Towards Efficient k-Mer Set Operations via Function-Assigned Masked Superstrings

Ondřej Sladký
ETH Zürich, Switzerland
Computer Science Institute of Charles University, Prague, Czechia

Pavel Veselý
Computer Science Institute of Charles University, Prague, Czechia

Karel Břinda
Inria, Irisa, Univ. Rennes, 35042 Rennes, France

Abstract

The design of efficient dynamic data structures for large k-mer sets belongs to central challenges of sequence bioinformatics. Recent advances in compact k-mer set representations via simplitigs/Spectrum-Preserving String Sets, culminating with the masked superstring framework, have provided data structures of remarkable space efficiency for wide ranges of k-mer sets. However, the possibility to perform set operations remained limited due to the static nature of the underlying compact representations. Here, we develop f-masked superstrings, a concept combining masked superstrings with custom demasking functions f to enable k-mer set operations based on index merging. Combined with the FMSI index for masked superstrings, we obtain a memory-efficient k-mer membership index and compressed dictionary supporting set operations via Burrows-Wheeler Transform merging. The framework provides a promising theoretical solution to a pressing bioinformatics problem and highlights the potential of f-masked superstrings to become an elementary data type for k-mer sets.

2012 ACM Subject Classification Applied computing → Bioinformatics; Theory of computation → Pattern matching

Keywords and phrases k-mer sets, data structures, set operations, masked superstrings

Funding *Ondřej Sladký*: Supported by GA ČR project 22-22997S and ERC-CZ project LL2406 of the Ministry of Education of Czech Republic.
Pavel Veselý: Supported by GA ČR project 22-22997S, ERC-CZ project LL2406 of the Ministry of Education of Czech Republic, and Center for Foundations of Modern Computer Science (Charles Univ. project UNCE 24/SCI/008).
Karel Břinda: Supported by French National Research Agency (ANR) under Grant ANR-24-CE45-1226 for the REALL project.

1 Introduction

To store and analyze the vast volumes of DNA sequencing data [55], modern bioinformatics methods increasingly rely on k-mers to bypass the computationally expensive sequence alignment and mitigate data heterogeneity. k-mer-based methods are used across many applications such as large-scale data search [6, 12, 30], metagenomic classification [13, 56], infectious disease diagnostics [7, 11], or transcript abundance quantification [8, 44]. All such applications depend on the efficiency of the underlying k-mer data structures.

The design of efficient k-mer data structures is a central topic of contemporary sequence bioinformatics [16, 39], with the following challenges. First, as even a single k-mer set can be large, potentially containing up to hundreds of billions distinct k-mers [30], the main challenge is to provide data structures that are efficient in space and time simultaneously. Second, as modern genomic databases undergo rapid development, thanks to their growing content and curation efforts, rapid and space-efficient updates across k-mer indexes are increasingly needed to avoid repetitive and costly index recomputations. This includes scenarios such as k-mer set operations across sets, and additions or removals of individual k-mers.

The space efficiency has been particularly well addressed using the (repetitive) Spectrum-Preserving String Sets, (r)SPSS, that exploit k-mer non-independence [15]. In particular, the key observation is that genomic k-mer sets can be typically represented by k-long substrings of a small number of (arbitrarily long) strings, which is the so-called *spectrum-like property (SLP)* [16]. Building on this observation, several textual representations of k-mer sets were proposed. The first ones were based on de Bruijn graphs, to which we jointly refer as *(repetitive) Spectrum Preserving String Sets* or *(r)SPSS* [9, 10, 17, 18, 46, 51, 52]. (r)SPSS are currently highly standard and many data structures for k-mer sets base upon them [2, 5, 41, 45, 50],

Masked superstrings (MS) have provided additional space gains [53], by the virtue of a better k-mer set compaction. The core improvement over (r)SPSS lies in modeling the structure of k-mer sets by overlap graphs instead of de Bruijn graphs, thus being able to exploit overlaps of *any* length. MS represent k-mer sets using an approximately shortest superstring of all k-mers and a binary mask to avoid false positive k-mers. MS generalize any existing (r)SPSS representation as these can always be encoded as MS, but provide further compression power, especially for sets without the SLP, such as those arising from sketching or subsampling. The resulting representation is well indexable using a technique called Masked Burrows Wheeler Transform [54], resulting in a k-mer data structure with $2 + o(1)$ bits per k-mer under the SLP.

However, the lack of dynamicity of both (r)SPSS and MS – and of their derived data structures – has been limiting their wider applicability. To the best of our knowledge, the only supported operations were union, either via merging (r)SPSS of several k-mer sets resulting in an rSPSS of their union, or by the Cdbg-Tricks [25] to calculate the union unitigs from unitigs of mutliple original k-mer sets. However, other operations than union, such as intersection or symmetric difference have never been considered.

Here, we develop a dynamic variant of masked superstrings (and thus also of (r)SPSS) called the *f-masked superstrings (f-MS)*. The key idea is to equip masked superstrings with so-called demasking functions f for more flexible mask interpretation (**Section 3**). When complemented with the concatenation and mask recast operations (**Section 4.1**) and possibly with additional compaction (**Section 4.2**), this provides support for union, intersection, and symmetric difference and thus any set operation with k-mer sets, resulting in a complete

algebraic type for k-mer sets (**Section 4.3**). We demonstrate applicability of the concept on the FMSI index for masked superstrings [54] (**Section 5**) and provide a proof-of-concept implementation in the FMSI program (**Section 6**).

1.1 Problem formulation

We focus on the problem of developing a time- and space-efficient index for k-mer sets with support for set operations.

(i) **k-mer set index construction**, with time complexity linear with the number of k-mers.
(ii) **k-mer membership queries**, with low bits-per-k-mer memory requirements, approaching 2 bits per distinct k-mer for datasets following the Spectrum-Like Property [16] and proportionally for more complex datasets.
(iii) **Set operations**, including union, intersection, difference, and symmetric difference.

