Fast Approximation of Frequent *k*-mers and Applications to Metagenomics

Leonardo Pellegrina[1,2], Cinzia Pizzi[1,3], and Fabio Vandin[1,4⋆]

[1] Department of Information Engineering, University of Padova, Via Gradenigo 6/A, Padova, 35131 (Italy)
[2] pellegri@dei.unipd.it, +39-0498277954
[3] cinzia.pizzi@dei.unipd.it, +39-0498277921
[4] fabio.vandin@unipd.it, +39-0498277946

Abstract. Estimating the abundances of all k-mers in as e t of biological sequences is a fundamental and challenging problem with many applications in biological analysis. Although several methods have been designed for the exact or approximate solution of this problem, they all require to process the entire data set, which can be extremely expensive for high-throughput sequencing data sets. Although in some applications it is crucial to estimate all k-mers and their abundances, in other situations it may be sufficient to report only frequent k-mers, which appear with relatively high frequency in a data set. This is the case, for example, in the computation of k-mers' abundance-based distances among data sets of reads, commonly used in metagenomic analyses. In this study, we develop, analyze, and test a sampling-based approach, called Sampling Algorithm for K-mErs approxIMAtion (SAKEIMA), to approximate the frequent k-mers and their frequencies in a high-throughput sequencing data set while providing rigorous guarantees on the quality of the approximation. SAKEIMA employs an advanced sampling scheme and we show how the characterization of the Vapnik-Chervonenkis dimension, a core concept from statistical learning theory, of a properly defined set of functions leads to practical bounds on the sample size required for a rigorous approximation. Our experimental evaluation shows that SAKEIMA allows to rigorously approximate frequent k-mers by processing only a fraction of a data set and that between high-throughput sequencing data sets. Overall, SAKEIMA is an efficient and rigorous tool to estimate k-mers' abundances providing significant speedups in the analysis of large sequencing data sets.

Keywords: k-mer analysis · sampling algorithm · VC dimension · metagenomics

1 Introduction

The analysis of substrings of length k, called *k-mers*, is ubiquitous in biological sequence analysis and is among the first steps of processing pipelines for a wide spectrum of applications, including: de novo assembly [Pevzner et al., 2001; Zerbino and Birney, 2008], error correction [Kelley et al., 2010; Salmela et al., 2016], repeat detection [Li and Waterman, 2003], genome comparison [Sims et al., 2009], digital normalization [Brown et al., 2012], RNA-seq quantification [Patro et al., 2014; Zhang and Wang, 2014], metagenomic reads classification [Wood and Salzberg, 2014] and binning [Girotto

⋆ Corresponding author.

et al., 2016], fast search-by-sequence over large high-throughput sequencing repositories [Solomon and Kingsford, 2016]. A fundamental task in k-mer analysis is to compute the frequency of all k-mers, with the goal to distinguish frequent k-mers from infrequent k-mers [Marçais and Kingsford, 2011; Melsted and Pritchard, 2011]. For example, this task is relevant in the analysis of high-throughput sequencing data, since infrequent k-mers are often assumed to result from sequencing errors. For several applications, the computation of k-mers frequencies is among the most computationally demanding steps of the analysis.

Many algorithms have been proposed for computing the exact frequency of all k-mers, such as Tallymer [Kurtz et al., 2008], Jellyfish [Marçais and Kingsford, 2011], BFCounter [Melsted and Pritchard, 2011], DSK [Rizk et al., 2013], KAnalyze [Audano and Vannberg, 2014], Turtle [Roy et al., 2014], KMC 3 [Kokot et al., 2017], and Squeakr-exact [Pandey et al., 2017]. These methods typically perform a linear scan of the sequences to analyze, and use a combination of parallelism and efficient data structures (such as Bloom filters and Hash tables) to maintain membership and counting information associated to all k-mers. Since the computation of exact k-mer frequencies is computationally demanding, in particular for large sequence analysis or for high-throughput sequence datasets, recent methods have focused on providing approximate solution to the problem, improving the time and memory requirements. KmerStream [Melsted and Halldórsson, 2014], Kmerlight [Sivadasan et al., 2016] and ntCard [Mohamadi et al., 2017] proposed streaming approaches for the approximation of the k-mer frequencies histogram. KmerGenie [Chikhi and Medvedev, 2013] performs a linear scan of the input, counting a random subset (chosen before processing the dataset) of all possible k-mers to approximate the abundance histogram, providing an exploratory tool to choose the value of k. khmer [Zhang et al., 2014] and the recently proposed Squeakr [Pandey et al., 2017] rely on probabilistic data structures to approximate the counts of individual k-mers. With the only exception of KmerGenie, all these methods processes *all* the k-mer occurrences in the input dataset; in addition, all the aforementioned approximate methods that report the counts of individual k-mers do not provide simultaneous estimates with rigorous guarantees for all the counts k-mers that are provided in output.

All the methods cited above try to estimate the frequency of *all* k-mers or of all k-mers that appear at least few times (e.g., twice) in the dataset. While this is crucial in some applications (e.g., in genome assembly k-mers that occur exactly once often represents sequencing errors and it is therefore important to estimate the count of all observed k-mers), in other applications this is less justified. For example, in the comparison of high-throughput sequencing metagenomic datasets, *abundance-based distances or dissimilarities* (e.g., the Bray-Curtis dissimilarity) between k-mer counts of two datasets are often used [Benoit et al., 2016; Danovaro et al., 2017; Dickson et al., 2017] to assess the distance between the corresponding datasets. In contrast to *presence-based distances* [Ondov et al., 2016] (e.g., Jaccard distance), abundance-based distances take into account the frequency of each k-mer, with frequent k-mers contributing more to the distance than k-mers that appear with low frequency, but still more than a handful of times, in the dataset. Thus, two natural questions are (i) whether the results obtained considering all k-mers can be estimated by considering the abundances of frequent k-mers only, and (ii) whether the abundances of frequent k-mers can be computed more efficiently than the counts of all k-mers. Recently, preliminary work [Hrytsenko et al., 2018] has shown that, for the cosine distance and $k = 12$, the answer to the first question is positive, and in Section 4 we show that this indeed the case for larger values of k and other abundance-based distances as well as presence-based distances (e.g., the Jaccard distance). To the best of our knowledge, the second question is hitherto unexplored. In addition, considering only frequent k-mers allows to focus on the most reliable information in a metagenomic dataset, since a high stochastic variability in low frequency k-mers is to be expected due to the sampling process inherent in sequencing.

A natural approach to reduce time and memory requirements for frequency estimation problems is to process only a portion of the data, for example by *sampling* some parts of a dataset. Sampling approaches are appealing because infrequent k-mers naturally tend to appear with lower probability in a sample, allowing to directly focus on frequent k-mers in subsequent steps. However, major

challenges in sampling approaches are (i) to provide rigorous guarantees relating the results obtained by processing the sample and the results that would be obtained from the whole dataset, and (ii) to provide effective bounds on the size of the sample required to achieve such guarantees. The application of sampling to k-mers is even more challenging than in other scenarios since, for values of k in the typical range of interest to applications (e.g., 20-60), even the most frequent k-mers have relatively low frequency in the data. To the best of our knowledge, no approach based on sampling a portion of the input dataset has been proposed to approximate frequent k-mers and their frequencies while providing rigorous guarantees.

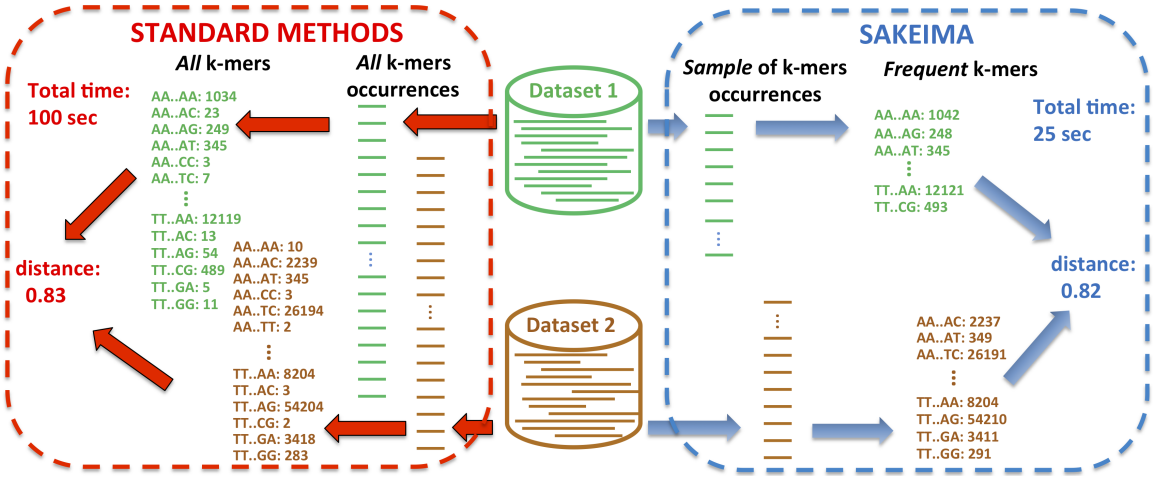

Fig. 1: SAKEIMA computes a fast and rigorous approximation of the *frequent* k-mers in a high-throughput sequencing dataset by sampling a fraction of all k-mer occurrences in a dataset, providing a significant speed-up for the computation of k-mer's abundance-based distances between datasets of reads (e.g., in metagenomics).

Our Contribution. We study the problem of approximating frequent k-mers, i.e., k-mers that appear with frequency above a user-defined threshold θ in a high-throughput sequencing dataset. In these regards, our contributions are fourfold. First, we define a rigorous definition of approximation, governed by an accuracy parameter ε. Second, we propose a new method, Sampling Algorithm for K-mErs approxIMAtion (SAKEIMA), to obtain an approximation to the set of frequent k-mers using *sampling*. SAKEIMA (see Fig.1) is based on a sampling scheme that goes beyond naïve sampling of k-mers and allows to estimate k-mers of relatively low frequency considering only a fraction of all k-mers occurrences in the dataset. Third, we provide analytical bounds to the sample size needed to obtain rigorous guarantees on the accuracy of the estimated k-mer frequencies, with respect to the ones measured on the entire dataset. Our bounds are based on the notion of VC dimension, a fundamental concept from statistical learning theory, which has been used to design efficient algorithms to identify frequent patterns in other scenarios [Riondato and Upfal, 2014; Riondato and Kornaropoulos, 2016; Servan-Schreiber et al., 2018]. To our knowledge, ours is the first method that applies concepts from *statistical learning* to provide a rigorous approximation of the k-mers frequencies. Fourth, we use SAKEIMA to extract frequent k-mers from metagenomic datasets from the Human Microbiome Project (HMP) and to approximate abundance-based and presence-based distances among such datasets, showing that SAKEIMA allows to accurately estimate such distances by analyzing only a fraction of the entire dataset, resulting in a significant speed-up.

Our approach is orthogonal to previous work: any exact or approximate algorithm can be applied to the sample extracted by SAKEIMA, that can therefore be used *before* applying previously proposed

methods, thus reducing their computational requirements while providing rigorous guarantees on the results w.r.t. to the entire dataset. While we present our methodology in the case of finding frequent k-mers from a set of sequences representing a high-throughput sequencing dataset of short reads, our results can be applied to datasets of long reads and to whole-genome sequences as well.

