Fast Approximations of Frequent k-Mers and Applications

Fabio Vandin

Dept. of Information Engineering
University of Padova

fabio.vandin@unipd.it
www.dei.unipd.it/~ vandinfa/

DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

Motivation

Dataset of Reads

$\mathcal{D} =$

ATACCGATG

CCGTCATG

AGAAATGCC

TCAATCAGC

ATGTGATGC

...

Goal: Compute counts of all *k*-mers

(i.e. k=3)

(ATG , 4)

(TCA , 3)

(GAA , 1)

...

Applications:

1. metagenomic reads classification
2. error correction
3. repeat detection
4. genome comparison
5. ...

Challenges:

1. size of datasets
2. $\mathcal{O}(4^k)$ distinct k-mers

Motivation

Many efficient approaches for exact or approximate counting are available:

`Jellyfish` (Marçais et al., 2011), `DSK` (Rizk et al, 2013), `KMC` (Kokot et al, 2017), `Squeakr` (Pandey et al, 2017), `KmerStream` (Melsted et al, 2014), `BFCounter` (Melsted and Pritchard, 2011) `khmer` (Zhang et al, 2014), `Kmerlight` (Sivadasan et al, 2016), `ntCard` (Mohamadi et al, 2017), `KmerGenie` (Chikhi et al, 2013), `KAnalyze` (Audano and Vannberg, 2017), `Turtle` (Roy et al., 2014),...

Based on efficient and succint data structures for storing distinct k-mers, parallelism, ...

Common to all: analyse *all* data, obtain counts of *all* k-mers

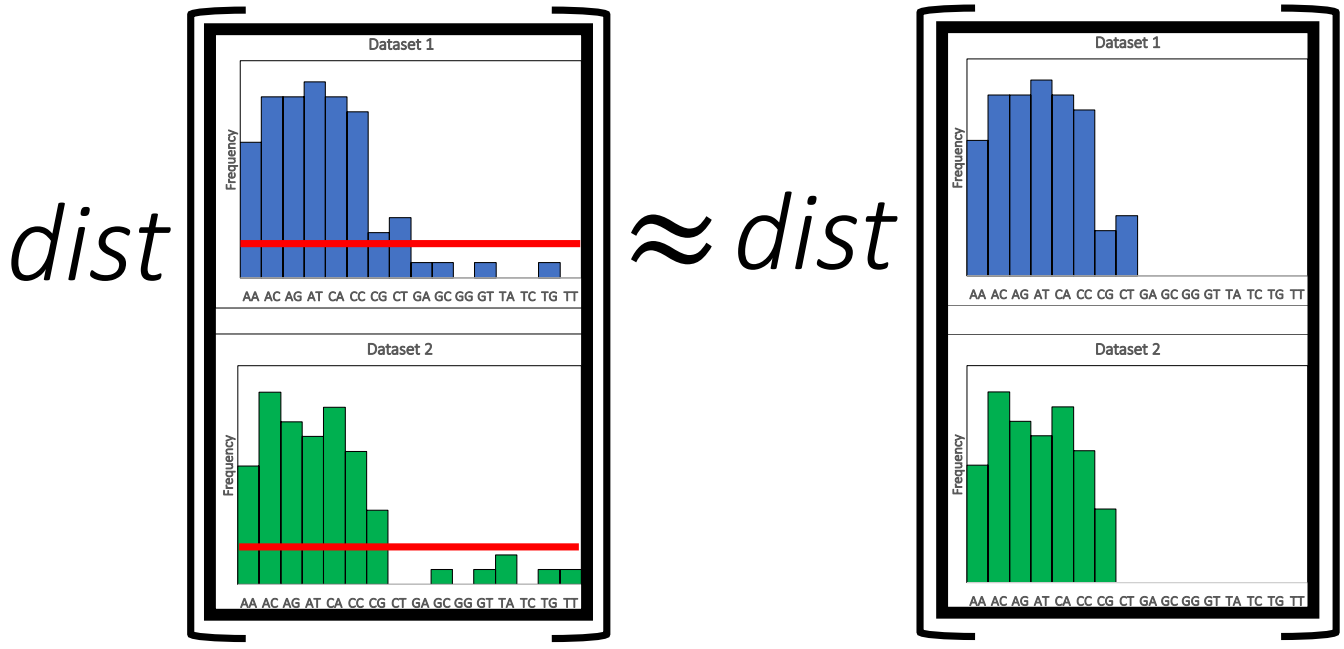
Is this really needed?

Abundance-based Distances between Metagenomic Datasets

BC distance between k-mers $\mathcal{S}_1$ of $\mathcal{D}_1$ and k-mers $\mathcal{S}_2$ of $\mathcal{D}_2$:

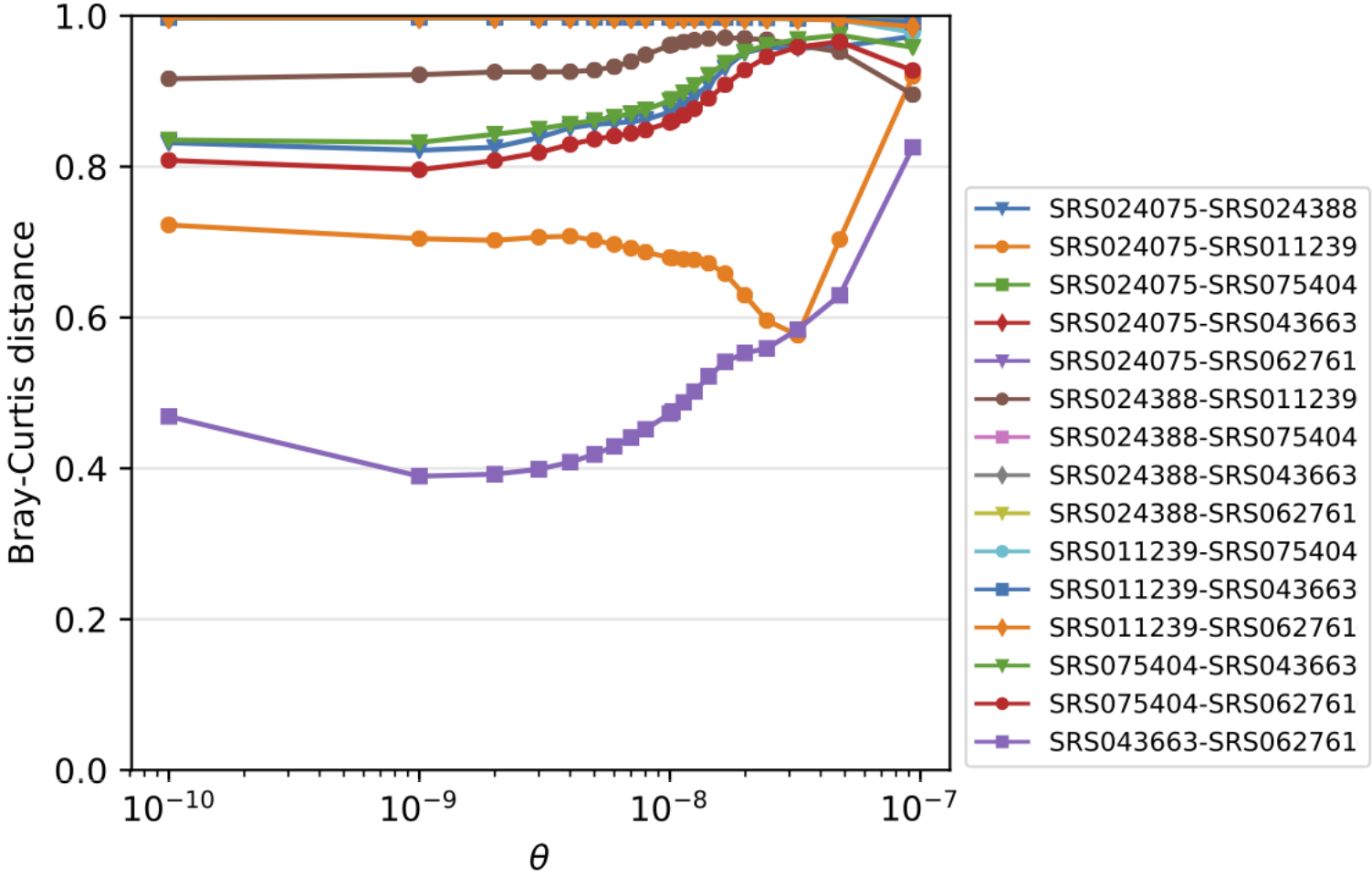
$$BC(\mathcal{D}_1, \mathcal{D}_2, \mathcal{S}_1, \mathcal{S}_2) = 1 - 2\frac{\sum_{K \in \mathcal{S}_1 \cap \mathcal{S}_2} \min\{o_{\mathcal{D}_1}(K), o_{\mathcal{D}_2}(K)\}}{\sum_{K \in \mathcal{S}_1} o_{\mathcal{D}_1}(K) + \sum_{K \in \mathcal{S}_2} o_{\mathcal{D}_2}(K)}$$

Do we really need to get the counts of *all* k-mers?



Abundance-based Distances between Metagenomic Datasets

What about computing BC distance between k-mers of $\mathcal{D}_1$ and k-mers of $\mathcal{D}_2$ considering *only* k-mers with *frequency* $\geq \theta$?



Our contributions

Two algorithms to approximate **frequent** k-mers:

- SAKEIMA (酒今) : Sampling Algorithm for K-mErs approxIMAtion [Pellegrina, Pizzi, V., RECOMB 2019, JCB 2020]
- SPRISS: SamPling Reads algorIthm to eStimate frequent k-merS [Santoro*,Pellegrina*, V., RECOMB 2021]

$\rightarrow$ process only a **random sample** of the dataset

$\rightarrow$ provide **rigorous approximations** *(statistical learning theory)*

$\rightarrow$ easily adaptable to any existing k-mer counting algorithm

Outline

1. **Problem definition**
2. Naïve sampling approach
3. SAKEIMA (酒今)
4. SPRISS

Preliminaries

$$\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} \ , \ \sigma = |\Sigma| = 4$$

$$\mathcal{D} = \{\mathtt{ACTACTACGT},$$
$$\mathtt{CCGTAGTGT},$$
$$\mathtt{AGAAATGCC},$$
$$\mathtt{TCAATCAGC},$$
$$\mathtt{ATGTGATGC},$$
$$\ldots \quad \}$$

For $k = 5$: $\mathcal{P}_{\mathcal{D},k} = \{\mathtt{ACTAC},$
$\mathtt{CTACT},$
$\mathtt{TACTA},$
$\mathtt{ACTAC},$
$\ldots \quad \}$

$$t_{\mathcal{D},k} = |\mathcal{P}_{\mathcal{D},k}| = \ \# \ k\text{-mers in } \mathcal{D}$$