1.2 Related Work

Many works have recently focused on data structures for single k-mer sets and their collections; we refer to [16, 39] for recent surveys. Here, we primarily focus on those that are exact and offer some kind of dynamicity, i.e., an efficient support for set operations or, at least, insertions/deletions of individual k-mers. The recently introduced Conway-Bromage-Lyndon (CBL) structure [42] builds on the work of Conway and Bromage [19] on sparse bit-vector encodings and combines them with smallest cyclic rotations of k-mers (a.k.a. Lyndon words), which yields a dynamic and exact k-mer index supporting set operations, such as union, intersection, and difference, as well as insertions or deletions of k-mers.

While, to the best of our knowledge, other k-mer indexes do not support efficient set operations such as the intersection or difference, other tools, including BufBoss [1], DynamicBoss [4], and FDBG [20] allow for efficient insertions and deletions of individual k-mers. Bifrost [26], VARI-merge [43], Metagraph [30], or the very recent Cdbg-Tricks [25] support insertions but not deletions. We note that there are many more highly efficient but static data structures for individual or multiple k-mer sets, e.g., [2, 3, 5, 22, 23, 26, 36, 40, 41]. Finally, one can also use the textual representations of (r)SPSS, with efficient string indexes such as the FM index [24], BWA [32–35], or SSHash [45], but this only yields static indexes.

Another line of work, e.g. [21, 29, 31, 37, 38, 47, 48], focused on k-mer counting, where we additionally require to compute the k-mer frequencies. Out of the many k-mer counters available, GenomeTester4 [29], KMC3 [31], or Meryl [47] support operations such as union, intersection, or difference on multisets. However, k-mer counters require substantially larger memory than the most efficient k-mer indexes or heavily utilize disk.

Baseline approaches for performing set operations. We note that one can add support for set operations to a static data structure in a straightforward way: One option is extracting the k-mer sets from the input indexes, performing the given operation with the sets, and computing the new index for the resulting set; however, this process requires substantial time and memory. Another option is to keep the indexes for input k-mer sets and process a k-mer query on the set resulting from the operation by asking each index for the presence/absence of the k-mer in each input set, which is however time and memory inefficient as all the indexes need to be loaded into memory, as also noted in [27]. Therefore, we seek to perform set operations without the costly operation of recomputing the index or making multiple queries to original indexes.

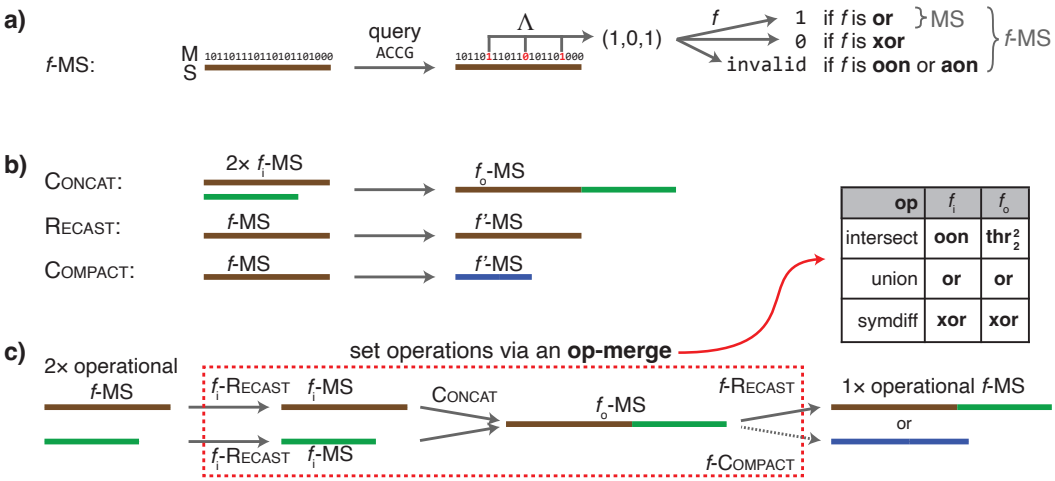

Figure 1 Overview of the f-MS framework. a) The concept of f-MS. For a given k-mer, the corresponding mask bits $\Lambda(S, M, Q)$ are evaluated using a demasking function f. **b) Low-level operations.** A $f \to f'$ Recast changes mask under function f to another mask under function f', while preserving the represented k-mer set. Concat merges two superstrings and masks. Both operations may conceptually be performed either on the original f-MS or on their associated indexes. **c) Set operations.** An operation **op** is performed by a sequence of Concat and Recast applied to the input f-MS, with operation-specific input and output functions (see Tab. 1). The Recast, which serves for keeping the f-MS operational for its data structure, may be replaced by compaction with the same target function.

2 Preliminaries

Strings and k-mers. We use constant-size alphabets Σ, typically the nucleotide alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ (unless stated otherwise). The set Σ^* contains all finite strings over Σ, with ϵ representing the empty string. For a given string $S \in \Sigma^*$, $|S|$ denotes its length, and $|S|_c$ the number of occurrences of the letter c in S. For two strings S and T, $S + T$ denotes their concatenation. A *k-mer* is a k-long string over Σ. For a string S and a fixed length k, the k-mers *generated* by S are all k-long substrings of S, and similarly for a set of strings $\mathcal{R}$, they are those generated by the individual strings $S \in \mathcal{R}$. We present our framework in the uni-directional model, where all k-mers are treated as distinct (unless stated otherwise).

