S3, modified with an extra column for sourmash gather.

The **NIST-IMMSA benchmark** [McIntyre et al, 2017] compares **metagenomic classifiers** and contains both **biological** and **simulated** metagenomic **datasets** where the species composition (the **truth set**) is **known**. The original publication compares and evaluates **11 tools** using a variety of classification approaches (**k-mer composition**, **alignment**, **markers**). We evaluated sourmash **gather** using the **simulated datasets** and found that it presents **better precision** and **recall** than the tools previously benchmarked.

$$\text{precision} = \frac{\text{species identified correctly}}{\text{species identified}}$$

$$\text{recall} = \frac{\text{species identified correctly}}{\text{species in truth set}}$$

Future Work

A **dual** approach to the Sequence Bloom Tree is to build a **reverse index**, a **mapping** of **hashed k-mers** to **signatures**. This trades **memory usage** for **speed** and is similar to how Kraken [Wood and Salzberg, 2014] performs the LCA assignment.

Taxonomic classifiers are **sensitive to changes in taxonomy** (both from **new species** being added as well from **reassignments**). Many tools provide **prepared databases** but it is not easy (or possible) to **update** them. sourmash indexes are **online** and can be updated **without complete recalculation**, and we are working in **automating the process** of **downloading** new changes from public databases and **publishing** updated indexes. We are also working on more comprehensive benchmarks and improving the computational performance of the method.

gather is currently available in prereleases of sourmash 2.0. We are working on releasing 2.0 soon, but it can already be installed from **PyPI** and **bioconda**:

```
$ pip install --pre sourmash
$ conda install -c bioconda -c conda-forge sourmash
```

References

Broder, Andrei Z. 1997. **“On the Resemblance and Containment of Documents.”** In Compression and Complexity of Sequences 1997. Proceedings, 21–29. IEEE. doi.org/10.1109/SEQUEN.1997.666900.

McIntyre, Alexa B. R., Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, et al. 2017. **“Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers.”** Genome Biology 18: 182. doi.org/10.1186/s13059-017-1299-7.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. **“Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.”** Genome Biology 17: 132. doi:10.1186/s13059-016-0997-x.

Solomon, Brad, and Carl Kingsford. 2016. **“Fast Search of Thousands of Short-Read Sequencing Experiments.”** Nature Biotechnology 34 (3): 300–302. doi.org/10.1038/nbt.3442.

Titus Brown, C., and Luiz Irber. 2016. **“sourmash: A Library for MinHash Sketching of DNA.”** The Journal of Open Source Software 1 (5). doi:10.21105/joss.00027.

Wood, Derrick E., and Steven L. Salzberg. 2014. **“Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments.”** Genome Biology 15 (3): R46. doi.org/10.1186/gb-2014-15-3-r46.