2 Preliminaries

Let a dataset $\mathcal{D}$ be a bag of n reads $\mathcal{D} = \{r_0, \ldots, r_{n-1}\}$, where each read r_i, $0 \leq i \leq n-1$, is a string of length n_i from an alphabet Σ of cardinality $|\Sigma| = \sigma$. For $j \in \{0, \ldots, n_i - 1\}$, let $r_i[j]$ be the j-th character of r_i. For a given integer $k \leq \min_i\{n_i : r_i \in \mathcal{D}\}$, we define a k-mer A as a string of length k from Σ, that is $A \in \Sigma^k$. We say that a k-mer A *appears* in r_i at position $j \in \{0, \ldots, n_i - k\}$ if $r_i[j+h] = A[h], \forall h \in \{0, \ldots, k-1\}$. For every $i, 0 \leq i \leq n-1$, and every $j \in \{0, \ldots, n_i - k\}$, we define the indicator function $\phi_{r_i,A}(j)$ that is 1 if the k-mer A appears in r_i at position j, while $\phi_{r_i,A}(j) = 0$ otherwise. The total number of k-mers in $\mathcal{D}$ is $t_{\mathcal{D},k} = \sum_{i=0}^{n-1}(n_i - k + 1)$. We define the *support* $o_{\mathcal{D}}(A)$ of a k-mer A as the number of distinct positions in $\mathcal{D}$ where A appears: $o_{\mathcal{D}}(A) = \sum_{i=0}^{n-1}\sum_{j=0}^{n_i-k}\phi_{r_i,A}(j)$. We define the *frequency* $f_{\mathcal{D}}(A)$ of A in $\mathcal{D}$ as the ratio between the number of distinct positions where A appears in $\mathcal{D}$ and the total number of k-mers in $\mathcal{D}$: $f_{\mathcal{D}}(A) = o_{\mathcal{D}}(A)/t_{\mathcal{D},k}$.

2.1 Frequent k-mers and Approximations

We are interested in obtaining the set $FK(\mathcal{D}, k, \theta)$ of frequent k-mers in a dataset $\mathcal{D}$ with respect to a minimum frequency threshold θ, defined as follows.

Definition 1. *Given a dataset $\mathcal{D}$, an integer $k > 0$, and a frequency threshold $\theta \in (0, 1]$, the set $FK(\mathcal{D}, k, \theta)$ of* Frequent k-Mers in $\mathcal{D}$ w.r.t. θ *is the collection of all k-mers with frequency at least θ in $\mathcal{D}$ and of their corresponding frequencies in $\mathcal{D}$:*

$$FK(\mathcal{D}, k, \theta) = \{(A, f_{\mathcal{D}}(A)) : f_{\mathcal{D}}(A) \geq \theta\}. \tag{1}$$

$FK(\mathcal{D}, k, \theta)$ can be computed with a single scan of all the k-mers occurrences in $\mathcal{D}$ maintaining the k-mers supports in an appropriate data structure; however, when $\mathcal{D}$ is extremely large and k is not small, the exact computation of $FK(\mathcal{D}, k, \theta)$ is extremely demanding in terms of time and memory, since the number of k-mers grows exponentially with k. In this case, a fast to compute *approximation* of the set $FK(\mathcal{D}, k, \theta)$ may be preferable, provided it ensures rigorous guarantees on its quality. In this work, we focus on the following approximation.

Definition 2. *Given a dataset $\mathcal{D}$, an integer $k > 0$, a frequency threshold $\theta \in (0, 1]$, and a constant $\varepsilon \in (0, \theta)$, an* ε-approximation *of $FK(\mathcal{D}, k, \theta)$ is a collection $C = \{(A, f_A) : f_A \in (0, 1]\}$ such that:*

- *for any $(A, f_{\mathcal{D}}(A)) \in FK(\mathcal{D}, k, \theta)$ there is a pair $(A, f_A) \in C$;*
- *for any $(A, f_A) \in C$ it holds that $f_{\mathcal{D}}(A) \geq \theta - \varepsilon$;*
- *for any $(A, f_A) \in C$ it holds that $|f_{\mathcal{D}}(A) - f_A| \leq \varepsilon/2$.*

The definition above guarantees that every frequent k-mer of $\mathcal{D}$ is in the approximation and that no k-mer with frequency $< \theta - \varepsilon$ is in the approximation. The third condition guarantees that the estimated frequency f_A of A in the approximation is close (i.e, within $\varepsilon/2$) to the frequency $f_{\mathcal{D}}(A)$ of A in $\mathcal{D}$. It is easy to show that obtaining a ε-approximation of $FK(\mathcal{D}, k, \theta)$ with absolute certainty requires to process all k-mers in $\mathcal{D}$.

2.2 Simple Sampling-Based Algorithms and Bounds

We aim to provide an approximation to $FK(\mathcal{D}, k, \theta)$ with *sampling*, by processing only *randomly selected portions* of $\mathcal{D}$. The simplest sampling scheme is the one in which a random sample is a bag P of m positions taken uniformly at random, with replacement, from the set $P_{\mathcal{D},k} = \{(i, j) : i \in [0, n-1], j \in [0, n_i-k]\}$ (note that $|P_{\mathcal{D},k}| = t_{\mathcal{D},k}$) of all positions where k-mers occurs in the dataset $\mathcal{D}$, corresponding to m occurrences of k-mers (with repetitions) taken uniformly at random. Given such sample P, an integer $k > 0$, and a minimum frequency threshold $\theta \in (0, 1]$ one can define the set of frequent k-mers (and their frequencies) in the sample P as $FK(P, k, \theta) = \{(A, f_P(A)) : f_P(A) \geq \theta\}$, where $f_P(A)$ is the frequency of k-mer A in the sample.

Obtaining a ε-approximation from a random sample with absolute certainty is impossible, thus we focus on obtaining a ε-approximation with probability $1 - \delta > 0$, where $\delta \in (0, 1)$ is a *confidence* parameter, whose value is provided by the user. Intuitively, the set $FK(\mathcal{D}, k, \theta)$ of frequent k-mers is well approximated by the set of frequent k-mers in a random sample P when P is sufficiently large. One natural question regards how many samples are needed to obtain the desired ε-approximation. By using Hoeffding's inequality [Mitzenmacher and Upfal, 2017] to bound the deviation of the frequency of a k-mer A in the sample from $f_{\mathcal{D}}(A)$ and a union bound on the maximum number σ^k of k-mers, where $\sigma = |\Sigma|$, we have the following result that provides a first such bound, and a corresponding first algorithm to obtain a ε-approximation to $FK(\mathcal{D}, k, \theta)$.

Proposition 1. *Consider a sample P of size m of $\mathcal{D}$. If $m \geq \frac{2}{\varepsilon^2}\left(\ln\left(2\sigma^k\right) + \ln\left(\frac{1}{\delta}\right)\right)$ for fixed $\varepsilon \in (0, \theta), \delta \in (0, 1)$, then, with probability $\geq 1-\delta$, $FK(P, k, \theta-\varepsilon/2)$ is a ε-approximation of $FK(\mathcal{D}, k, \theta)$.*

Proof. We first prove that when $m \geq \frac{2}{\varepsilon^2}\left(\ln\left(2\sigma^k\right) + \ln\left(\frac{1}{\delta}\right)\right)$, then, with probability $\geq 1-\delta$, for every k-mer A simultaneously we have $|f_P(A) - f_{\mathcal{D}}(A)| \leq \varepsilon/2$.

For an arbitrary k-mer A, given the definition of $f_P(A)$ we have that $f_P(A) = \sum_{(i,j)\in P} \phi_{r_i,A}(j)/m$ where $\sum_{(i,j)\in P} \phi_{r_i,A}(j)$ is the sum of m 0-1 independent random variables. Since $\mathbb{E}[\phi_{r_i,A}(j)] = f_{\mathcal{D}}(A)$, we have that $\mathbb{E}[f_P(A)] = f_{\mathcal{D}}(A)$, and by Hoeffding inequality [Mitzenmacher and Upfal, 2017] we have

$$\Pr(|f_P(A) - f_{\mathcal{D}}(A)| \geq \varepsilon) = \Pr\left(\left|\sum_{(i,j)\in P} \phi_{r_i,A}(j) - m f_{\mathcal{D}}(A)\right| \geq m\varepsilon\right) \leq 2e^{\frac{-2m^2\varepsilon^2}{m}} = 2e^{-2m\varepsilon^2}. \quad (2)$$

Now define the event $E_A =$ "$|f_P(A) - f_{\mathcal{D}}(A)| \leq \varepsilon/2$" and let $\bar{E}_A$ be the complementary event. From Equation 2 and the choice of m, $\Pr(\bar{E}_A) \leq 2e^{-m\varepsilon^2/2} = \delta/\sigma^k$. By union bound, the probability that at least one $\bar{E}_A$ holds is bounded by $\sum_{A\in\Sigma^k} \Pr(\bar{E}_A) \leq \delta$. Therefore with probability at least $1 - \delta$ all events E_A hold.

We now prove that when $|f_P(A) - f_{\mathcal{D}}(A)| \leq \varepsilon/2$ for every k-mer A, then $FK(P, k, \theta - \varepsilon/2)$ is a ε-approximation of $FK(\mathcal{D}, k, \theta)$. Consider an arbitrary pair $(A, f_{\mathcal{D}}(A)) \in FK(\mathcal{D}, k, \theta)$. By the definition of $FK(\mathcal{D}, k, \theta)$ we have that $f_{\mathcal{D}}(A) \geq \theta$, and, since $|f_P(A) - f_{\mathcal{D}}(A)| \leq \varepsilon/2$, we have that $f_P(A) \geq \theta - \varepsilon/2$, that is there is a pair $(A, f_A) \in FK(P, k, \theta - \varepsilon/2)$. Now consider a k-mer A with $f_{\mathcal{D}}(A) < \theta - \varepsilon$: since $|f_P(A) - f_{\mathcal{D}}(A)| \leq \varepsilon/2$ we have that $f_P(A) \leq f_{\mathcal{D}}(A) + \varepsilon/2 < \theta - \varepsilon/2$, that is there is no pair $(A, f_A) \in FK(P, k, \theta - \varepsilon/2)$. □

In addition, by using known results in statistical learning theory [Vapnik and Chervonenkis, 1971; Mitzenmacher and Upfal, 2017] relating the VC dimension (see Section 3 for its definition) of a family of functions to a newly derived bound on the family of functions $\{f_{\mathcal{D}}(A)\}$, we obtain the following improved bound and algorithm. (The derivation is in Appendix.)

Proposition 2. *Let P be a sample of size m of $\mathcal{D}$. For fixed $\varepsilon \in (0, \theta), \delta \in (0, 1)$, if $m \geq \frac{2}{\varepsilon^2}\left(1 + \ln\left(\frac{1}{\delta}\right)\right)$ then $FK(P, k, \theta-\varepsilon/2)$ is an ε-approximation for $FK(\mathcal{D}, k, \theta)$ with probability $\geq 1-\delta$.*

3 Advanced and Practical Bounds and Algorithms for *k*-mer Approximations

Athough the bound of Proposition 2 significantly improves the simple bounds of Section 1, since the factor $\ln(2\sigma^k)$ has been reduced to 1, it still has an inverse quadratic dependency with respect to the accuracy parameter ε, that is problematic when the quantities to estimate are small. In these cases, one needs a small ε to produce a meaningful approximation (since $\varepsilon < \theta$), and the inverse quadratic dependence of the sample size from ε often results in a sample size larger than the entire input, defeating the purpose of sampling. The case of k-mers is particularly challenging, since the sum $\sum_{A\in\Sigma^k} f_{\mathcal{D}}(A)$ of all k-mers frequencies is exactly 1. Therefore the higher the number of distinct k-mers appearing in the input, the lower their frequencies will be, with the consequence that θ (and therefore ε) typically needs to be set to a very low value. For example, a typical dataset from the Human Microbiome Project (HMP) has $n \approx 10^8$ reads of (average) length ≈ 100: therefore if we are interested in k-mers for $k = 31$, by setting $\delta = 0.05$ the bound of Section 2.2 gives $\varepsilon \approx 10^{-5}$, that is only k-mers with frequency $\geq 10^{-5}$ could be reliably reported by sampling. However, in datasets we considered, no or a very small number (≤ 30) of k-mers have frequency $\geq 10^{-5}$, therefore according to the result from Section 2.2 we cannot obtain a meaningful approximation of k-mers and their frequencies. In the remainder of this section we develop more refined sampling schemes and estimation techniques leading to a practical sampling-based algorithm.