Masked superstrings. Given a k-mer set K, a *masked superstring* (MS) [53] consists of a pair (S, M), where S is an arbitrary superstring of the k-mers in K and M is a binary mask of the same length. An occurrence of a k-mer in an MS is said to be on if there is 1 at the corresponding position in the mask (i.e., the initial position of the occurrence), and off otherwise. The set of k-mers generated by S are referred to as the *appearing k-mers*, and only those that have at least one on occurrence are *represented k-mers*, i.e. those from the represented set K. All masks M that represent a given K in a combination with a given superstring are called *compatible*.

Occurrence function. Let (M, S) be an MS and suppose that our objective is to determine whether a given k-mer Q is among the MS-represented k-mers. Conceptually, this process consists of two steps: (1) identify the occurrences of Q in S, and (2) verify using the mask M whether at least one occurrence of Q is on. We can formalize this process via a so-called *occurrence function*.

Fn f	Definition	Comprh.	Use cases
or	`1` if $\|\Lambda\|_1 > 0$ `0` if $\|\Lambda\|_1 = 0$	yes	• f for MS (Sec. 3), (r)SPSS (App. A) • f_i, f_o for union (Sec. 4.3)
xor	`1` if $\|\Lambda\|_1$ is odd `0` if $\|\Lambda\|_1$ is even	yes	• f_i, f_o for sym. difference (Sec. 4.3)
$\mathbf{thr}_a^b$	`1` if $a \leq \|\Lambda\|_1 \leq b$ `0` otherwise	iff $a = 1$	• $[a, b]$-threshold • f_o for intersection (Sec. 4.3)
oon	`1` if $\|\Lambda\|_1 = 1$ `0` if $\|\Lambda\|_1 = 0$ `invalid` otherwise	yes	• One-or-Nothing (corresponds to min-1 masks) • f in FMSI-dict queries (Sec. 3) • f_i for intersection (Sec. 4.3)
aon	`1` if $\Lambda \neq \epsilon \wedge \|\Lambda\|_0 = 0$ `0` if $\|\Lambda\|_1 = 0$ `invalid` otherwise	yes	• All-or-Nothing (corresponds to max-1 masks) • f in FMSI-memb queries (Sec. 3)

Table 1 Overview of demasking functions. $\Lambda(f, S, M)$ is abbreviated as Λ. Only comprehensive functions can be a target of recasting.

▶ **Definition 1.** *For a superstring S, a mask M, and a k-mer Q, the* occurrence function $\Lambda(S, M, Q) \to \{0, 1\}^*$ *is a function returning a finite binary sequence with the mask symbols of the corresponding occurrences, i.e.,*

$$\Lambda(S, M, Q) := \left(M_i \,\middle|\, S_i \cdots S_{i+k-1} = Q\right) . \tag{1}$$

In this notation, verifying k-mer presence corresponds to evaluating the composite function '$\mathbf{or} \circ \Lambda$'; that is, a k-mer is present if $\Lambda(S, M, Q)$ is non-empty and the logical **or** operation on these values yields 1. Thus, the set of all MS-represented k-mers can be expressed as

$$K = \{Q \in \Sigma^k \,|\, \mathbf{or}(\Lambda(S, M, Q)) = 1\}. \tag{2}$$

▶ **Example 2.** Consider the k-mer set $K = \{\texttt{ACG}, \texttt{GGG}\}$. One possible superstring is `ACGGGG`, with three compatible masks: `101100`, `100100`, `101000`, resulting in the mask-cased encodings `AcGGgg`, `AcgGgg`, `AcGggg`, respectively. Given a k-mer `GGG` and masked superstring `AcgGgg`, $\Lambda(S, M, Q) = (0, 1)$ and $\mathbf{or}((0, 1)) = 1$, therefore `GGG` is considered represented.

3 Function-Assigned Masked Superstrings

This work is based on two important observations on masked superstrings in the context of their applications for k-mer-set indexing.

First, **or** is one member of a large class of functions that could be used to demask k-mers in masked superstrings: for instance, MS could have been defined using the **xor** function, with a k-mer considered present if the number of its ON occurrences is odd. In fact, any symmetric Boolean function k-mer demasking can serve the role.

Second, different data structures may impose different constrains on f-masked superstrings, and this can be embedded directly into demasking functions via a special return value, `invalid`. Additionally, we treat the non-appearing k-mers as non-represented by requiring $f(\varepsilon) = 0$.

▶ **Definition 3.** *We call a symmetric function $f : \{0, 1\}^* \to \{0, 1, \texttt{invalid}\}$ with $f(\varepsilon) = 0$ a k-mer demasking function.*

We now generalize the concept of MS into so-called f-masked superstrings (f-MS).

▶ **Definition 4.** *Given a demasking function f, a superstring S, and a binary mask M, such that $|M| = |S|$, we call a triplet $\mathcal{S} = (f, S, M)$ a* function-assigned masked superstring *or* f-masked superstrings*, and abbreviate it as* f-MS*.*
If $f(\Lambda(S, M, Q)) = \texttt{invalid}$ for any k-mer Q, we call the f-MS invalid*.*

Now, for a valid f-MS, we generalize Equation (2) for k-mer decoding as

$$K = \{Q \in \Sigma^k \mid f(\Lambda(S, M, Q)) = 1\}. \tag{3}$$

We note that by the validity requirement, f imposes structural constraints on the valid masks, e.g., **oon** requires that only a single occurrence of a reprensented k-mer is ON and all other occurrences are OFF. Still, there may be many valid masks, leaving room for an application-specific mask optimization.

▶ **Example 5.** Consider Example 2 with the set of 3-mers $K = \{\texttt{ACG}, \texttt{GGG}\}$ and the masked superstring `AcGGgg`. For the query k-mer $Q = \texttt{GGG}$, the occurrence function gives $\Lambda(S, M, Q) = (1, 1)$. The result of demasking $f(\Lambda(S, M, Q))$ then depends on the specific function f: for **or** it evaluates to 1 and thus `GGG` is represented; however, for **xor** the result would be 0 and `GGG` would not be represented.