Goal: $o_{\mathcal{D}}(K) = \ \#$ occurences of K in $\mathcal{P}_{\mathcal{D},k}$

$$f_{\mathcal{D}}(K) = o_{\mathcal{D}}(K)/t_{\mathcal{D},k}$$

Preliminaries

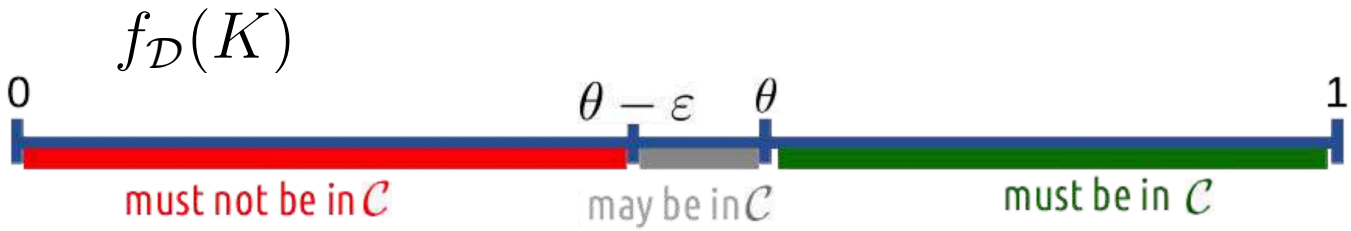
Definition: Set $FK(\mathcal{D}, k, \theta)$ of *frequent k-mers in $\mathcal{D}$ w.r.t θ*:

$$FK(\mathcal{D}, k, \theta) = \{(K, f_{\mathcal{D}}(K)) : f_{\mathcal{D}}(K) \geq \theta\}$$

Approximation of $FK(\mathcal{D}, k, \theta)$

Definition: For $\varepsilon < \theta$, an *ε-approximation* of $FK(\mathcal{D}, k, \theta)$ is a collection $C = \{(K, f_K) : f_K \in (0, 1]\}$ s.t.:

- Contains all K with $f_{\mathcal{D}}(K) \geq \theta$
- Contains no K with $f_{\mathcal{D}}(K) \leq \theta - \varepsilon$
- $|f_{\mathcal{D}}(K) - f_K| \leq \varepsilon/2, \forall K \in C$.



Fast computation? Random sampling $\rightarrow$ Approximation with probability $\geq 1 - \delta$

Outline

1. Problem definition
2. **Naïve sampling approach**
3. SAKEIMA (酒今)
4. SPRISS

Naïve sampling approach

Random sample $\mathcal{P}$ of $\mathcal{P}_{\mathcal{D},k}$ and compute $FK(\mathcal{P}, k, \theta - \varepsilon/2)$

$$\mathcal{P}_{\mathcal{D},k} = \{\texttt{AATAC}, \texttt{ATACC}, \texttt{TACCG}, \texttt{ACCGA}, \texttt{AATAC}, \ldots\} \rightarrow \mathcal{P} = \{\texttt{AATAC}, \texttt{ATACC}, \texttt{ACCGA}, \ldots\}$$

How many samples do we need?

Naïve sampling approach

Theorem: $FK(\mathcal{P}, k, \theta - \varepsilon/2)$ is an ε-approximation of $FK(\mathcal{D}, k, \theta)$ with probability $\geq 1 - \delta$ if

$$m \geq \frac{2}{\varepsilon^2}\left(\ln(2\sigma^k) + \ln\left(\frac{1}{\delta}\right)\right)$$

Improved bound:

$$m \geq \frac{2}{\varepsilon^2}\left(1 + \ln\left(\frac{1}{\delta}\right)\right)$$

Naïve sampling approach

Theorem: $FK(\mathcal{P}, k, \theta - \varepsilon/2)$ is an ε-approximation of $FK(\mathcal{D}, k, \theta)$ with probability $\geq 1 - \delta$ if

$$m \geq \frac{2}{\varepsilon^2}\left(\ln(2\sigma^k) + \ln\left(\frac{1}{\delta}\right)\right)$$

Improved bound:

$$m \geq \frac{2}{\varepsilon^2}\left(1 + \ln\left(\frac{1}{\delta}\right)\right)$$

Naïve sampling approach

Theorem: $FK(\mathcal{P}, k, \theta - \varepsilon/2)$ is an ε-approximation of $FK(\mathcal{D}, k, \theta)$ with probability $\geq 1 - \delta$ if

$$m \geq \frac{2}{\varepsilon^2}\left(\ln(2\sigma^k) + \ln\left(\frac{1}{\delta}\right)\right)$$

Improved bound:

$$m \geq \frac{2}{\varepsilon^2}\left(1 + \ln\left(\frac{1}{\delta}\right)\right)$$

$\#$ of k-mers to process $> \frac{1}{\varepsilon^2}$! Not practical since $\varepsilon < \theta$.