3.1 Sampling Bags of Positions and VC dimension Bound.

We propose a method to provide an efficiently computable approximation to $FK(\mathcal{D}, k, \theta)$ when the minimum frequency θ is low, by properly defining samples so that any k-mer A will appear in a sample with probability higher than $f_{\mathcal{D}}(A)$, thus lessening the dependence of the sample size from $1/\varepsilon^2$. For this to be achievable, we need to relax the notion of approximation defined in Section 2. In particular, the guarantees, provided by our method, in such relaxed approximation are that *all* k-mers with frequency above θ', with θ' slightly higher than θ, are reported in output, and that no k-mer having frequency below $\theta - \varepsilon$ is reported in output. (See Proposition 5 for the definition of θ'.) Our experiments show that the fraction of k-mers having frequency $\in [\theta, \theta')$ which are non reported is very small. Our method works by sampling *bags of positions* instead of single positions. In particular, an element of the sample is now a set of ℓ positions chosen independently at random from the set $P_{\mathcal{D},k}$ of all positions.

Let $I_\ell = \{(i_1, j_1), (i_2, j_2), \ldots, (i_\ell, j_\ell)\}$ be a bag of ℓ positions for k-mers in $\mathcal{D}$, chosen uniformly at random from the set $P_{\mathcal{D},k}$. We define the indicator functions $\hat{\phi}_A(I_\ell)$ that, for a given bag I_ℓ of ℓ positions, is equal to 1 if k-mer A appears in *at least* one of the ℓ positions in I_ℓ and is equal to 0 otherwise. That is $\hat{\phi}_A(I_\ell) = \min\left\{1, \sum_{(i,j)\in I_\ell} \phi_{r_i,A}(j)\right\}$. We define the *$\ell$-positions sample* P_ℓ as a bag of m bags $\{I_{\ell,0}, I_{\ell,1}, \ldots, I_{\ell,m-1}\}$, where each $I_{\ell,j}, 0 \leq j \leq m-1$ is a bag of ℓ positions, sampled independently, and

$$\hat{f}_{P_\ell}(A) = \frac{1}{m} \sum_{I_{\ell,i}\in P_\ell} \frac{\hat{\phi}_A(I_{\ell,i})}{\ell}. \tag{3}$$

Intuitively, $\hat{f}_{P_\ell}(A)$ is the biased version of the unbiased estimator

$$f_{P_\ell}(A) = \frac{1}{m} \sum_{I_{\ell,i}\in P_\ell} \frac{\sum_{(i,j)\in I_{\ell,i}} \phi_{r_i,A}(j)}{\ell} \tag{4}$$

of $f_{\mathcal{D}}(A)$, where the bias arises from considering a value of 1 every time $\sum_{(i,j)\in I_{\ell,i}} \phi_{r_i,A}(j) > 1$.

In our analysis we use the VC dimension [Vapnik, 1998; Vapnik and Chervonenkis, 1971], a statistical learning concept that measures the expressivity of a family of binary functions. We define

a range space Q as a pair $Q = (X, R_X)$ where X is a finite or infinite set and R_X is a finite or infinite family of subsets of X. The members of R_X are called *ranges*. Given $D \subset X$, the *projection* of R_X on D is defined as $proj_{R_X}(D) = \{r \cap D : r \in R_X\}$. We say that D is *shattered by* R_X if $proj_{R_X}(D) = 2^{|D|}$. The *VC dimension of* Q, denoted as $VC(Q)$, is the maximum cardinality of a subset of X shattered by R_X. If there are arbitrary large shattered subsets of X shattered by R_X, then $VC(Q) = \infty$.

A finite bound on the VC dimension of a range space Q implies a bound on the number of random samples required to obtain a good approximation of its ranges, defined as follows.

Definition 3. *Let $Q = (X, R_X)$ be a range space and let D be a finite subset of X. For $\varepsilon \in (0, 1]$, a subset B of D is an* ε-approximation of D *if for all $r \in R_X$ we have: $\left| \frac{|D \cap r|}{|D|} - \frac{|B \cap r|}{|B|} \right| \leq \varepsilon/2$.*

The following result [Mitzenmacher and Upfal, 2017] relates ε and the probability that a random sample of size m is an ε-approximation for a range space of VC dimension at most v.

Proposition 3 ([Mitzenmacher and Upfal, 2017]). *There is an absolute positive constant c such that if (X, R_X) is a range-space of VC dimension at most v, D is a finite subset of X, and $0 < \varepsilon, \delta < 1$, then a random subset $B \subset D$ of cardinality m with $m \geq \frac{4c}{\varepsilon^2}\left(v + \ln\left(\frac{1}{\delta}\right)\right)$ is a ε-approximation of D with probability at least $1 - \delta$.*

The universal constant c has been experimentally estimated to be at most 0.5 [Löffler and Phillips, 2009].

We now prove an upper bound to the VC dimension $VC(Q)$ of the range space Q associated to the class of functions $\hat{\phi}_A$ that grows sub-linearly with respect to ℓ. To this aim, we first define the range space associated to bags of ℓ positions of k-mers.

Definition 4. *Let $\mathcal{D}$ be a dataset of n reads and let k and ℓ be two integers ≥ 1. We define $Q = (X_{\mathcal{D},k,\ell}, R_{\mathcal{D},k,\ell})$ to be the following range space:*

- *$X_{\mathcal{D},k,\ell}$ is the set of all bags of ℓ positions of k-mers in $\mathcal{D}$, that is the set of all possible subsets, with repetitions, of size ℓ from from $P_{\mathcal{D},k}$;*
- *$R_{\mathcal{D},k,\ell} = \{P_{\mathcal{D},\ell}(A) | A \in \Sigma^k\}$ is the family of sets of starting positions of k-mers, such that for each k-mer A, the set $P_{\mathcal{D},\ell}(A)$ is the set of all bags of ℓ starting positions in $\mathcal{D}$ where A appears at least once.*

We prove the following results on the VC dimension of the above range space.

Proposition 4. *Let Q the range space from Definition 4. Then: $VC(Q) \leq \lfloor \log_2(\ell) \rfloor + 1$.*

Proof. If $VC(Q) \geq v$, then there must exists a set $Z \subseteq X_{\mathcal{D},k,\ell}$ with $|Z| \geq v$ that is shattered. This means that 2^v subsets of Z must be in projection of $R_{\mathcal{D},k,\ell}$ on Z. If this is true, then every element of Z needs to belong to exactly 2^{v-1} such sets. Therefore, every element of Z needs to contain at least $\ell = 2^{v-1}$ distinct k-mers. This implies that $v \leq \log_2(\ell) + 1$, and the thesis follows. □

Using the result above, we prove the following.

Proposition 5. *Let $\ell \geq 1$ be an integer and P_ℓ be a bag of m bags of ℓ positions of $\mathcal{D}$ with*

$$m \geq \frac{2}{(\ell\varepsilon)^2}\left(\lfloor \log_2 \min(2\ell, \sigma^k) \rfloor + \ln\left(\frac{1}{\delta}\right)\right). \tag{5}$$

Then, with probability at least $1 - \delta$:

- *for any k-mer $A \in FK(\mathcal{D}, k, \theta)$ such that $f_{\mathcal{D}}(A) \geq \theta' = 1 - (1 - \ell\theta)^{1/\ell}$ it holds $\hat{f}_{P_\ell}(A) \geq \theta - \varepsilon/2$;*
- *for any k-mer A with $\hat{f}_{P_\ell}(A) \geq \theta - \varepsilon/2$ it holds $f_{\mathcal{D}}(A) \geq \theta - \varepsilon$;*

- *for any k-mer* $A \in FK(\mathcal{D}, k, \theta)$ *it holds* $f_{\mathcal{D}}(A) \geq \hat{f}_{P_\ell}(A) - \varepsilon/2$*;*
- *for any k-mer* A *with* $\hat{f}_{P_\ell}(A) - \varepsilon/2 \geq 0$*, it holds* $f_{\mathcal{D}}(A) \geq 1 - (1 - \ell(\hat{f}_{P_\ell}(A) - \varepsilon/2))^{1/\ell}$*;*
- *for any k-mer* A *with* $\ell(\hat{f}_{P_\ell}(A) + \varepsilon/2) \leq 1$ *it holds* $f_{\mathcal{D}}(A) \leq 1 - (1 - \ell(\hat{f}_{P_\ell}(A) + \varepsilon/2))^{1/\ell}$*.*

Proof. For a given k-mer A, consider the event $E_A =$ "$|\mathbb{E}[\hat{f}_{P_\ell}(A)] - \hat{f}_{P_\ell}(A)| \leq \varepsilon/2$". Note that it is equivalent to "$|\mathbb{E}[\ell\hat{f}_{P_\ell}(A)] - \ell\hat{f}_{P_\ell}(A)| \leq \ell\varepsilon/2$" and that $\ell\hat{f}_{P_\ell}(A) = \frac{1}{m}\sum_{i=0}^{m-1}\hat{\phi}_A(I_{\ell,i})$, therefore $\mathbb{E}[\ell\hat{f}_{P_\ell}(A)] = \mathbb{E}[\hat{\phi}_A(I_{\ell,i})]$. Now note that if for the range space $Q = (X_{\mathcal{D},k,\ell}, R_{\mathcal{D},k,\ell})$ we consider $r_A = P_{\mathcal{D},\ell}(A)$, we have that $\frac{|X_{\mathcal{D},k,\ell} \cap r_A|}{|X_{\mathcal{D},k,\ell}|} = \mathbb{E}[\hat{\phi}_A(I_{\ell,i})]$, since $I_{\ell,i}$ is a bag of ℓ positions taken uniformly at random among all possible such bags and therefore $\mathbb{E}[\hat{\phi}_A(I_{\ell,i})]$ is the fraction of bags of ℓ positions that contain at least a position where A occurs (i.e., $\mathbb{E}[\hat{\phi}_A(I_{\ell,i})]$ is w.r.t. the uniform distribution over bags of ℓ positions). Therefore, combining Proposition 4 and Proposition 3, for the given choice of m we have that with probability $1-\delta$ it holds that $|\mathbb{E}[\ell\hat{f}_{P_\ell}(A)] - \ell\hat{f}_{P_\ell}(A)| \leq \ell\varepsilon/2, \forall A$, or, equivalently, $|\mathbb{E}[\hat{f}_{P_\ell}(A)] - \hat{f}_{P_\ell}(A)| \leq \varepsilon/2, \forall A$: we assume that this holds in the rest of the proof.

Consider a k-mer A with frequency $f_{\mathcal{D}}(A)$ in $\mathcal{D}$. From the definition of $\hat{f}_{P_\ell}(A)$, we have $\mathbb{E}[\hat{f}_{P_\ell}(A)] \leq \mathbb{E}[f_{P_\ell}(A)] = f_{\mathcal{D}}(A)$. Let $X_i = \hat{\phi}_A(I_{\ell,i})/\ell$ be the random variable taking value $1/\ell$ if the k-mer A appears at least once in the ℓ positions of $I_{\ell,i}$, and value 0 otherwise. We have that: $\mathbb{E}\left[\hat{f}_{P_\ell}(A)\right] = \frac{1}{m}\sum_{I_{\ell,i}\in P_\ell}\mathbb{E}\left[X_i\right] = \frac{1}{m}\sum_{I_{\ell,i}\in P_\ell}\frac{1}{\ell}\Pr\left(X_i \geq 1\right) = \left(1 - (1 - f_{\mathcal{D}}(A))^\ell\right)/\ell$. Now consider a k-mer A with $f_{\mathcal{D}}(A) \geq 1 - (1-\ell\theta)^{1/\ell}$. By the derivation above we have that $\mathbb{E}\left[\hat{f}_{P_\ell}(A)\right] \geq \theta$, and therefore its frequency $\hat{f}_{P_\ell}(A)$ in the sample P_ℓ satisfies $\hat{f}_{P_\ell}(A) \geq \theta - \varepsilon/2$, that completes the proof of the first part.