Operational f-MS with respect to different applications. Different use cases may require different properties of masks. Since the choice of a demasking function can be also viewed as a restriction of the set of acceptable masks for a given superstring and k-mer set, we can choose different demasking function to enforce certain properties of the masks. We call the demasking functions that enforce the desired properties of masks as *operational*. The concept of operational demasking functions plays a central role in the whole f-MS framework, see Figure 1.

As an example, considering the FMSI index [54], for its version for membership queries, only the **aon** function is operational. This is due to a certain query speed optimization, If this optimization is omitted, all demasking functions are operational. For the dictionary version, only the **oon** function is operational. For more details, see Section 5.

4 Abstract algebraic framework for k-mer set operations

The dynamization of a data structure can be conceptually split into two parts – the dynamization of the representation and the dynamization of the data structure internals itself. Here, we give a general approach to dynamize the representation, based on two low-level operations for their manipulation – mask recasting and concatenation, where we assume that the concatenation in the data structure can be done efficiently. In the context of different data structures, these can have very different forms, and we discuss a specific variant for BWT in Section 5.

4.1 Concat and Recast: The Two Essential Low-Level Operations

Here, we describe two elementary operations, concatenation and recasting, that we use to perform set operations, as described in Section 4.3. While concatenation effectively merges two f-MS, recasting facilitates adjusting f-MS into the desired form, e.g., before executing a set operation or to make it operational after performing the operation.

4.1.1 Concat

For full generality, we define concatenation on f-MS as concatenating the underlying superstrings and masks for all possible input and output functions f.

▶ **Definition 6.** *Given f-MS (f_i, S_1, M_1) and (f_i, S_2, M_2), we define (f_i, f_o)-concatenation as the operation taking these two f_i-MS and producing the result $(f_o, S_1 + S_2, M_1 + M_2)$.*

Note that Definition 6 can be easily extended to more than two input f-MS. In the case that all the functions are the same, i.e., $f = f_i = f_o$, we call it *f-concatenation* or just *concatenation* if f is obvious from the context.

Furthermore, note that while the set of appearing k-mers of $S_1 + S_2$ clearly contains the union of appearing k-mers of S_1 and of S_2, additional new occurrences of k-mers may appear at the boundary of the two superstrings. These newly occurring k-mers may not be appearing in any of the superstrings S_1 and S_2. The occurrences of appearing k-mers of $S_1 + S_2$ that overlap both input superstrings are called *boundary occurrences*.

▶ **Example 7.** Consider **aon**-MS `AcGGgg`, representing 3-mers $K_1 = \{\texttt{ACG}, \texttt{GGG}\}$, and **aon**-MS `Ggg`, representing $K_2 = \{\texttt{GGG}\}$. Their **aon**-concatenation is `AcGGggGgg`, with two boundary occurrences of GGG; causing the result to be an invalid **aon**-MS even though the input **aon**-MS were valid. Similarly, boundary occurrences of k-mers can change the presence in the result for some functions, illustrating the need of a careful treatment of concatenation.

4.1.2 Recast

When performing operations with f-MS, recasting is used to change the demasking function f without altering the represented k-mer set and the underlying superstring; here, we note that recomputing the superstring is the most resource-demanding operation [53].

In many cases, recasting triggers recomputation of the mask, which is however a less expensive operation. We first note that recasting to an arbitrary demasking function may not be possible with a fixed superstring as there is no valid mask, e.g., for a function that requires the represented k-mers to have at least two ON occurrences. Nevertheless, recasting is possible for a large class of demasking functions, specifically functions with the property that all appearing k-mers are encodable using a compatible mask as present or absent, irrespective of the number of their occurrences; we call such functions *comprehensive*:

▶ **Definition 8.** *A demasking function f is* comprehensive *if for every $n > 0$, there exist $x, y \in \{0,1\}^n$ such that $f(x) = 0$ and $f(y) = 1$.*

The definition directly implies the possibility of recasting an f-MS by changing the function f to any comprehensive f'. This also implies an efficient initialization procedure to create a desired operational f-MS for a comprehensive f: First, compute an **or**-MS as described in [53] and then recast it to f.

▶ **Lemma 9** (recasting)**.** *Let (f, S, M) be a valid f-MS representing a k-mer set K. Then, for every comprehensive demasking function f', there exists a valid mask M' such that (f', S, M') represents K.*

Proof. Since $f(\epsilon) = 0$ for any demasking function f, we know that non-appearing k-mers are not represented in the f-MS (f, S, M) and will correctly be non-represented in the resulting f'-MS. For any appearing k-mer Q in S, we find all of its $n_Q > 0$ occurrences and then use Definition 8 to set the mask bits of M' at these occurrences either to x with $f(x) = 0$ if $Q \notin K$, or to y with $f(y) = 1$ otherwise (since f is symmetric by Definition 3, the order does not matter). ◀

The proof gives a general algorithm for recasting to any comprehensive function as for each appearing k-mer we can check whether it is represented in the f-MS of origin, and find the number of ON occurrences for it to be represented (or not) in the new f-MS. Moreover, for all comprehensive functions mentioned in Table 1, recasting can be done either by maximizing the number of 1s in the mask (**aon**), or by minimizing the number of 1s (all other comprehensive functions in Table 1). Both can be achieved in linear time using a two- or single-pass algorithms, respectively, as described in [53].

4.2 Compact: an Optional Low-Level Operation

After performing a number of set operations, the resulting superstring may be much longer than the size of the k-mer set it represents. Then it is desirable to change also the superstring alongside recasting the mask. We call this operation of changing (f, M, S) into (f', M', S') while preserving the represented set K *compaction*, and typically require $|S'| < |S|$. Lemma 9 implies that if f' is comprehensive we can use any superstring of K as our S', and therefore compact can be performed by any algorithm for superstring computation. We give an example of how compaction can be performed on a BWT-based index in Section 5.