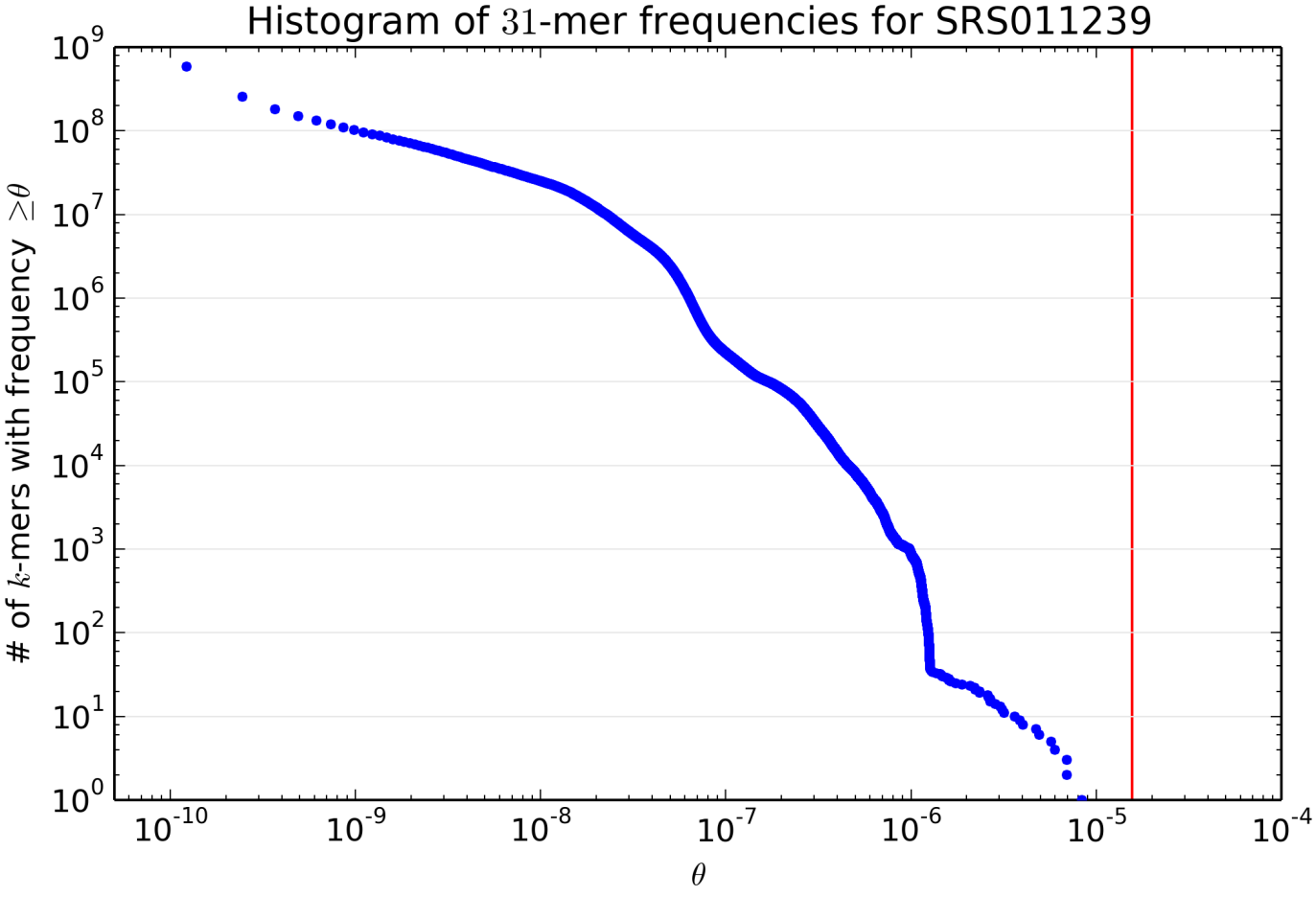
Naïve sampling approach

Th

ε-ap

≥ 1

Imp



of k-mers to process $> \frac{1}{\varepsilon^2}$! Not practical since $\varepsilon < \theta$.

Outline

1. Problem definition
2. Naïve sampling approach
3. SAKEIMA (酒今)
4. SPRISS

SAKEIMA (酒今)

Random sample $\mathcal{P}_\ell$: m samples (=*bags*) each one containing ℓ k-mers.

$$\mathcal{P}_{\mathcal{D},k} = \{\texttt{AATAC},\ \texttt{ATACC},\ \texttt{TACCG},\ \texttt{ACCGA},\ \texttt{AATAC},\ \texttt{GGCCA},\ \ldots\ \}$$

$$\rightarrow$$

With $\ell = 2$:

$$\mathcal{P}_\ell = \Big\{\{\texttt{AATAC},\ \texttt{GGCCA}\},\ \{\texttt{ATACC},\ \texttt{ACCGA}\},\ \ldots\ \Big\}$$

$\hat{f}_{P_\ell}(K) :=$ fract. of bags of $\mathcal{P}_\ell$ with at least one K

$\hat{f}_{P_\ell}(K)/\ell =$ *biased estimator* of $f_{\mathcal{D}}(K)$:

$$\mathbb{E}\left[\hat{f}_{P_\ell}(K)/\ell\right] = 1 - (1 - \ell f_{\mathcal{D}}(K))^{1/\ell} \approx f_{\mathcal{D}}(K)$$

SAKEIMA (酒今)

Proposition: Let $\ell \geq 1$ and $\mathcal{P}_\ell$ be a sample of m bags of size ℓ of $\mathcal{P}_{\mathcal{D},k}$ with

$$m \geq \frac{2}{(\ell\varepsilon)^2}\left(\lfloor \log_2(2\ell) \rfloor + \ln\left(\frac{1}{\delta}\right)\right).$$

Then, with probability at least $1-\delta$, the k-mers with frequency *in the sample* $\geq \theta - \varepsilon/2$ contain:

- All K with $f_{\mathcal{D}}(K) \geq \theta' \approx \theta$
- No K with $f_{\mathcal{D}}(K) \leq \theta - \varepsilon$

Note: number of k-mers to process: $\mathcal{O}(m\ell) = \mathcal{O}\left(\frac{\log(\ell)}{\ell\varepsilon^2}\right)$
$\rightarrow$ by properly setting ℓ we obtain practical sample sizes!