For the second part, consider a k-mer A with $f_{\mathcal{D}}(A) < \theta - \varepsilon$. By the derivation above, we have that $\mathbb{E}[\hat{f}_{P_\ell}(A)] \leq \mathbb{E}[f_{P_\ell}(A)] = f_{\mathcal{D}}(A) < \theta - \varepsilon$. Since $|\mathbb{E}[\hat{f}_{P_\ell}(A)] - \hat{f}_{P_\ell}(A)| \leq \varepsilon/2, \forall A$, we have that $\hat{f}_{P_\ell}(A) < \theta - \varepsilon/2$, which proves the second part of the result.

The third result follows from $|\mathbb{E}[\hat{f}_{P_\ell}(A)] - \hat{f}_{P_\ell}(A)| \leq \varepsilon/2$ and $\mathbb{E}[\hat{f}_{P_\ell}(A)] \leq f_{\mathcal{D}}(A)$.

The last two results follow from $|\mathbb{E}[\hat{f}_{P_\ell}(A)] - \hat{f}_{P_\ell}(A)| \leq \varepsilon/2$ and $\mathbb{E}[\hat{f}_{P_\ell}(A)] = (1-(1-f_{\mathcal{D}}(A))^\ell)/\ell$. □

Note that from Proposition 5 the set $\{(A, f_{P_\ell}(A)) : \hat{f}_{P_\ell}(A) \geq \theta - \varepsilon/2\}$ is *almost* a ε-approximation to $FK(\mathcal{D}, k, \theta)$: in particular, there may be k-mers A for which $\mathbb{E}[\hat{f}_{P_\ell}(A)] = (1-(1-f_{\mathcal{D}}(A))^\ell)/\ell < \theta$ while $f_{\mathcal{D}}(A) = \mathbb{E}[f_{P_\ell}(A)] \geq \theta$ and such that for the given sample P_ℓ we have $\hat{f}_{P_\ell}(A) \approx \mathbb{E}[\hat{f}_{P_\ell}(A)] - \varepsilon/2$. While this can happen, we can limit the probability of this happening by appropriately choosing ℓ, and still enjoy the reduction in sample size of the order of $\frac{\log_2 \ell}{\ell^2}$ w.r.t. Proposition 2 obtained by considering bags of bags of ℓ positions. In particular, this result allows the user to set θ, ε, δ, and ℓ to effectively find, with probability at least $1-\delta$, *all* frequent k-mers A for which $f_{\mathcal{D}}(A) \geq \theta'$ and do not report any k-mer with frequency below $\theta - \varepsilon$, while still being able to report in output almost all k-mers with frequency $\in [\theta, \theta')$. Our experimental analysis (Section 4) shows that in practice choosing ℓ close from below to $1/\theta$ is very effective to obtain such result. Then, the third, fourth, and fifth guarantees from Proposition 5 state that we can use the biased estimates $\hat{f}_{P_\ell}(A)$ to derive *guaranteed upper and lower bounds* to $f_{\mathcal{D}}(A)$ that will be much tighter than the one obtained using the bounds of Section 2.2. We will show how to obtain further improved upper and lower bounds to $f_{\mathcal{D}}(A)$ in Section 3.3. Such lower bounds ℓb_A can be used, for example, to prove that the set $\{(A, f_{P_\ell}(A)) : \ell b_A \geq \theta - \varepsilon\}$ enjoys the same last four guarantees from Proposition 5 while the first one holds for a $\theta' < 1 - (1-\ell\theta)^{1/\ell}$; therefore, when false negatives are problematic, the set $\{(A, f_{P_\ell}(A)) : \ell b_A \geq \theta - \varepsilon\}$ can be used to obtain a different approximation of $FK(\mathcal{D}, k, \theta)$ with fewer false negatives.

3.2 SAKEIMA: An Efficient Algorithm to Approximate Frequent k-mers

We now present our SAKEIMAthat builds on Proposition 5 and efficiently samples a bag P_ℓ of bags of ℓ-positions from $\mathcal{D}$ to obtain an approximation of the set $FK(\mathcal{D}, k, \theta)$ with probability $1 - \delta$, where δ is a parameter provided by the user.

Algorithm 1: SAKEIMA

Input: dataset $\mathcal{D}$, total number of k-mers $t_{\mathcal{D},k}$ in $\mathcal{D}$,
frequency threshold θ, accuracy parameter $\varepsilon \in (0, \theta)$,
confidence parameter $\delta \in (0, 1)$, integer $\ell \geq 1$.
Output: approximation $\{(A, f_A)\}$ of $FK(\mathcal{D}, k, \theta)$ with probability $\geq 1 - \delta$.

1 $m \leftarrow \left\lceil \frac{2}{(\ell\varepsilon)^2} \left(\lfloor \log_2 \min(2\ell, \sigma^k) \rfloor + \ln\left(\frac{2}{\delta}\right) \right) \right\rceil$; $\lambda \leftarrow \frac{m\ell}{t_{\mathcal{D},k}}$;
2 $T \leftarrow$ empty hash table;
3 **forall** *reads* $r_i \in \mathcal{D}$ **do**
4 **forall** $j \in [0, n_i - k]$ **do**
5 $A \leftarrow k$-mer in position j of read r_i;
6 $a \leftarrow Poisson(\lambda)$;
7 **if** $a > 0$ **then** $T[A] \leftarrow T[A] + a$;
8 $\mathcal{O} \leftarrow \emptyset$; $t \leftarrow \sum_{A \in T} T[A]$;
9 $P_\ell \leftarrow$ random partition of the occurrences in T into m bags;
10 **forall** k*-mers* $A \in T$ **do**
11 $f_A \leftarrow T[A]/t$;
12 $P_A \leftarrow$ bags of P_ℓ where A appears at least once;
13 $\hat{f}_A \leftarrow |P_A|/(m\ell)$;
14 **if** $\hat{f}_A \geq \theta - \varepsilon/2$ **then** $\mathcal{O} \leftarrow \mathcal{O} \cup (A, f_A)$;
15 **return** $\mathcal{O}$;

SAKEIMA is described in Algorithm 1. Athough SAKEIMA performs a linear scan of the input dataset, it practically reduces the number of k-mers that need to be processed with the following stategy. SAKEIMA performs a pass on the stream of k-mers appearing in $\mathcal{D}$, and for each position in the stream it draws the number a of times that the position appears in the sample P_ℓ independently at random from the Poisson distribution $Poisson(\lambda)$ of parameter $\lambda = m\ell/t_{\mathcal{D},k}$. SAKEIMA stores such values, if strictly positive, in a counting structure T (lines 3-7) that keeps, for each k-mer A, the total number of occurrences of A in the sample P_ℓ. Note that $t_{\mathcal{D},k}$ can be computed with a very quick linear scan of the dataset, where n_i is computed for every $r_i \in \mathcal{D}$ without extracting and processing (e.g., inserting or updating information for) k-mers; in alternative a lower bound to $t_{\mathcal{D},k}$ can be used, simply resulting in a number of samples higher than needed. For each k-mer A appearing at least once in the sample, the unbiased estimate f_A is computed in line 11 as the number $T[A]$ of occurrences of A in the sample P_ℓ divided by the total number of positions in the sample t. The biased estimate $\hat{f}_A$ can be computed partitioning the $T[A]$ occurrences of A into m bags $I_{\ell,0}, \ldots, I_{\ell,m-1}$; $\hat{f}_A$ is then simply the ratio between the number of bags where A appears at least once and $m\ell$. We describe a more efficient way of computing such biased estimate at the end of this section. Then SAKEIMA flags A as frequent if $\hat{f}_A \geq \theta - \varepsilon/2$ (line 14) and, in this case, the couple (A, f_A) is added to the output set $\mathcal{O}$ (line 15), since f_A is the best (and unbiased) estimate to $f_{\mathcal{D}}(A)$.

Note that SAKEIMA does not sample m bags of *exactly* ℓ positions each, since the number of occurrences of each position in $\mathcal{D}$ in the sample P_ℓ is sampled independently from a Poisson distribution, even if the expected number of total occurrences sampled from the algorithm is $m\ell$. However, the independent Poisson distributions used by SAKEIMA provide an accurate approximation of the

random sampling of *exactly* $m\ell$ positions used in the analysis of Section 3.1. In particular, this holds when one focuses on the events of interests for our approximation of Section 3.1 (e.g., the event "there exists a k-mer A such that $|\mathbb{E}[\hat{f}_{P_\ell}(A)] - \hat{f}_{P_\ell}(A)| > \varepsilon/2$"). In fact, a simple adaptation of a known result (Corollary 5.11 of [Mitzenmacher and Upfal, 2017]) on the relation between sampling with replacement and the use of independent Poisson distributions gives the following.

Proposition 6. *Let E be an event whose probability is either monotonically increasing or monotonically decreasing in the number of sampled positions. If E has probability p when the independent Poisson distributions are used, then E has probability at most $2p$ when the sampling with replacement is used.*

As a simple corollary, the output $\mathcal{O}$ features the guarantees of Proposition 5 with probability $\geq 1-\delta'$, with $\delta' = 2\delta$.

The technique we just described can be used to avoid the exact computation of $\hat{f}_A$, which requires to maintain and update the counters for the m buckets; in fact, we can approximate the number of occurrences of a k-mer A, appearing $T[A]$ times in the random sample of `SAKEIMA` into a given bucket as a sample from $Poisson(T[A]/m)$. This means that the number of buckets where A will be inserted *at least once* is well approximated by a sample from $Binomial(m, 1 - e^{-T[A]/m})$, which models the number of successes in m independent trials with probability of success $1 - e^{-T[A]/m}$. Due to this second Poisson approximation, we obtain that the output $\mathcal{O}$ provides the guarantees of Proposition 5 with probability $\geq 1 - \delta''$, with $\delta'' = 4\delta$. In terms of Algorithm 1, such modification simply requires to substitute $\frac{2}{\delta}$ with $\frac{4}{\delta}$ in line 1, to remove line 9, and to substitute lines 12-13 with "$\hat{f}_A \leftarrow Binomial(m, 1 - e^{-T[A]/m})/(m\ell)$". This also allows to efficiently compute multiple values of $\hat{f}_A$, corresponding to different values of ℓ, by simply taking samples from binomial distributions of different appropriate parameters. (In particular, if one samples a total t of k-mers, then the value m to be used for both parameters of the binomial distribution is t/ℓ.) The next section shows why this is useful.

3.3 Improved Lower and Upper Bounds to k-mers Frequencies

Note that Proposition 5 guarantees that we can obtain upper and lower bounds to $f_{\mathcal{D}}(A)$ for every $A \in FK(\mathcal{D}, k, \theta)$ from the sample of bags of ℓ positions. These bounds are meaningful only in specific ranges of the frequencies; for example, the lower bound from the third guarantee in Proposition 5 is meaningful when the frequency of A is fairly low, i.e $f_{\mathcal{D}}(A) \approx 1/\ell$, while for very frequent k-mers they could be a multiplicative factor $1/\ell$ away from than the correct value. For example, if a k-mer is very frequent and appears in all bags of ℓ k-mers in a sample S, its corresponding lower bound is still only $1/\ell - \varepsilon/2$.