4.3 k-Mer Set Operations via f-MS and its Low-Level Operations

Union. As implicitly shown in [53], concatenating MS, which are **or**-MS in our notation, acts as union on the represented sets, that is, the resulting represented set is the union of the original represented sets. Specifically, the boundary occurrences of k-mers do not affect the result by the definition of **or**. This allows **or**-MS to generalize (r)SPSS representations, since any set of k-mers in the (r)SPSS representation can be directly viewed as an **or**-MS by concatenating the individual simplitigs/matchtigs.

We show that **or** is the only comprehensive demasking function that acts as union on the represented sets, up to the interchange of the meaning of 0s and 1s; see Appendix A for details. We further demonstrate this uniqueness even on the level of matchtigs and therefore, **or**-MS are the only f-MS that generalize (r)SPSS representations.

Symmetric difference. Next, we observe that **xor** naturally acts as the symmetric difference set operation, i.e., concatenating two **xor**-MS results in a **xor**-MS representing the symmetric difference of the original sets. Indeed, recall that using **xor** implies that a k-mer is represented if and only if there is a odd number of ON occurrences of that k-mer. Observe that the boundary occurrences of k-mers do not affect the resulting represented set as those have zeros in the mask. Thus, if a k-mer is present in both sets, it has an even number of ON occurrences in total and hence, is not represented in the result. Likewise, if a k-mer belongs to exactly one input set, it has an odd number of ON occurrences in this input set and an even number (possibly zero) in the other; thus, it is represented in the result. As any appearing k-mer is either boundary or appears in one of the input MS, the result corresponds to the symmetric difference.

Intersection. After seeing functions for union and symmetric difference operations, it might seem natural that there is a comprehensive function f such that f-concatenation yields intersection. This is however not the case as there exists no comprehensive demasking function acting as intersection, which we show in Appendix B. We can circumvent the non-existence of a single demasking function acting as intersection by using a different function on the output than on the input. To this end, we will need two different demasking functions:

- The **oon** (abbreviation of one-or-nothing) function is a demasking function that returns 1 if there is exactly one 1 in the input, 0 if there are no 1s, and `invalid` if there is more than a single ON occurrence of the k-mer.
- The $\mathbf{thr}_a^b$ function (an abbreviation of threshold), where $0 < a \leq b$, is a demasking function that returns 1 whenever it receives an input of at least a ones and at most b ones and 0 otherwise. Note that unless $a = 1$, $\mathbf{thr}_a^b$ functions are not comprehensive. The corresponding f-MS are denoted $\mathbf{thr}_a^b$ -MS.

We use **oon**-MS to represent the input masked superstrings and treat the result of the concatenation as $\mathbf{thr}_2^2$-MS, that is, we apply (**oon**, $\mathbf{thr}_2^2$)-concatenation. Since the represented k-mers are those that have precisely one ON occurrence in the original **oon**-MS, they are those with two ON occurrences in the result and therefore are correctly recognized by the $\mathbf{thr}_2^2$ function.

Further generalizations and optimization. Other set operations can be performed by chaining intersection, union, and symmetric difference. For example, the asymmetric difference $A - B$ can be obtained as $A\Delta(A \cap B)$. Such an approach is possible for *any set operation* on *any number of input sets*.

Furthermore, the approach with **oon**-MS for intersection can be utilized to obtain more efficient computation of any symmetric operation for any number of input sets. If we represent each input as **oon**-MS, then after concatenation of all of them, the number of ON occurrences counts the number of sets, where the k-mer appeared and so the symmetric set operation can be recognized with appropriate demasking function. Furthermore, the same such mask and superstring can be used for different operations, e.g., for intersection and union at once, just with different f.

Finally, there are more demasking functions that could be used as operational in various other use cases. For instance, the **and** function, for which an appearing k-mer is represented if all of its occurrences are ON, could be used to allow ON occurrences of non-represented k-mers, potentially making the masks of **and**-MS better compressible.

5 Example Application: Dynamization of the FMSI Index

The f-MS framework may be used for the dynamization of many different data structures. Here, we will demonstrate such a procedure on the FMSI index introduced in our concurrent work [54]. The FMSI index [54] for a k-mer set K is constructed from its masked superstring (M, S) either maximizing the number of ones for membership queries (FMSI-memb), or minimizing them for dictionary queries (FMSI-dict). The FMSI index consists of the Burrows-Wheeler transform (BWT) [14] of S with an associated rank data structure [28], and the *SA-transformed mask* M' which is a bit-vector of the same length as S, where $M'[i] = M[j_i - 1 \mod |M|]$ where j_i is the starting position of the lexicographically i-th suffix of S. FMSI can be constructed in linear time through Masked BWT [54], a tailored variant of the classical BWT [14]. FMSI can index a k-mer Q in $O(k)$ time by first computing the range of occurrences of Q in the suffix-array coordinates. Here, we only consider the most memory-efficient version of FMSI index, which requires $2 + o(1)$ bits of memory per distinct k-mer under the spectrum-like property and at most $3 + o(1)$ bits per superstring character in the general case [54]. In addition, we consider the rank data structure also for the SA-transformed mask, which does not asymptotically increase complexity, that is, costs only another $o(1)$ bits per superstring character [28].

5.1 Step 1: Extending the FMSI index to f-MS

The first step to an f-MS dynamization of a data structure is to identify its operational demasking function. For FMSI-memb and FMSI-dict, this is **aon** and **oon**, respectively, but the implementation can be generalized for any demasking function to be operational. Afterwards, we simply implement the evaluation of the operational demasking functions, possibly with some pre- and post-processing which may be specific to the operation and to the internals of the data structure.