Proof: based on VC-dimension of bags of k-mers.

Can we do better?

SAKEIMA (酒今) is great, but still requires to stream over all the reads in the dataset

What about sampling *reads* instead of k-mers?

Challenge: sampling reads introduces correlations among sampled k-mers

Outline

1. Problem definition
2. Naïve sampling approach
3. SAKEIMA (酒今)
4. SPRISS

SPRISS

SamPling Reads algorIthm to eStimate frequent k-merS

Naïve sampling approach: requires more reads than in the dataset!

Idea: sample bags of reads, each bag with ℓ reads

Algorithm 1: SPRISS$(\mathcal{D}, k, \theta, \delta, \varepsilon, \ell)$

Data: $\mathcal{D}$, k, $\theta \in (0,1]$, $\delta \in (0,1)$, $\varepsilon \in (0,\theta)$, integer $\ell \geq 1$

Result: Approximation A of $FK(\mathcal{D}, k, \theta)$ with probability at least $1-\delta$

$m \leftarrow \left\lceil \frac{2}{\varepsilon^2} \left(\frac{1}{\ell \ell_{\mathcal{D},k}} \right)^2 \left(\lfloor \log_2 \min(2\ell\ell_{\max,\mathcal{D},k}, \sigma^k) \rfloor + \ln\left(\frac{1}{\delta}\right) \right) \right\rceil$;

$S \leftarrow$ sample of exactly $m\ell$ reads drawn from $\mathcal{D}$;

$T \leftarrow \texttt{exact\_counting}(S, k)$;

$S_\ell \leftarrow$ random partition of S into m bags of ℓ reads each;

$A \leftarrow \emptyset$;

forall $(K, o_S(K)) \in T$ **do**

  $S_K \leftarrow$ number of bags of S_ℓ where K appears;

  $\hat{f}_{S_\ell}(K) \leftarrow S_K/(m\ell\ell_{\mathcal{D},k})$;

  $f_{S_\ell}(K) \leftarrow o_S(K)/(m\ell\ell_{\mathcal{D},k})$;

  if $\hat{f}_{S_\ell}(K) \geq \theta - \varepsilon/2$ **then** $A \leftarrow A \cup (K, f_{S_\ell}(K))$;

return A;

Proposition: The output of SPRISS is *almost* an ε-approximation of $FK(\mathcal{D}, k, \theta)$ with probability $\geq 1-\delta$.

Proof: based on the *pseudo-dimension* of bags of reads.

SPRISS

Efficient Implementation:

- # of reads where a k-mer appear in a bag is well approximated by a Poisson approximation $\rightarrow$ **no need to explicitly create the bags**;
- most k-mers appear at most once in a read $\rightarrow$ frequency $\hat{f}_{S_\ell}(K)$ is well approximated with a Binomial approximation that **only requires the number of occurrences of K in sample S**

Final approach:

1. obtain sample S of m reads
2. use an exact k-mer counter to obtain frequency $f_S(K)$ of k-mers *in S*
3. use approximations to derive $\hat{f}_{S_\ell}(K)$ from $f_S(K)$
4. report in output k-mers with $\hat{f}_{S_\ell}(K) \geq \theta - \varepsilon/2$ (estimated frequency: previous slide)

Experimental Results: Accuracy and Resources

6 largest datasets from HMP (https://hmpdacc.org/HMASM/)

dataset	label	$t_{\mathcal{D},k}$	$\lvert\mathcal{D}\rvert$	$\max_{n_i}$	avg_{n_i}
SRS024075(s)	HMP1	$8.82 \cdot 10^9$	$1.38 \cdot 10^8$	95	93.88
SRS024388(s)	HMP2	$7.92 \cdot 10^9$	$1.20 \cdot 10^8$	101	96.21
SRS011239(s)	HMP3	$8.13 \cdot 10^9$	$1.24 \cdot 10^8$	101	95.69
SRS075404(t)	HMP4	$7.75 \cdot 10^9$	$1.22 \cdot 10^8$	101	93.51
SRS043663(t)	HMP5	$9.15 \cdot 10^9$	$1.31 \cdot 10^8$	100	100.00
SRS062761(t)	HMP6	$8.26 \cdot 10^9$	$1.18 \cdot 10^8$	100	100.00

