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, UC Davis
Stanford Microbiome Symposium - 2018/09/24

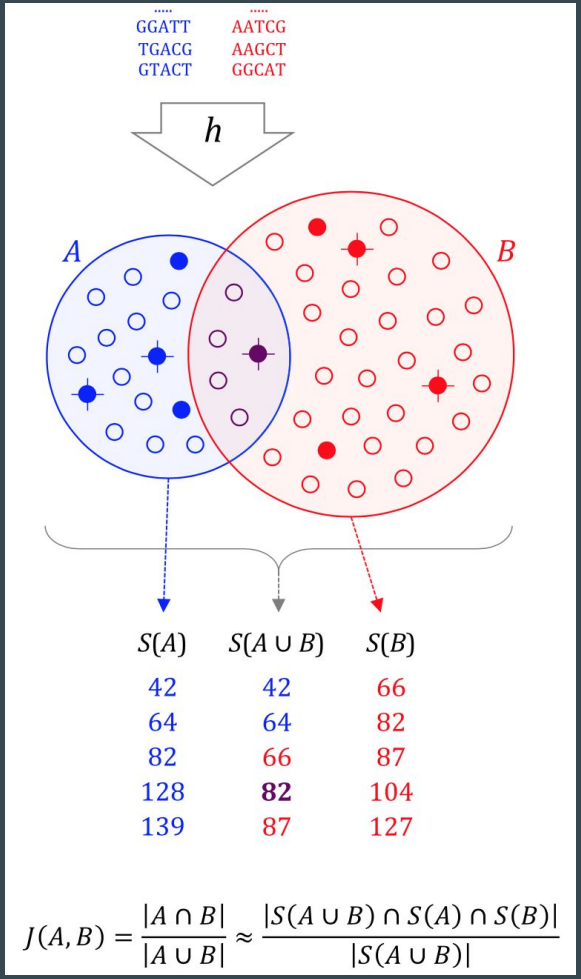
Taxonomic classification

- What are the species present in a sample?
- And what are their abundances?
- Current methods
  - K-mer composition (Kraken)
  - Alignment (Diamond-MEGAN)
  - Markers (PhyloSift)

MinHash

Mash (Ondov et al, 2016), sourmash (Brown and Irber, 2016)

- How similar are two datasets?
- A mix of k-mer composition and markers
  - Uses hashed k-mers...
  - ... but only a systematically consistent subset
- sourmash: a library for MinHash sketching
- Scaled MinHash: accounts for genomic complexity
  - Bacteria: ~10k hashes
  - Metagenome: ~10M hashes
  - (not fixed size anymore!)
- From large dataset to small signature