Efficient membership queries with arbitrary demasking functions. As FMSI-memb retrieves only the mask symbol at an arbitrary occurrence, it requires all the symbols of a k-mer to be the same. Thus, the only comprehensive operational f-MS is **aon**-MS, which returns 1 if it receives a list of ones, 0 is a list without a one, and `invalid` otherwise (see Table 1). This f-MS corresponds to **or**-MS with maximized number of ones.

However, the search in FMSI can be easily extended to return all occurrences of a k-mer. In such case, all possible demasking functions are operational. In Lemma 10, we demonstrate that if the function can be evaluated efficiently, k-mer presence in FMSI can be determined in constant time after performing backwards search.

▶ **Lemma 10.** *Consider a query for k-mer Q on an f-MS (f, S, M) representing a k-mer set K, such that f can be computed in $O(1)$ from the number of* ON *and* OFF *occurrences of Q. Let M' be the corresponding SA-tranformed mask [54] and let $[i, j)$ be the range of sorted rotations of S starting with a k-mer Q. Then the presence or absence of Q in K can be determined in $O(1)$ time.*

Proof. From [54, Lemma 1], $M'[x]$ for $x \in [i, j)$ corresponds to the mask symbol of a particular occurrence of Q. Therefore, $|\Lambda(S, M, Q)|_1 = \mathrm{rank}_1(M', j) - \mathrm{rank}_1(M', i)$, which can be computed in $O(1)$ time using two rank queries on the mask; here, $\mathrm{rank}_1(M', i) = \sum_{a=0}^{i-1} M'[a]$ is the number of ones on coordinates $0, \ldots, i-1$ in M', computed by the rank data structure. Furthermore, $|\Lambda(S, M, Q)|_0 = |\Lambda(S, M, Q)| - |\Lambda(S, M, Q)|_1 = j - i - |\Lambda(S, M, Q)|_1$. Since f is commutative, $f(\Lambda(S, M, Q))$ can be evaluated from the two quantities in constant time. ◀

Dictionary queries. For dictionary queries, FMSI requires the mask to contain only a single ON occurrence of a k-mer. This exactly corresponds to **oon**-MS (introduced in Section 4.3, see Table 1). After recasting to this f-MS, dictionary queries function in the same way as in FMSI-dict by retrieving the rank in the SA-transformed mask of the first lexicographic occurrence [54].

5.2 Step 2: Adding Concat and Recast

Recast. Recasting of an f-MS can be done directly in SA coordinates without additional memory. For each position of BWT that was not previously visited, we first retrieve the k-mer by traversing the inverse LF mapping [54, Alg. 3] and then obtain the range of this k-mer. Then from Lemma 9, we find the number of ones needed to make this k-mer represented if it was represented in f-MS of origin. This takes time $O(|S| + k|K'|)$, where K' are all appearing k-mers. Then, we need to update the auxiliary data structures (rank and select), possibly by recomputing them.

Concat. Performing the operation on indexes boils down to merging two BWTs using any algorithm for BWT merging, for example [27] which runs in linear time. To merge the SA-tranformed masks, we attach the mask symbols to the corresponding characters of BWT,

in the same way as for FMSI construction [54, Alg. 1] and perform BWT merge; hence, the existing algorithms for BWT merging can be adjusted in a straightforward way to merge the SA-tranformed masks. The BWT merging is then followed by recomputation of the ranks necessary for the functioning of FMSI.

5.3 Step 3 (optional): Adding Compact

If an f-MS contains too many redundant copies of individual k-mers, e.g., if an f-MS is obtained by concatenating multiple input f-MS, it might be desirable to *compact* it, i.e., reoptimize its support superstring. This can be done by recasting to **oon**-MS, exporting this **oon**-MS to its superstring form by reversing the LF-mapping, removing all segments of support superstring not containing any ON occurrence of a k-mer, which can be done by single-pass algorithm, and reindexing the result. Although this algorithm requires reindexing of the result, it uses memory linear in the total size of the input superstrings. We leave it to future work to design more efficient algorithms for compaction directly in the FMSI index without the need to export the f-MS.

6 Implementation and Experimental Evaluation

To demonstrate feasibility of using f-MS for set operations in practice, we developed a prototype implementation of the framework, embedded it into the FMSI k-mer-set index [54], and evaluated it using two selected genomes.

6.1 Implementation in the FMSI tool

We implemented the f-MS framework as an experimental extension of the FMSI program [54] (`https://github.com/OndrejSladky/fmsi`). In its basic form, FMSI takes an input masked superstring, constructs an FMSI index over it and enables performing membership and dictionary k-mer queries. For membership queries, we extended the original FMSI implementation [54, Alg. 3] to reflect the new Lemma 10. Dictionary queries are supported as in the original implementation from [54].

Set operations follow the workflow described in Figure 1. The CONCAT operation is implemented via FMSI index merging: first, by exporting the individual masked superstrings, then, by concatenating them, and finally by reindexing the newly obtained masked superstring result, with the f_i, f_o functions tracked by the user. The RECAST operation is currently supported for **or**, **xor**, **oon**, and **aon**, and proceeds via three steps: exporting the masked superstring by FMSI, mask re-optimization by KmerCamel [53] (min-1 for **or**, **xor**, **oon**, and max-1 for **aon**), and reindexing the new f-masked superstring by FMSI. The COMPACT operation is supported directly in FMSI, and internally runs global greedy algorithm [53] to approximate shortest superstring (in the case of FMSI-dict, followed additionally RECAST to **aon** to stay operational).