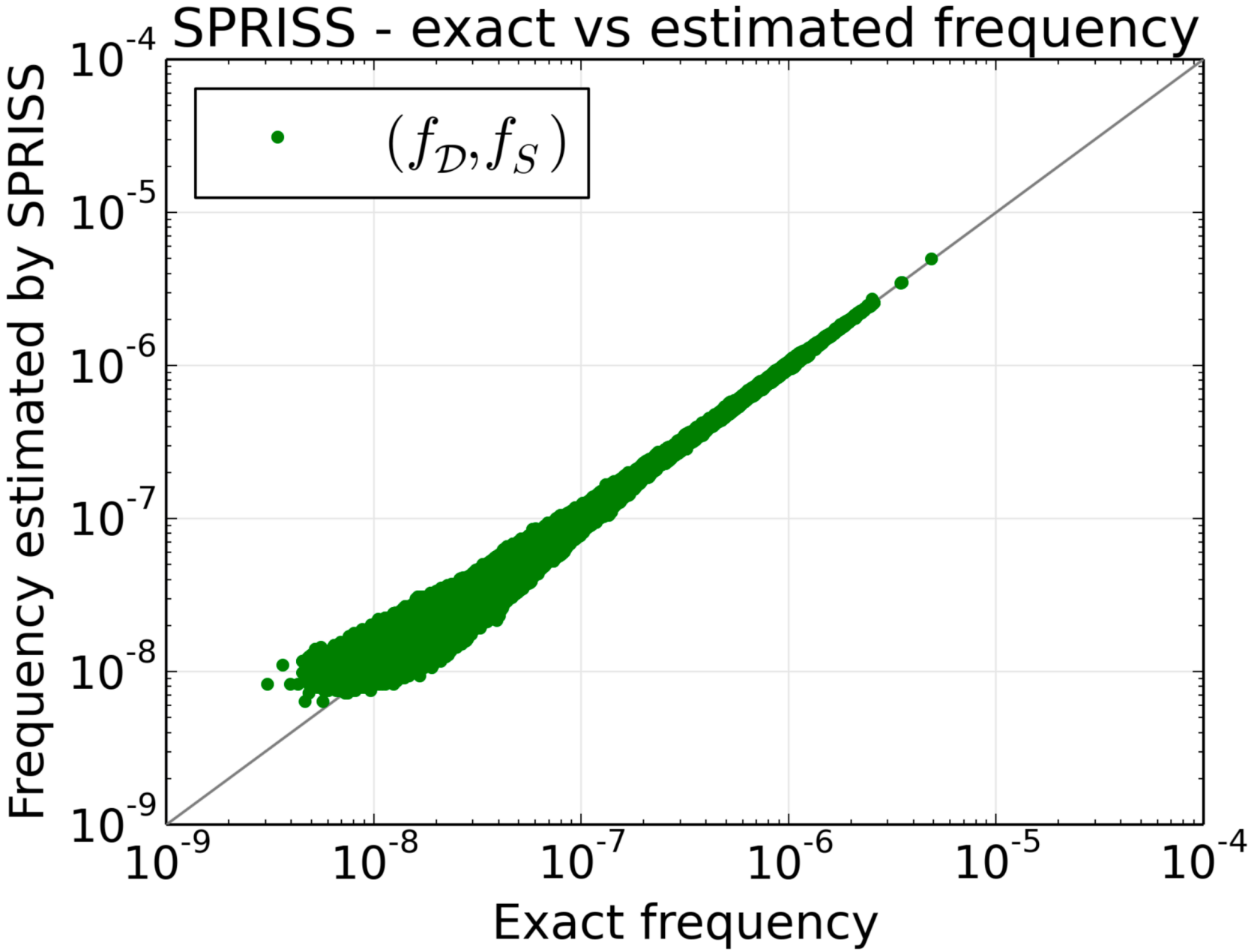
Comparison

- exact counter: KMC (Kokot et. al, 2017)
- SAKEIMA (酒今) : implemented on top of Jellyfish2 (Marçais et al., 2011)
- SPRISS: implemented on top of KMC