However, Proposition 5 can be generalized to obtain tighter upper and lower bounds to the frequency of all k-mers. For given ℓ, ε, and δ, let m as given in Proposition 5. Note that the total number of k-mer's positions in the sample P_ℓ is $m\ell$. Let $\mathcal{L}$ be a set of integer values $\mathcal{L} = \{\ell_i\}$ with $\ell_i \in [1, m\ell], \forall i = 0, \ldots, |\mathcal{L}| - 1$. Now, for every $\ell_i \in \mathcal{L}$, we can partition the *same* $m\ell$ k-mers that are in P_ℓ into $m_i = m\ell/\ell_i$ partitions having size ℓ_i. Let P_{ℓ_i} be such a random partition of such positions into m_i bags of ℓ_i positions each. Note that each P_{ℓ_i} is a "valid" sample (i.e., a sample of independent bags of positions, each obtained by uniform sampling with replacement) for Proposition 5, even if the P_{ℓ_i}'s are not independent. From each P_{ℓ_i}, we define a maximum deviation ε_i from Proposition 5 as $\varepsilon_i = \frac{1}{\ell_i}\sqrt{\frac{2}{m_i}\left(\lfloor \log_2(\min(2\ell_i, \sigma^k)) \rfloor + \ln\left(|\mathcal{L}|/\delta\right)\right)}$. We have the following result.

Proposition 7. *With probability at least $1 - \delta$, for all k-mers A simultaneously and for all the random partitions induced by $\mathcal{L}$ it holds*

- $f_{\mathcal{D}}(A) \geq \max\{\hat{f}_{P_{\ell_i}}(A) - \varepsilon_i/2 : i = 0, \ldots, |\mathcal{L}| - 1\}$;
- $f_{\mathcal{D}}(A) \geq \max\{1 - (1 - \ell(\hat{f}_{P_{\ell_i}}(A) - \varepsilon_i/2))^{1/\ell} : i = 0, \ldots, |\mathcal{L}| - 1$ *and* $\hat{f}_{P_{\ell_i}}(A) - \varepsilon_i/2 \geq 0\}$;

– $f_{\mathcal{D}}(A) \leq \min\{1-(1-\ell(\hat{f}_{P_{\ell_i}}(A)+\varepsilon_i/2))^{1/\ell} : i=0,\ldots,|\mathcal{L}|-1 \text{ and } \hat{f}_{P_{\ell_i}}(A)+\varepsilon_i/2 \leq 1/\ell\}$.

Proof. Combining proposition 4 and Proposition 3 and by union bound on the $|\mathcal{L}|$ values of i, we have that with probability $1-\delta$ it holds that $|\mathbb{E}[\hat{f}_{P_{\ell_i}}(A)] - \hat{f}_{P_{\ell_i}}(A)| \leq \varepsilon/2, \forall A$ and $\forall i = 0,\ldots,|\mathcal{L}|-1$: we assume that this holds in the rest of the proof. To prove the lower bound, note that since $\mathbb{E}[\hat{f}_{P_{\ell_i}}(A)] = (1-(1-f_{\mathcal{D}}(A))^{\ell})/\ell$, from the above we have that

$$(1-(1-f_{\mathcal{D}}(A))^{\ell})/\ell \geq \hat{f}_{P_{\ell_i}}(A) - \varepsilon_i/2$$

that is equivalent to

$$f_{\mathcal{D}}(A) \geq 1-(1-\ell(\hat{f}_{P_{\ell_i}}(A)-\varepsilon_i/2))^{1/\ell}$$

when $\hat{f}_{P_{\ell_i}}(A) - \varepsilon_i/2 \geq 0$. The proof of the upper bound is analogous. □

In our experiments, we use $\mathcal{L} = \{\ell_i\}$ with $\ell_i = \ell/2^i, \forall i \in [0, \lfloor \log_2 \ell \rfloor - 1]$; in this case, note that $P_{\ell_0} = P_\ell$. Using this scheme, we can compute upper and lower bounds for k-mers having frequencies of many different orders of magnitude, but any (application dependent) distribution can be specified by the user. Then, these upper and lower bounds can be used to obtain different approximations of $FK(\mathcal{D}, k, \theta)$ with different guarantees. For example, by reporting all k-mers (and their frequencies) that have an upper bound $\geq \theta$, we have an approximation that guarantees that all k-mers A with $f_{\mathcal{D}}(A) \geq \theta$ are in the approximation.

4 Experimental Results

In this section we present the results of our experimental evaluation for `SAKEIMA`. Section 4.1 describes the datasets, our implementation for `SAKEIMA`[5], and the baseline for comparisons. In Section 4.2, we report the results for computing the approximation of the frequent k-mers using `SAKEIMA`. Section 4.3 reports the results of using our approximation to compute abundance-based and presence-based distances between metagenomic datasets.

4.1 Datasets and Implementation

We considered six datasets from the Human Microbiome Project (HMP)[6], one of the largest publicly available collection of metagenomic datasets from high-throughput sequencing. In particular, we selected the three largest datasets of `stool` and the three largest of `tongue dorsum` (Table 1). These datasets constitute the most challenging instances, due to their size, and provide a test case with different degrees of similarities among datasets. We implemented `SAKEIMA` in `C++` as a modification of Jellyfish [Marçais and Kingsford, 2011] (the version we used is 2.2.10[7]), a very popular and efficient algorithm for exact k-mer counting. Doing so, our algorithm enjoys the succinct counting data structure provided by Jellyfish publicly available implementation. We remark that our sampling-based approach can be used in combination with any other highly tuned method available for exact, approximate, and parallel k-mer counting. For this reason, we only compare `SAKEIMA` with the exact counting performed by Jellyfish, since they share the underlying characteristics, allowing us to evaluate the impact of `SAKEIMA`'s sampling strategy.

For running time and memory we computed the average from 10 runs. When comparing Jellyfish and `SAKEIMA` using 1 worker, we show the CPU time, while when using multiple threads we show the overall running time. We did not include the time to compute $t_{\mathcal{D},k}$ in our experiments since we assume it is provided in input (for example, computed while the dataset of read is created). In cases

[5] Available at https://github.com/VandinLab/SAKEIMA
[6] https://hmpdacc.org/HMASM
[7] https://github.com/gmarcais/Jellyfish

when it is not known in advance, $t_{\mathcal{D},k}$ can be computed by simply scanning all the k-mers without counting them. We computed the time required for this task for the datasets we consider and it was always small (i.e., always less than 175 seconds with 1 worker, and than 70 seconds with 32 workers) compared to the time for counting k-mers.

For the computation of the abundance-based distances from the k-mer counts of two dataset, we implemented in `C++` a simple algorithm that loads the counts of one dataset in main memory and then performs one pass on the counts of the other dataset, producing the distances in output. We executed all our experiments on the same machine with 512 GB of RAM and 2.30 GHz Intel Xeon CPUs (with 64 cores in total), compiling both implementations with `GCC 8`. `SAKEIMA` can be used in combination with more efficient algorithms and implementations for the computation of these (and other) distances [Benoit et al., 2016], resulting in speed-ups analogous to the ones we present below. For all the experiments of `SAKEIMA`, given θ and a dataset $\mathcal{D}$, we fixed the parameters $\delta = 0.1$, $\varepsilon = \theta - 2/t_{\mathcal{D},k}$, and we fix $\ell = \lfloor 0.9/\theta \rfloor$.

4.2 Approximation of the Frequent k-mers

We fixed $k = 31$, and we compared `SAKEIMA` with the exact counting of all k-mers (from Jellyfish) in terms of:

(i) running time, including, for both algorithms, the time required to write the output on disk;

(ii) memory requirement. We also assessed the accuracy of the output of `SAKEIMA`.

Figure 2 shows the average running times and peak memory as function of θ, using 1 worker. Note that for the exact counting algorithm these metrics do not depend on θ, since it always counts all k-mers. `SAKEIMA` is always faster than the exact counting, with a difference that increases when θ increases and a speed-up around 2 even for $\theta = 2 \cdot 10^{-8}$. The memory requirement of `SAKEIMA` reduces when θ increases, and for $\theta = 2 \cdot 10^{-8}$ it is half of the memory required by the exact counting. This is due to `SAKEIMA`'s sample size being much smaller than the dataset size (Figure 2(d)), therefore a large portion of extremely low frequency k-mers are naturally left out from the random sample and do not need to be accounted for in the counting data structure, as confirmed by counting the number of *distinct* k-mers that are inserted in the counting data structure by the two algorithms (Figure 2(c)). (The difference between the memory requirement and the number of distinct k-mers is given by Jellyfish's strategy to doubles the size of the counting data structure when it is full.)

Figure 3 shows the average running times of `SAKEIMA` and Jellyfish as function of θ and the number of workers w for counting k-mer from dataset `SRS043663`. The memory used by both approaches does not depend on w, therefore it is the same of Figure 2. We can see that increasing w reduces the running time of both approaches, and that the relative improvements provided by the sampling strategy of `SAKEIMA` is mantained. This shows that `SAKEIMA` is well suited to be combined with parallel approaches.

In terms of quality of the approximation, the output of `SAKEIMA` satisfied the guarantees given by Proposition 5 for all runs of our experiments, therefore with probability higher than $1 - \delta$. While `SAKEIMA` may incur in false negatives, its false negative ratio (i.e., the fraction of k-mers in $FK(\mathcal{D}, k, \theta)$ not reported by `SAKEIMA`) is always $\leq 3 \cdot 10^{-4}$ (Figure 4(a)), even if the sampling technique of Section 3.1 does not provide rigorous guarantees on such quantity. Therefore `SAKEIMA` is very effective in reporting almost all frequent k-mers. As mentioned in Section 3.3, `SAKEIMA` can be easily modified so to report all frequent k-mers in output, even if at the cost of reporting also more k-mers with frequency between $\theta - \varepsilon$ and θ. In addition, the estimated frequencies f_A reported by `SAKEIMA` are always close to the true values $f_{\mathcal{D}}(A)$, with a small maximum deviation $|f_A - f_{\mathcal{D}}(A)|$ (Figure 4(b)), and an even smaller average deviation (Figure 4(c)). In addition, the upper and lower bounds computed as in Section 3.3 provide small confidence intervals always containing the value $f_{\mathcal{D}}(A)$ (e.g., Figure 4(d) for dataset `SRS062761`), and could be used to obtain sets of k-mers with various guarantees from the sample used by `SAKEIMA`.

