sourmash

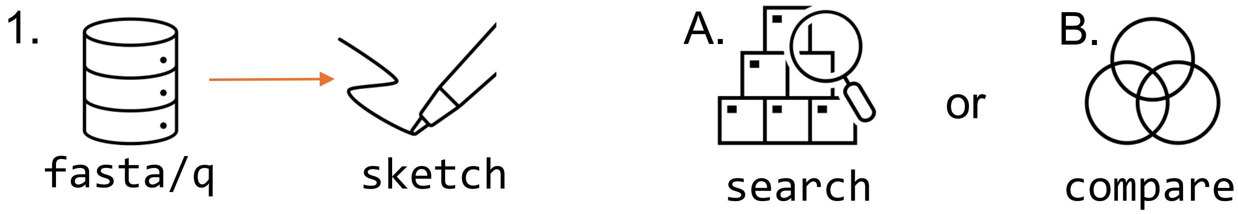
the cheat sheet

A quick lookup to sketching, comparison, and searching of metagenomic samples.

In partnership with Penn State

Learn more: https://sourmash.readthedocs.io/

Overview



commands

`sourmash sketch`	produce a signature file for the following moltypes: `dna`, `protein`, `translate`
`sourmash sig`	obtain information on signatures produced, combine with `describe` to get signature information or `manifest` to obtain md5 identifiers of sketches
`sourmash search`	report the similarity percentages of a query signature or `--md5` identifier in a database signature
`sourmash compare`	report similarity indexes between two signatures

parameters

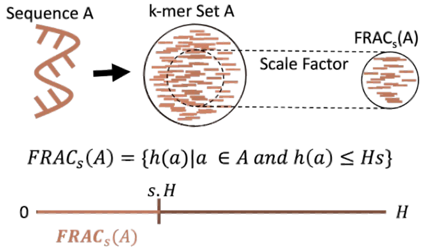
k	identify size of k-mers (sequence subset of size k), required with `sourmash sketch`, `compare`
scaled	Identify the scale factor, reduces original k-mer set, to keep all k-mers use scaled=1, used with `sourmash sketch`

Note: required to use `-p` before parameters

options

`--containment`	utilized in either `sourmash search` or `compare` commands to report containment index
`--ani`	approximate average nucleotide identity from `--containment`
`--dna`	identify moltype of signature input as `--dna`, `--protein` is also available
`--singleton`	`sourmash sketch` a signature file containing a sketch for each sequence within a fasta file
`--o`	output filename for signature when using `sourmash sketch` or similarity estimates matrix filename for `sourmash compare`

sketch

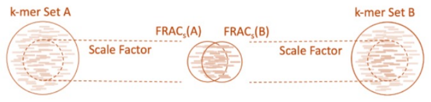


$$FRAC_s(A) = \{h(a) | a \in A \; and \; h(a) \leq Hs\}$$

- The FracMinHash function is used to produce a scaled down representation of the original sequencing data
- This approach is great for data of differing sizes

```
sourmash sketch dna genome.fasta -p k=31,scaled=500
```

compare



$$C_{frac}(A,B) \coloneqq \frac{|\boldsymbol{FRAC_s}(A) \cap \boldsymbol{FRAC_s}(B)|}{|\boldsymbol{FRAC_s}(A)|(1-(1-s)^{|A|})} \mathbb{1}_{|\boldsymbol{FRAC_s}(A)|>0}$$

Produce a matrix that reports the similarity index, containment, between two sequencing datasets

```
sourmash compare sample_1.sig sample_2.sig --containment
```

search sig query

Search for highly similar sequences between genomes or genomes between two metagenomic samples

```
sourmash search sample_1.sig sample_2.sig --containment
```

search md5 query

Each sketch is assigned an md5 identifier. Use the `sourmash search --md5` option to report similarity of a sketch of interest within a signature file:

1. Produce a manifest file

```
sourmash sig manifest sample_1.sig -o MANIFEST.csv
```

2. Open MANIFEST file and choose an md5 identifier

		md5			
		X			
		Y			

3. `sourmash search` command is modified to search similarity of an md5 in a database

```
sourmash search --md5 X sample_2.sig --containment
```