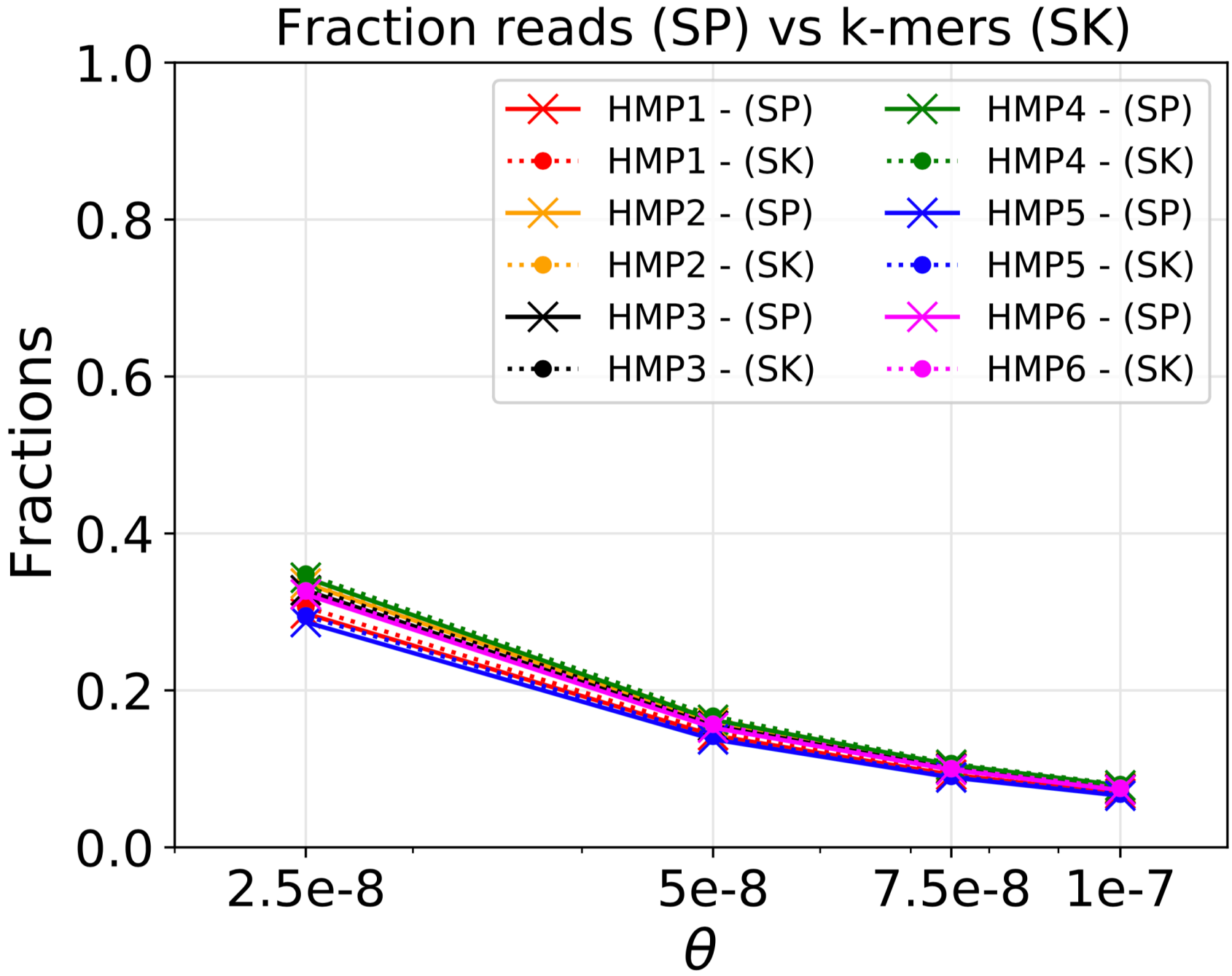
Parameters: $k = 31$, $\delta = 0.1$, $\varepsilon = \theta - 2/t_{\mathcal{D},k}$, $\ell = \lfloor 0.9/(\theta \ell_{\mathcal{D},k}) \rfloor$