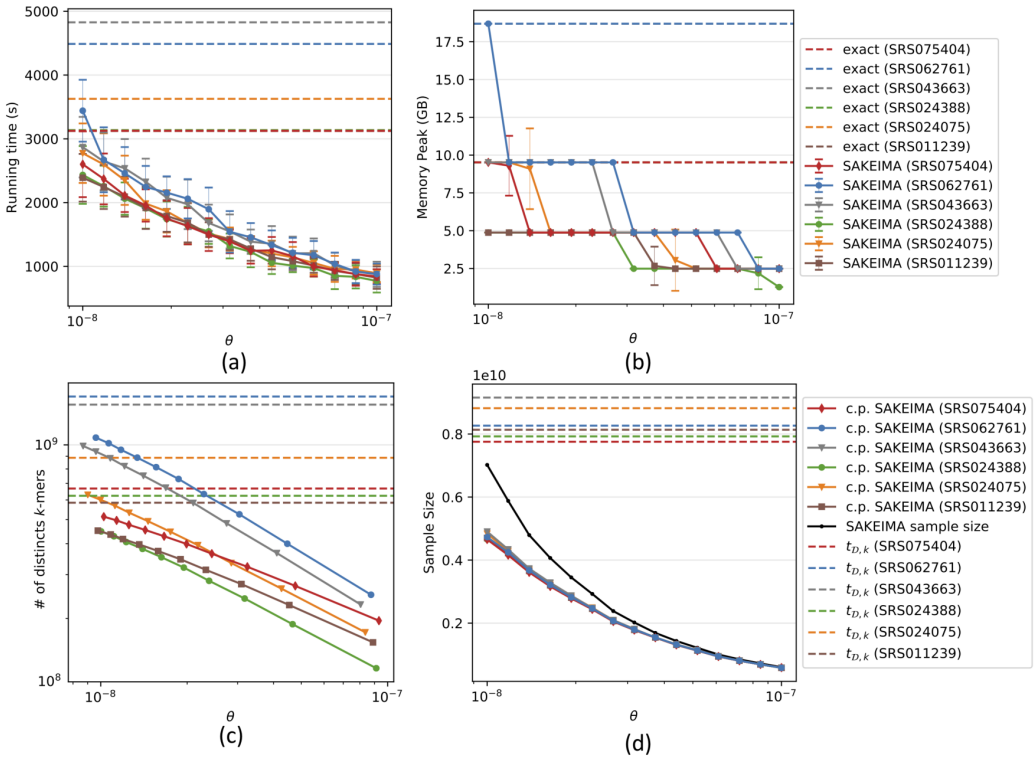

Fig. 2: Running time, memory requirements, and number of distinct k-mers counted, for SAKEIMA and exact counting as function of θ. (a) Running time (average ± 2 standard deviations from 10 runs). (b) Memory requirement (the standard deviation is not shown when all the 10 runs have the same peak memory). (c) Number of distinct k-mers counted. (d) Sample sizes of SAKEIMA, total size $t_{\mathcal{D},k}$ of the datasets, and number (c.p.) of dataset's distinct covered positions (i.e., included in SAKEIMA's sample), as function of θ.

4.3 Application to Metagenomics: Computation of Ecological Distances

We evaluate the use of SAKEIMA to speed up the computation of commonly used k-mer based ecological distances [Benoit et al., 2016] between datasets of Next-Generation Sequencing (NGS) reads. We present results for the Bray-Curtis distance; analogous results hold for other distances (see Appendix).

We first investigated how the distances change when those are computed considering only the *frequent* k-mers (w.r.t. a frequency threshold θ) instead that the full spectrum of k-mers appearing in the data. Therefore, given a pair of datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ and θ, we computed the sets $\mathcal{O}_1 = FK(\mathcal{D}_1, k, \theta)$ and $\mathcal{O}_2 = FK(\mathcal{D}_2, k, \theta)$ using Jellyfish and then computed a generalized version of the distances for all pairs of datasets we used for our experiments. For the Bray-Curtis distance, this generalization is defined as: $BC(\mathcal{D}_1, \mathcal{D}_2, \mathcal{O}_1, \mathcal{O}_2) = 1 - 2\frac{\sum_{A \in \mathcal{O}_1 \cap \mathcal{O}_2} \min\{o_{\mathcal{D}_1}(A), o_{\mathcal{D}_2}(A)\}}{\sum_{A \in \mathcal{O}_1} o_{\mathcal{D}_1}(A) + \sum_{A \in \mathcal{O}_2} o_{\mathcal{D}_2}(A)}$.

Note that when $\theta \leq 10^{-10}$ then $FK(\mathcal{D}, k, \theta)$ coincides with the set of *all* k-mers, for any of the datasets we tested. The results (Figure 5(a)) show that for θ up to 5×10^{-8} the values of the distances are fairly stable and therefore one can use only frequent k-mers for such values of θ to compute the distances, and that for θ up to 10^{-7} the relation between distances of different pairs of datasets are almost always conserved. We underline that the exact counting approach needs to count *all* the k-mers and only afterwards can filter the infrequent ones before writing them to disk to compute

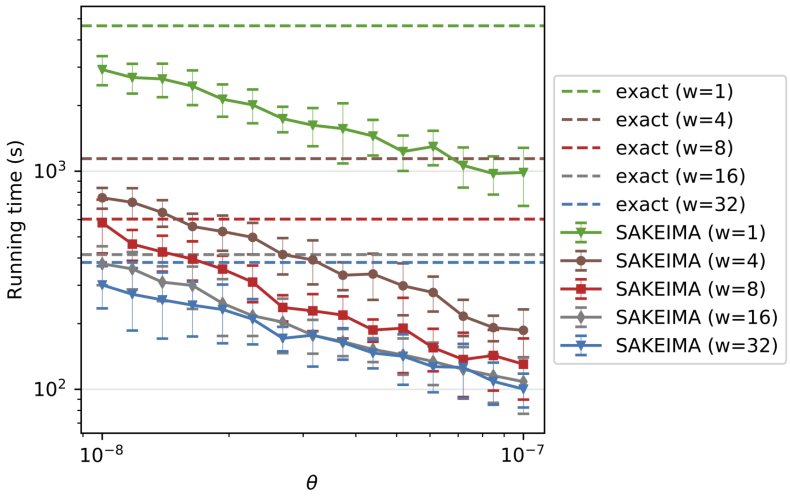

Fig. 3: Running time for SAKEIMA and exact counting for dataset SRS043663, as function of θ and the number of workers w.

$FK(\mathcal{D}, k, \theta)$. We then used SAKEIMA to extract approximations (of k-mers and their frequencies) of $FK(\mathcal{D}_1, k, \theta)$ and $FK(\mathcal{D}_2, k, \theta)$ and used such approximations to compute the distances among datasets (Figure 5(b)). Strikingly, the distances computed from the output of SAKEIMA are very close to their exact variant (Figure 5(c)). Interestingly this holds also for the Jaccard distance, a presence-based distance that does not depend neither on k-mer abundances nor on k-mer ranking by frequencies.

We then compared, for different values of θ, the total running time required to compute the approximations of the frequent k-mers using SAKEIMA for all datasets in Table 1 and all distances among such datasets using SAKEIMA approximations with the running time required when the exact counting algorithm is used for the same tasks. SAKEIMA reduces the computing time by more than 75% (Figure 5(d)). This result comes from both the efficiency of SAKEIMA and from the fact that by focusing on the the most frequent k-mers we greatly reduce the number of distinct k-mers that need to be processed for computing the distances. Therefore SAKEIMA can be used for a very fast comparison of metagenomic datasets while preserving the ability of distinguishing similar datasets from different datasets.

5 Conclusion

We presented SAKEIMA, a sampling-based algorithm to approximate frequent k-mers and their frequencies with rigorous guarantees on the quality of the approximation. We show that SAKEIMA can be used to speed up the analysis of large high-throughput sequencing metagenomic datasets, in particular to compute abundance-based distances among such datasets. Interestingly SAKEIMA allows to compute accurate approximations also for presence-based distances (e.g., the Jaccard distance), even if for such distances other, potentially faster, tools [Ondov et al., 2016] are available. SAKEIMA can be combined with any highly optimized method that counts all k-mers in a set of strings, including recent parallel methods designed for comparative metagenomics [Benoit et al., 2016]. While we presented results for k-mers from datasets of short reads, SAKEIMA can also be used for the analysis of spaced seeds [Břinda et al., 2015], large datasets of long reads, and whole genome sequences.

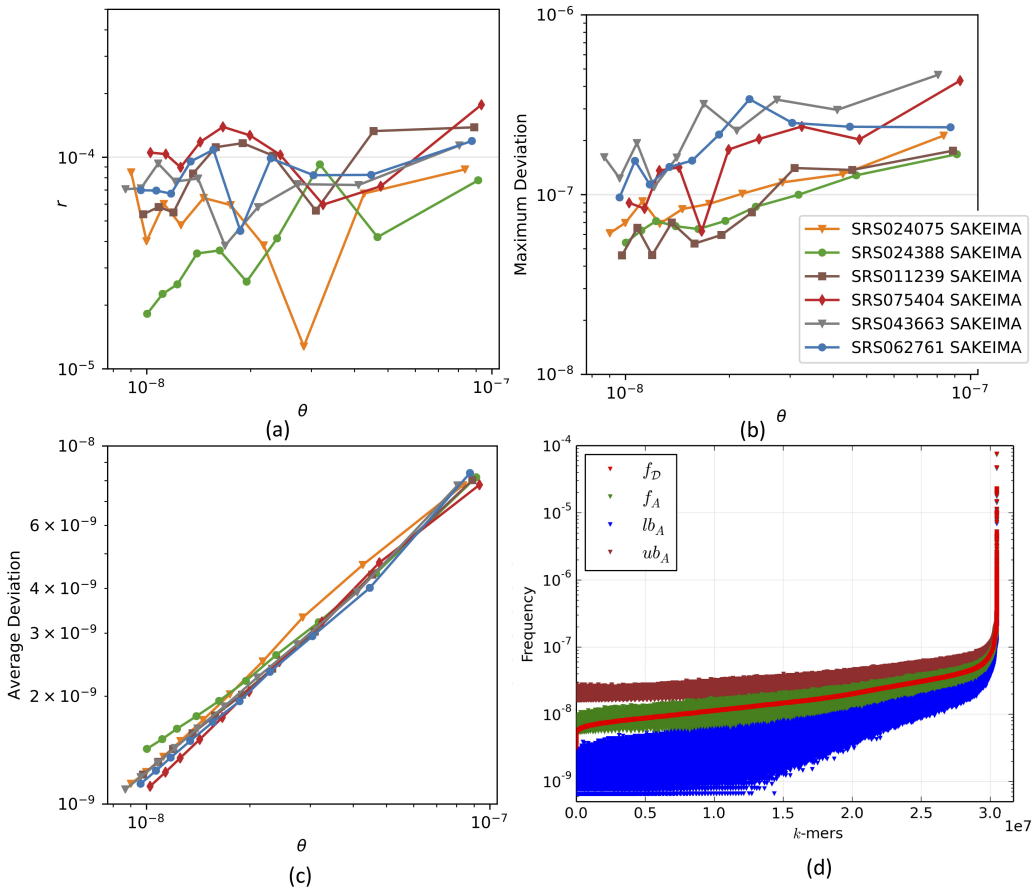

Fig. 4: Quality of the approximation of $FK(\mathcal{D}, k, \theta)$ produced by `SAKEIMA`. (a) False negative rate, i.e., the fraction r of k-mers in $FK(\mathcal{D}, k, \theta)$ not reported by `SAKEIMA`. (b) Maximum deviation $|f_A - f_{\mathcal{D}}(A)|$ of the estimates reported by `SAKEIMA` for various θ. (c) Average value of $|f_A - f_{\mathcal{D}}(A)|$ for the k-mers A reported by `SAKEIMA` for various θ. (d) Frequencies and bounds for dataset `SRS062761` and $\theta = 10^{-8}$ shown for k-mers sorted in increasing order of exact frequencies. Red: exact frequencies $f_{\mathcal{D}}(A)$. Green: estimate f_A of $f_{\mathcal{D}}(A)$ from `SAKEIMA`. Blue: lower bound lb_A to $f_{\mathcal{D}}(A)$ from `SAKEIMA`. Brown: upper bound ub_A to $f_{\mathcal{D}}(A)$ from `SAKEIMA`.

6 Acknowledgments

This work is supported, in part, by the University of Padova grants *SID2017* and *STARS: Algorithms for Inferential Data Mining*.

7 Author Disclosure Statement

No competing financial interests exist.