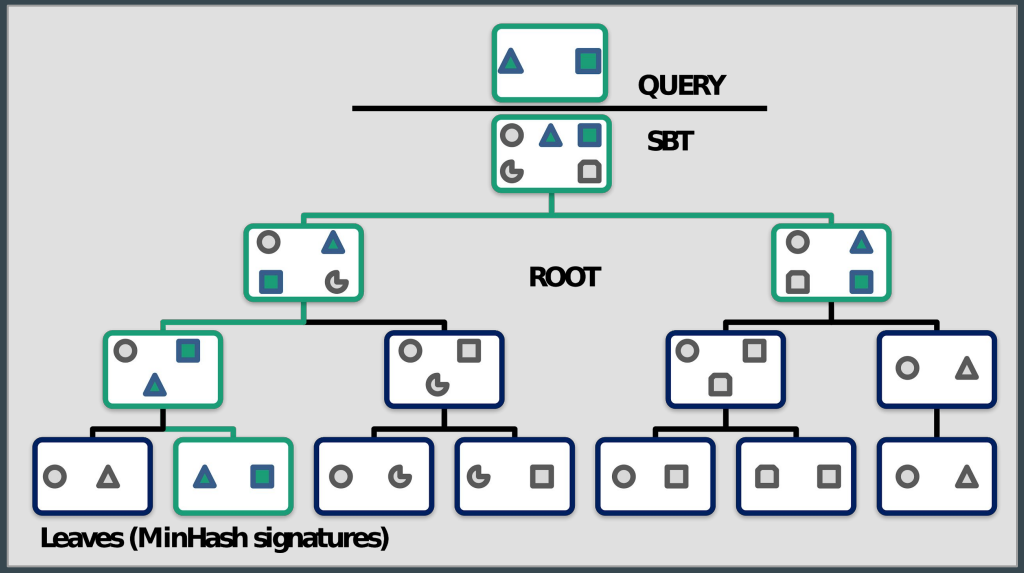


Ondov et al., 2016

Sequence Bloom Tree

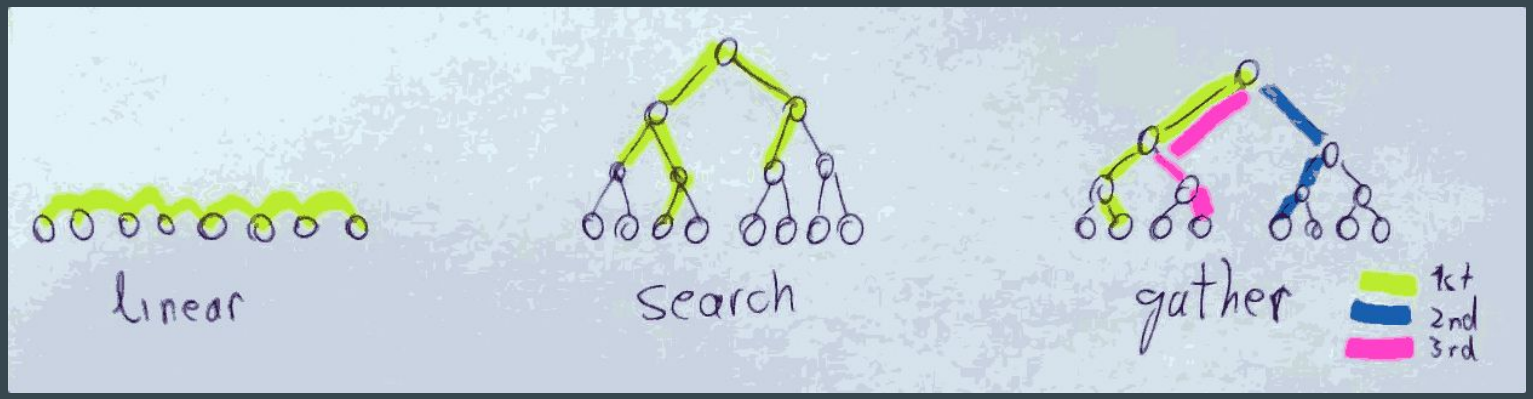
Solomon and Kingsford, 2016

- Hierarchical index
- Leaf nodes
  - Original: all k-mers from a dataset (in a Bloom Filter)
  - sourmash: minimum hash values (in a MinHash)
- Internal nodes: union of all elements in children
- Find all datasets above a similarity threshold with query
- From **one** to **many signatures**



sourmash gather

- Same search index
- Distinct search method
  - search: breadth-first with truncation
  - gather: multiple best-first searches
- When a match is found:
  - Save metadata for summary
  - Remove hashes from query
  - Find next best match
  - ... until no more hashes/threshold reached for query



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

sourmash gather: precision and recall

McIntyre et al. 2017 Additional File 4: Table S3, modified with an extra column for sourmash gather
(Phillip Brooks and Krista Ternus)

More info

- sourmash manual: sourmash.rtfd.org
- Code repo: github.com/dib-lab/sourmash
- This talk and poster: github.com/luizirber/2018-sms
- Poster session!
  - Also check Taylor's poster on spacegraphcats =]