Results = averages of 5 runs

Experimental Results: Accuracy



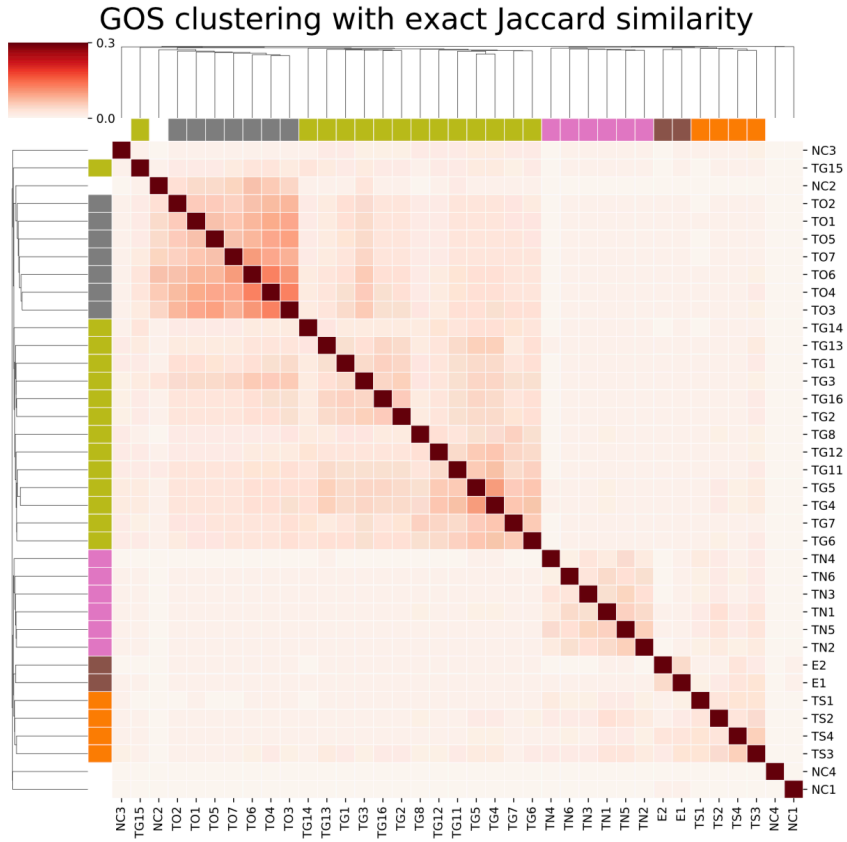
Experimental Results: Resources



Experimental Results: Comparison of Metagenomic Datasets

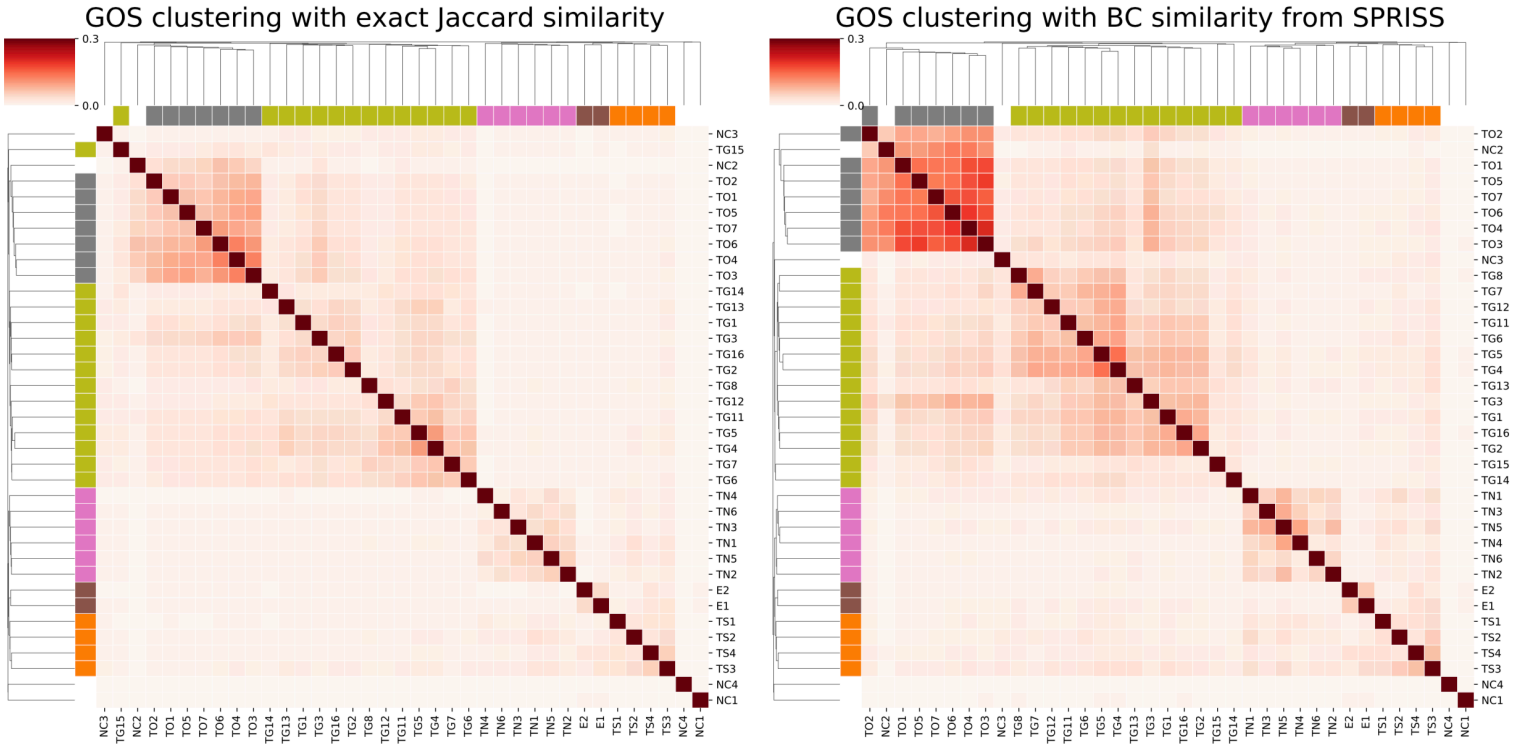
37 datasets from (Rusch et al., 2007), annotated with origin of the samples

Clustering: usually performed with *presence-based* distances (e.g., Jaccard distance) since abundance-based distances (e.g., BC distance) are more expensive



Use SPRISS to approximate the frequent k-mers?

Experimental Results: Comparison of Metagenomic Datasets



BC distance is more informative (inside-vs-outside cluster signal increases by 50%)!

SPRISS requires $< 40\%$ of the time of exact BC computation

Experimental Results: Finding Discriminative k-mers

Given two datasets $\mathcal{D}_1, \mathcal{D}_2$
Goal: find k-mers appearing more frequently in $\mathcal{D}_1$ than in $\mathcal{D}_2$, and viceversa

Given minimum frequency θ: the set $DK(\mathcal{D}_1, \mathcal{D}_2, k, \theta, \rho)$ of $\mathcal{D}_1$-of discriminative k-mers comprises k-mers K for which

1. $K \in FK(\mathcal{D}_1, k, \theta)$;
2. $f_{\mathcal{D}_1}(K) \geq 2 f_{\mathcal{D}_2}(K)$

Data from Liu et al., 2017 ($\theta = 2 \times 10^{-7}$)

dataset	$t_{\mathcal{D},k}$	$\lvert\mathcal{D}\rvert$	$\max_{n_i}$	avg_{n_i}
B73	$9.92 \cdot 10^{10}$	$4.50 \cdot 10^{8}$	250	250
Mo17	$9.97 \cdot 10^{10}$	$4.45 \cdot 10^{8}$	250	250

Exact computation (KMC): 10^4 sec
Approximation with SPRISS

- using 5% of data: $< 3\%$ false negatives in 1130 sec.
- using 10% of data: $< 2\%$ false negatives in 1970 sec.

Conclusions

Two algorithms to approximate **frequent** k-mers and applications

- SAKEIMA (酒今) : Sampling Algorithm for K-mErs approxIMAtion [Pellegrina, Pizzi, V., RECOMB 2019, JCB 2020]
- SPRISS: SamPling Reads algorIthm to eStimate frequent k-merS [Santoro*,Pellegrina*, V., RECOMB 2021]

Code:

- https://github.com/VandinLab/SAKEIMA
- https://github.com/VandinLab/SPRISS

Acknowledgements

Leonardo Pellegrina  Diego Santoro 

Cinzia Pizzi 

Fundings

MIUR of Italy