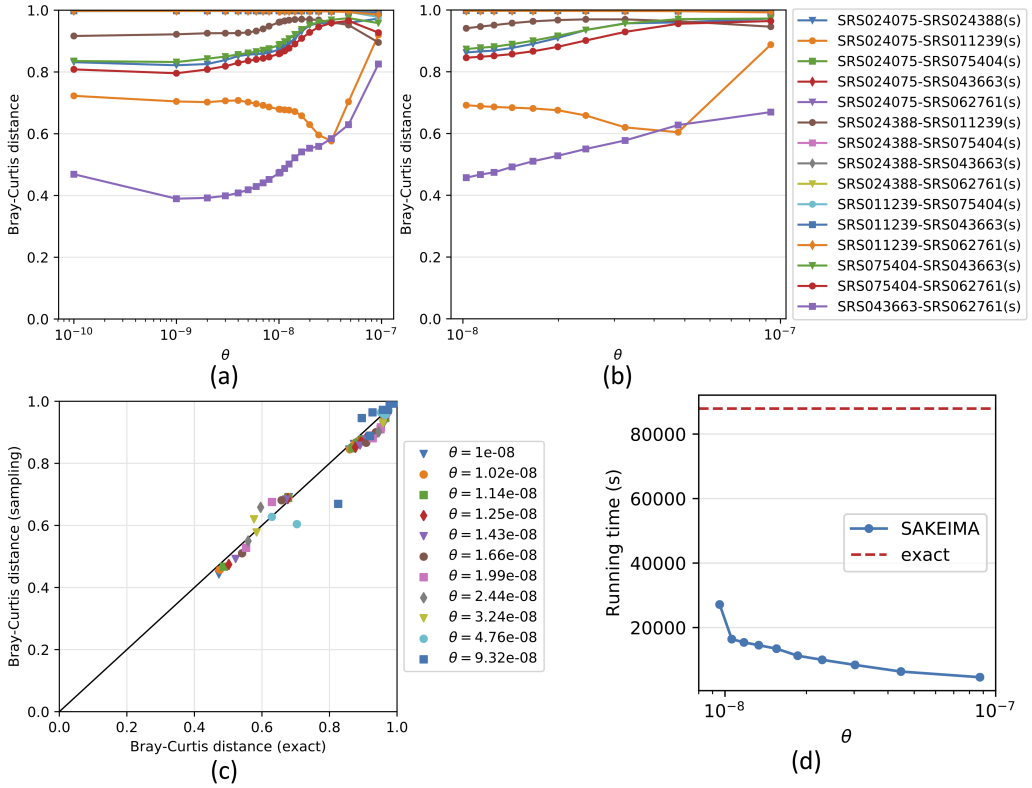

Fig. 5: Results for Bray-Curtis (BC) distances of metagenomic datasets. (a) BC distance computed using k-mers with frequency $\geq \theta$. (b) BC distances computed using the approximation of k-mers with frequency $\geq \theta$ from `SAKEIMA`. (c) Comparison of the BC distance using all k-mers with exact counts and the approximation of frequent k-mers by `SAKEIMA`. (d) Total time required by `SAKEIMA` and the exact approach to find frequent k-mers and compute all distances between datasets as a function of θ.

Bibliography

Audano, P. and Vannberg, F. (2014). Kanalyze: a fast versatile pipelined k-mer toolkit. *Bioinformatics*, 30(14):2070–2072.

Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., and Lemaitre, C. (2016). Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*, 2:e94.

Břinda, K., Sykulski, M., and Kucherov, G. (2015). Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics*, 31(22):3584–3592.

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv preprint arXiv:1203.4802*.

Chikhi, R. and Medvedev, P. (2013). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37.

Danovaro, R., Canals, M., Tangherlini, M., Dell'Anno, A., Gambi, C., Lastras, G., Amblas, D., Sanchez-Vidal, A., Frigola, J., Calafat, A. M., et al. (2017). A submarine volcanic eruption leads to a novel microbial habitat. *Nature ecology & evolution*, 1(6):0144.

Dickson, L. B., Jiolle, D., Minard, G., Moltini-Conclois, I., Volant, S., Ghozlane, A., Bouchier, C., Ayala, D., Paupy, C., Moro, C. V., et al. (2017). Carryover effects of larval exposure to different environmental bacteria drive adult trait variation in a mosquito vector. *Science advances*, 3(8):e1700585.

Girotto, S., Pizzi, C., and Comin, M. (2016). Metaprob: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics*, 32(17):i567–i575.

Hrytsenko, Y., Daniels, N. M., and Schwartz, R. S. (2018). Efficient distance calculations between genomes using mathematical approximation. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 546–546. ACM.

Kelley, D. R., Schatz, M. C., and Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. *Genome biology*, 11(11):R116.

Kokot, M., Długosz, M., and Deorowicz, S. (2017). Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761.

Kurtz, S., Narechania, A., Stein, J. C., and Ware, D. (2008). A new method to compute k-mer frequencies and its application to annotate large repetitive plant genomes. *BMC genomics*, 9(1):517.

Li, X. and Waterman, M. S. (2003). Estimating the repeat structure and length of dna sequences using ℓ-tuples. *Genome research*, 13(8):1916–1922.

Löffler, M. and Phillips, J. M. (2009). Shape fitting on point sets with probability distributions. In *European Symposium on Algorithms*, pages 313–324. Springer.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770.

Melsted, P. and Halldórsson, B. V. (2014). Kmerstream: streaming algorithms for k-mer abundance estimation. *Bioinformatics*, 30(24):3541–3547.

Melsted, P. and Pritchard, J. K. (2011). Efficient counting of k-mers in dna sequences using a bloom filter. *BMC bioinformatics*, 12(1):333.

Mitzenmacher, M. and Upfal, E. (2017). *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press.

Mohamadi, H., Khan, H., and Birol, I. (2017). ntcard: a streaming algorithm for cardinality estimation in genomics data. *Bioinformatics*, 33(9):1324–1330.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using minhash. *Genome biology*, 17(1):132.

Pandey, P., Bender, M. A., Johnson, R., and Patro, R. (2017). Squeakr: an exact and approximate k-mer counting system. *Bioinformatics*.

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5):462.

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An eulerian path approach to dna fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753.

Riondato, M. and Kornaropoulos, E. M. (2016). Fast approximation of betweenness centrality through sampling. *Data Mining and Knowledge Discovery*, 30(2):438–475.

Riondato, M. and Upfal, E. (2014). Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(4):20.

Rizk, G., Lavenier, D., and Chikhi, R. (2013). Dsk: k-mer counting with very low memory usage. *Bioinformatics*, 29(5):652–653.

Roy, R. S., Bhattacharya, D., and Schliep, A. (2014). Turtle: Identifying frequent k-mers with cache-efficient algorithms. *Bioinformatics*, 30(14):1950–1957.

Salmela, L., Walve, R., Rivals, E., and Ukkonen, E. (2016). Accurate self-correction of errors in long reads using de bruijn graphs. *Bioinformatics*, 33(6):799–806.

Servan-Schreiber, S., Riondato, M., and Zgraggen, E. (2018). Prosecco: Progressive sequence mining with convergence guarantees. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 417–426. IEEE.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106(8):2677–2682.

Sivadasan, N., Srinivasan, R., and Goyal, K. (2016). Kmerlight: fast and accurate k-mer abundance estimation. *arXiv preprint arXiv:1609.05626*.

Solomon, B. and Kingsford, C. (2016). Fast search of thousands of short-read sequencing experiments. *Nature biotechnology*, 34(3):300.

Vapnik, V. (1998). *Statistical learning theory. 1998*. Wiley, New York.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46.

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829.

Zhang, Q., Pell, J., Canino-Koning, R., Howe, A. C., and Brown, C. T. (2014). These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *PloS one*, 9(7):e101271.

Zhang, Z. and Wang, W. (2014). Rna-skim: a rapid method for rna-seq quantification at transcript level. *Bioinformatics*, 30(12):i283–i292.

8 Tables

Table 1: Datasets for our experimental evaluation. For each dataset $\mathcal{D}$ the table shows: the dataset name and site (`(s)` for `stool`, `(t)` for `tongue dorsum`); the total number $t_{\mathcal{D},k}$ of k-mers ($k = 31$) in $\mathcal{D}$; the number $|\mathcal{D}|$ of reads it contains; the maximum read length $\max_{n_i} = \max_i\{n_i | r_i \in \mathcal{D}\}$; the average read length $\text{avg}_{n_i} = \sum_{i=0}^{n-1} n_i/n$.

dataset	$t_{\mathcal{D},k}$	$\lvert\mathcal{D}\rvert$	$\max_{n_i}$	avg_{n_i}
`SRS024388(s)`	$7.92 \cdot 10^9$	$1.20 \cdot 10^8$	102	97.21
`SRS011239(s)`	$8.13 \cdot 10^9$	$1.24 \cdot 10^8$	102	96.69
`SRS024075(s)`	$8.82 \cdot 10^9$	$1.38 \cdot 10^8$	96	94.88
`SRS075404(t)`	$7.75 \cdot 10^9$	$1.22 \cdot 10^8$	102	94.51
`SRS062761(t)`	$8.26 \cdot 10^9$	$1.18 \cdot 10^8$	101	101.00
`SRS043663(t)`	$9.15 \cdot 10^9$	$1.31 \cdot 10^8$	101	101.00

Appendix

A Proof of Proposition 2

In this section we derive the proof of Proposition 2. In our analysis we use the VC dimension of the range space associated to k-mers. We now define the range space associated to k-mers and derive an upper bound to its VC dimension.

Definition 5. *Let $\mathcal{D}$ be a bag of n reads and let $k > 0$ be an integer. For any k-mer A, let $P_{\mathcal{D},k}(A)$ be the set of elements of $P_{\mathcal{D},k}$ corresponding to the occurrences of A in $\mathcal{D}$. We define the range space $Q = (X_{\mathcal{D},k}, R_{\mathcal{D},k})$ associated to the k-mers in $\mathcal{D}$ as follows:*

- *$X_{\mathcal{D},k}$ is the set of all occurrences of k-mers in $\mathcal{D}$, that is: $X_{\mathcal{D},k} = P_{\mathcal{D},k}$;*
- *$R_{\mathcal{D},k} = \{P_{\mathcal{D},k}(A) | A \in \Sigma^k\}$.*

Note that for any A, if we consider $r = P_{\mathcal{D},k}(A) \subseteq P_{\mathcal{D},k}$ we have $|X_{\mathcal{D},k} \cap r|/|X_{\mathcal{D},k}| = f_{\mathcal{D}}(A)$. Therefore, by taking $\mathcal{D} = X_{\mathcal{D},k}$ and $R_X = R_{\mathcal{D},k}$ in Definition 3, we have that an ε-approximation B of $X_{\mathcal{D},k}$ guarantees that $|f_{\mathcal{D}}(A) - f_B(A)| \leq \varepsilon/2$.

A trivial upper bound [Shalev-Shwartz and Ben-David, 2014] to the VC dimension v of the range space $Q = (X_{\mathcal{D},k}, R_{\mathcal{D},k})$ is given by $v \leq \lfloor \log_2 |R_{\mathcal{D},k}| \rfloor = \lfloor \log_2 \sigma^k \rfloor$. However, we derive the following tighter upper bound to v, that is instrumental in obtaining an improved bound on the number of samples required for a ε-approximation.

Proposition 8. *Let $\mathcal{D}$ be a bag of n reads, $k > 0$ an integer, and $Q = (X_{\mathcal{D},k}, R_{\mathcal{D},k})$ be the corresponding range space. Then the VC dimension $VC(Q)$ of Q is 1.*

Proof. The proof is by contradiction. Assume that $VC(Q) = v' > 1$: therefore there exists a set $X \subseteq X_{\mathcal{D},k}$ with $|X| \geq v'$ that can be shattered by $R_{\mathcal{D},k}$. In order to be shattered, there should exist at least $2^{v'}$ k-mers $A_1, A_2, \ldots, A_{2^{v'}}$ such that the projection of their corresponding ranges on X gives all subsets of X. Consider two subsets X', X'' of X for which $X' \neq X''$ and $X' \cap X'' \neq \emptyset$. Since X' and X'' must be in the projection of the ranges corresponding to $A_1, A_2, \ldots, A_{2^{v'}}$ on X, there must exist two distinct k-mers A_i and A_j for which $P_{\mathcal{D},k}(A_i) = X'$ and $P_{\mathcal{D},k}(A_j) = X''$. This is a contradiction, since if $X' \cap X'' \neq \emptyset$, then each position in $X' \cap X''$ must be the starting position for the two distinct k-mers A_i and A_j, while a position can be the starting position for only one k-mer. □

Proposition 2 simply derives from from Proposition 3 and from Proposition 8.

B Frequency Histograms of 31-mers

We show in Figure S1 the exact frequency histograms we computed with Jellyfish of the k-mers (with $k = 31$) for all the datasets we considered in our experiments. For every dataset we computed $\frac{1}{2}\sqrt{\frac{2}{t_{\mathcal{D},k}}\left(1 + \log(\frac{1}{\delta})\right)}$ (with $\delta = 0.05$), that is a lower bound to the frequency threshold $\theta - \varepsilon/2$ (drawn in the plots with red vertical lines) that can be obtained from the results of Section 2.2.

C Distances for Datasets of Reads

In our experimental evaluation we considered three abundance-based distances and one presence-based distance commonly used to compare metagenomic datasets [Benoit et al., 2016], and generalized them to the scenario in which only a set of all k-mers are observed. Let $\mathcal{O}$ be a subset of the set

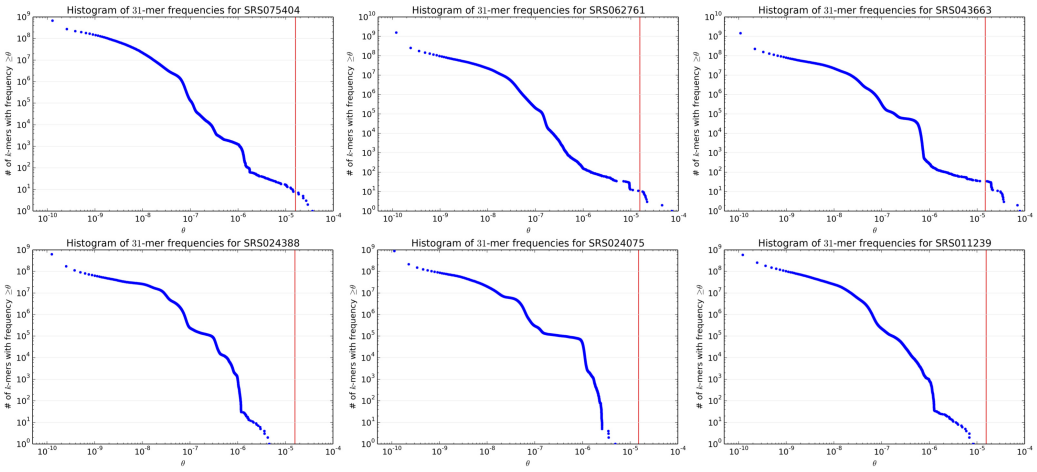

Fig. S1: Histograms of the exact frequencies of the datasets we tested. The vertical red line is drawn in correspondence of a lower bound to $\theta - \varepsilon/2 = \frac{1}{2}\sqrt{\frac{2}{t_{\mathcal{D},k}}\left(1+\log(\frac{1}{\delta})\right)}$ (with $\delta = 0.05$), that is the lowest achievable frequency threshold using the results of Section 2.2.

of all possible k-mers. Define the indicator functions: 1. $f_{\mathcal{D},\mathcal{O}}(A) = f_{\mathcal{D}}(A)$ if $A \in \mathcal{O}$, $f_{\mathcal{D},\mathcal{O}}(A) = 0$ otherwise; and 2. 3. $o_{\mathcal{D},\mathcal{O}}(A) = o_{\mathcal{D}}(A)$ if $A \in \mathcal{O}$, $o_{\mathcal{D},\mathcal{O}}(A) = 0$ otherwise.Given two datasets $\mathcal{D}_1, \mathcal{D}_2$, let $\mathcal{O}_1$ and $\mathcal{O}_2$ be the k-mers observed for $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. We considered the following distances:

- the Bray-Curtis distance: $BC(\mathcal{D}_1, \mathcal{D}_2, \mathcal{O}_1, \mathcal{O}_2) = 1 - 2\frac{\sum_{A\in\Sigma^k} \min\{o_{\mathcal{D}_1,\mathcal{O}_1}(A), o_{\mathcal{D}_2,\mathcal{O}_2}(A)\}}{\sum_{A\in\Sigma^k} o_{\mathcal{D}_1,\mathcal{O}_1}(A) + \sum_{A\in\Sigma^k} o_{\mathcal{D}_2,\mathcal{O}_2}(A)}$;
- the Whittaker distance: $W_t(\mathcal{D}_1, \mathcal{D}_2, \mathcal{O}_1, \mathcal{O}_2) = \frac{1}{2}\sum_{A\in\Sigma^k} |f_{\mathcal{D}_1,\mathcal{O}_1}(A) - f_{\mathcal{D}_2,\mathcal{O}_2}(A)|$;
- the Chord distance: $C_h(\mathcal{D}_1, \mathcal{D}_2, \mathcal{O}_1, \mathcal{O}_2) = \sqrt{2 - 2\sum_{A\in\Sigma^k} \frac{o_{\mathcal{D}_1,\mathcal{O}_1}(A) o_{\mathcal{D}_2,\mathcal{O}_2}(A)}{\sqrt{\sum_{A\in\Sigma^k} o_{\mathcal{D}_1,\mathcal{O}_1}(A)^2}\sqrt{\sum_{A\in\Sigma^k} o_{\mathcal{D}_2,\mathcal{O}_2}(A)^2}}}$;
- the Jaccard distance: $J_c(\mathcal{D}_1, \mathcal{D}_2, \mathcal{O}_1, \mathcal{O}_2) = 1 - \frac{|\mathcal{O}_1 \cap \mathcal{O}_2|}{|\mathcal{O}_1 \cup \mathcal{O}_2|}$.

For the Jaccard distance, we considered only k-mers appearing at least twice in the datasets, since k-mers with count 1 often represents sequencing errors and greatly affect the accuracy of presence-based distances, such as the Jaccard distance.

D Results for other distances

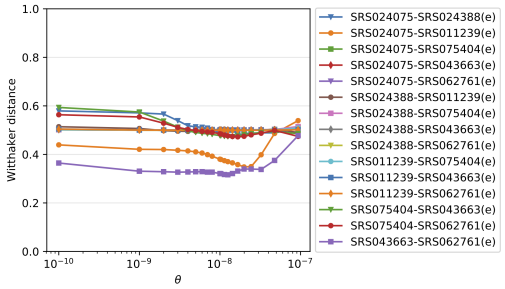

Fig. S2: Whittaker distance on exact frequent k-mers.

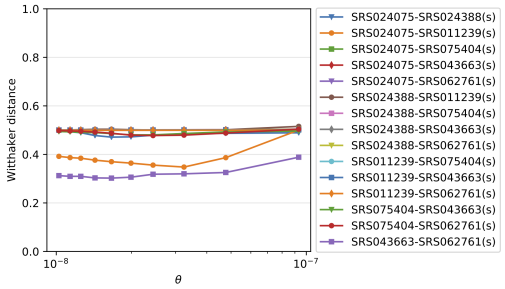

Fig. S3: Whittaker distance on output of `SAKEIMA`.

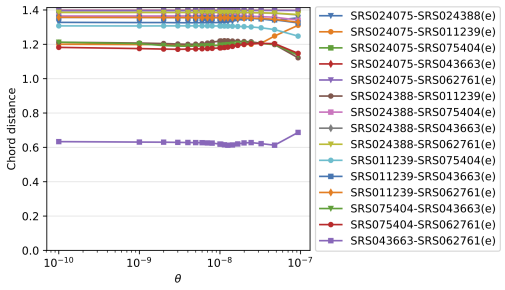

Fig. S4: Chord distance on exact frequent k-mers.

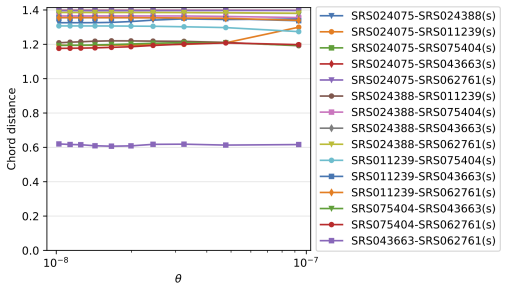

Fig. S5: Chord distance on output of `SAKEIMA`.

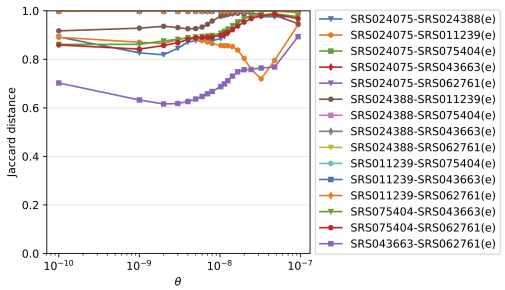

Fig. S6: Jaccard distance on exact frequent k-mers.

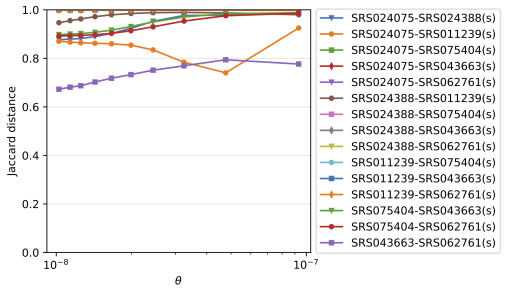

Fig. S7: Jaccard distance on output of `SAKEIMA`.

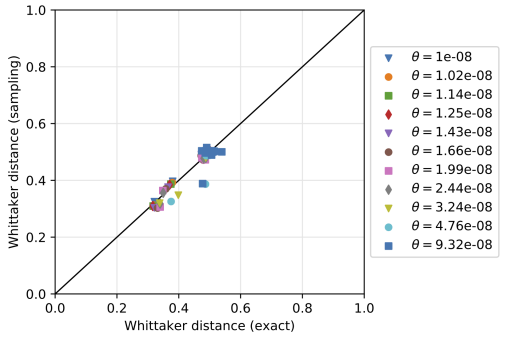

Fig. S8: Whittaker distance.

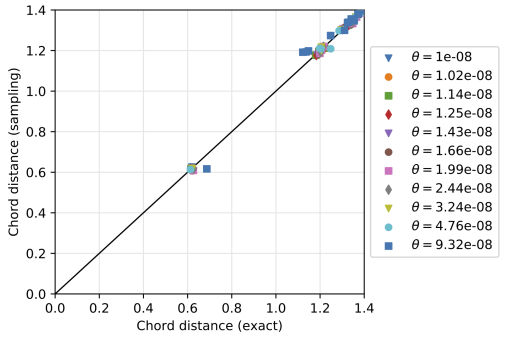

Fig. S9: Chord distance.

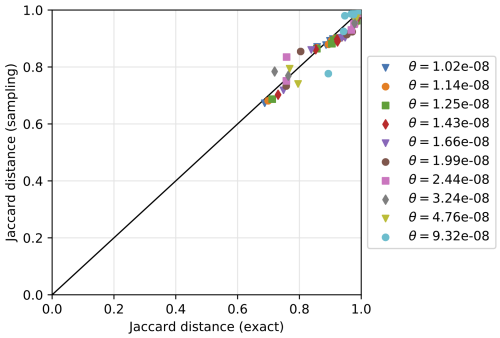

Fig. S10: Jaccard distance.