6.2 Experiments using selected genomes

We evaluated our framework using the *C. elegans* and *C. briggsae* genomes, with a primary focus on the memory requirements for processing membership queries. We started by indexing the initial **or**-MS is obtained by KmerCamel's global greedy algorithm [53]. Then, we constructed the indexes for source genomes with $k = 21$, and tested the union, intersection, and symmetric difference operations of the corresponding k-mer sets. Finally, we evaluated

the associated quantitative characteristics, including the superstring length of each computed f-MS, and the memory requirements to perform queries on the indexed individual and concatenated f-MS, both without and with compaction (measured by GNU time). and also compared FMSI to CBL [42] (commit `328bcc6`, 28 prefix bits). More details are provided in Appendix C.

The starting memory requirements of membership queries of the FMSI and CBL indexes were approximately 2.7 and 83 bits per distinct k-mer, respectively, for both of the genomes. After computing the union of both of the k-mer sets by FMSI index concatenation, the memory requirements of FMSI per distinct k-mer decreased slightly to about 2.5 bits per distinct k-mer in the union. For CBL, the memory requirements dropped to about 59 bits per distinct k-mer, still requiring over 20 times more than FMSI. In case of symmetric difference, the results are very similar, with FMSI requiring 2.6 bits per distinct k-mer and CBL 59 bits per distinct k-mer.

However, for intersection, which contains only less than 1% of the original k-mers, the memory consumption increases significantly to 524 bits per distinct k-mer for FMSI, even underperforming CBL whose performance increased to about 420 bits per distinct k-mer. This is caused by the resulting superstring length being much larger than the represented set, in which case compaction is necessary.

After compaction, the memory requirements of FMSI decreased to 43 bits per k-mer, almost 10 times less than CBL. Although this case necessitates compaction for FMSI, the compaction here is relatively cheap as it has to be done only if the result has significantly less k-mers than the original set. We also illustrate this case in Figure 2.

Overall, this experiment demonstrates that if the percentage of k-mers in the result is high, FMSI maintains its superior memory efficiency over CBL even after index merging, without the need for compaction. If the percentage of k-mers in the result is very low, compaction needs to be applied in order for FMSI to maintain low memory footprint.

7 Conclusion and Outlook

We have proposed f-masked superstrings (f-MS) as an algebraic data type for k-mer sets that allows for seamless execution of set operations. It is primarily based on equipping masked superstrings (MS) from [53] with a demasking function f, and we have thoroughly investigated several natural demasking functions, demonstrating that set operations on k-mer sets can be carried out simply by masked superstring concatenation or, if indexed, by merging their masked Burrows-Wheeler transform from [54]. This leads to a simple data structure that simultaneously allows for beyond worst-case compressibility, answering exact membership queries, and efficiently performing set operations on the k-mer sets, without the costly operation of recomputing the underlying representation. Another major advantage is the versality of our concept as it can in fact be combined with (repetitive) Spectrum Preserving String Sets [10, 46, 52] instead of (more general) masked superstrings. For instance, our approach can be applied to SPSS-based indexes, such as SSHash – simplitigs of k-mers split based on minimizers can be treated as f-MS and the same Concat and Recast operations can be performed on those to achieve dynamicity.

The main practical limitation of our work is the current implementation of the index merging, which is very slow and not using the state-of-the-art algorithms for BWT merging [27]. Furthermore, our proof-of-concept experiment is only meant to demonstrate feasibility, and we leave a more thorough evaluation, using various datasets and including a thorough comparison to other tools for set operations, such as CBL [42], to future work.

Our work opens up several research directions for future theoretical investigation. On the algorithmic level, our work relies on efficient algorithms for merging the Burrows-Wheeler transform. Additionally, we seek an algorithm for mask recasting that is not only very space-efficient, but also has faster running time, which can possibly be achieved by using additional data structures for the index, such as the kLCP array [49]. Moreover, it is open how to directly perform compaction with f-masked superstrings indexed with the masked Burrows-Wheeler transform [54], that does not necessitate to reverse the BWT and export the masked superstring, but rather work locally with the BWT of the superstring and the SA-transformed mask. Furthermore, as our work deals only with set operations, we open the question of performing single insertions and deletions in a more efficient way than performing these through set operations; note that for comprehensive demasking functions, deletions can be handled efficiently by changing the corresponding mask bits.

In conclusion, while the primary contributions of this paper are conceptual, they pave the path towards a space- and time-efficient library for k-mer sets that would include all of these features. In the light of advances in efficient superstring approximation algorithms and BWT-based indexing, we believe that the f-masked superstring framework is a useful step towards designing appropriate data structures for this library, which is now mainly an engineering challenge.

References

1. Jarno Alanko, Bahar Alipanahi, Jonathen Settle, Christina Boucher, and Travis Gagie. Buffering updates enables efficient dynamic de bruijn graphs. *Computational and structural biotechnology journal*, 19:4067–4078, 2021.
2. Jarno N Alanko, Simon J Puglisi, and Jaakko Vuohtoniemi. Small searchable κ-spectra via subset rank queries on the spectral Burrows-Wheeler transform. In *SIAM Conference on Applied and Computational Discrete Algorithms (ACDA23)*, pages 225–236. SIAM, 2023. doi:10.1137/1.9781611977714.20.
3. Jarno N Alanko, Jaakko Vuohtoniemi, Tommi Mäklin, and Simon J Puglisi. Themisto: a scalable colored k-mer index for sensitive pseudoalignment against hundreds of thousands of bacterial genomes. *Bioinformatics*, 39(Supplement_1):i260–i269, 2023. doi:10.1093/bioinformatics/btad233.
4. Bahar Alipanahi, Alan Kuhnle, Simon J Puglisi, Leena Salmela, and Christina Boucher. Succinct dynamic de bruijn graphs. *Bioinformatics*, 37(14):1946–1952, 2021.
5. Alexander Bowe, Taku Onodera, Kunihiko Sadakane, and Tetsuo Shibuya. Succinct de Bruijn graphs. In Benjamin J. Raphael and Jijun Tang, editors, *Algorithms in Bioinformatics - 12th International Workshop, WABI 2012, Ljubljana, Slovenia, September 10-12, 2012. Proceedings*, volume 7534 of *Lecture Notes in Computer Science*, pages 225–235. Springer, 2012. doi:10.1007/978-3-642-33122-0_18.
6. Phelim Bradley, Henk C Den Bakker, Eduardo PC Rocha, Gil McVean, and Zamin Iqbal. Ultrafast search of all deposited bacterial and viral genomic data. *Nature Biotechnology*, 37(2):152–159, 2019.
7. Phelim Bradley, N Claire Gordon, Timothy M Walker, Laura Dunn, Simon Heys, Bill Huang, Sarah Earle, Louise J Pankhurst, Luke Anson, Mariateresa De Cesare, et al. Rapid antibiotic-resistance predictions from genome sequence data for staphylococcus aureus and mycobacterium tuberculosis. *Nature Communications*, 6(1):10063, 2015. doi:10.1038/ncomms10063.

8. Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, 2016. doi:10.1038/nbt.3519.
9. Karel Břinda. Novel computational techniques for mapping and classification of Next-Generation Sequencing data. PhD thesis, Université Paris-Est, 2016. doi:10.5281/zenodo.1045317.
10. Karel Břinda, Michael Baym, and Gregory Kucherov. Simplitigs as an efficient and scalable representation of de Bruijn graphs. *Genome Biology*, 22(96), 2021. doi:10.1186/s13059-021-02297-z.
11. Karel Břinda, Alanna Callendrello, Kevin C Ma, Derek R MacFadden, Themoula Charalampous, Robyn S Lee, Lauren Cowley, Crista B Wadsworth, Yonatan H Grad, Gregory Kucherov, et al. Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nature Microbiology*, 5(3):455–464, 2020. doi:10.1038/s41564-019-0656-6.
12. Karel Břinda, Leandro Lima, Simone Pignotti, Natalia Quinones-Olvera, Kamil Salikhov, Rayan Chikhi, Gregory Kucherov, Zamin Iqbal, and Michael Baym. Efficient and robust search of microbial genomes via phylogenetic compression. *bioRxiv*, 2023.04.15.536996, 2023. doi:10.1101/2023.04.15.536996.
13. Karel Břinda, Kamil Salikhov, Simone Pignotti, and Gregory Kucherov. Prophyle 0.3.1.0. *Zenodo*, 5281, 2017. URL: https://prophyle.github.io, doi:10.5281/zenodo.5237391.
14. Michael Burrows and David Wheeler. A block-sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation, 1994.
15. Rayan Chikhi. K-mer data structures in sequence bioinformatics. HDR thesis, Institut Pasteur Ecole Doctorale "EDITE", 2021. URL: http://rayan.chikhi.name/pdf/RChikhi-HDR.pdf.
16. Rayan Chikhi, Jan Holub, and Paul Medvedev. Data structures to represent a set of k-long DNA sequences. *ACM Computing Surveys*, 54(1):17:1–17:22, 2022. doi:10.1145/3445967.
17. Rayan Chikhi, Antoine Limasset, Shaun Jackman, Jared T. Simpson, and Paul Medvedev. On the representation of de Bruijn graphs. In Roded Sharan, editor, *Research in Computational Molecular Biology*, pages 35–55, Cham, 2014. Springer International Publishing. doi:10.1007/978-3-319-05269-4_4.
18. Rayan Chikhi, Antoine Limasset, and Paul Medvedev. Compacting de Bruijn graphs from sequencing data quickly and in low memory. *Bioinformatics*, 32(12):i201–i208, 2016. doi:10.1093/bioinformatics/btw279.
19. Thomas C. Conway and Andrew J. Bromage. Succinct data structures for assembling large genomes. *Bioinformatics*, 27(4):479–486, 2011. doi:10.1093/bioinformatics/btq697.
20. Victoria G. Crawford, Alan Kuhnle, Christina Boucher, Rayan Chikhi, and Travis Gagie. Practical dynamic de Bruijn graphs. *Bioinformatics*, 34(24):4189–4195, 2018. doi:10.1093/bioinformatics/bty500.
21. Sebastian Deorowicz, Agnieszka Debudaj-Grabysz, and Szymon Grabowski. Disk-based k-mer counting on a pc. *BMC bioinformatics*, 14:1–12, 2013.
22. Jason Fan, Jamshed Khan, Giulio Ermanno Pibiri, and Rob Patro. Spectrum preserving tilings enable sparse and modular reference indexing. In Haixu Tang, editor, *Research in Computational Molecular Biology - 27th Annual International Conference, RECOMB 2023, Istanbul, Turkey, April 16-19, 2023, Proceedings*, volume 13976 of *Lecture Notes in Computer Science*, pages 21–40. Springer, 2023. doi:10.1007/978-3-031-29119-7_2.

23. Jason Fan, Jamshed Khan, Noor Pratap Singh, Giulio Ermanno Pibiri, and Rob Patro. Fulgor: a fast and compact k-mer index for large-scale matching and color queries. *Algorithms for Molecular Biology*, 19(1):3, 2024. doi:10.1186/S13015-024-00251-9.
24. Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, SFCS-00. IEEE Comput. Soc, 2000. doi:10.1109/sfcs.2000.892127.
25. Khodor Hannoush, Camille Marchet, and Pierre Peterlongo. Cdbgtricks: Strategies to update a compacted de bruijn graph. *bioRxiv*, 2024. URL: https://www.biorxiv.org/content/early/2024/05/28/2024.05.24.595676, arXiv:https://www.biorxiv.org/content/early/2024/05/28/2024.05.24.595676.full.pdf, doi:10.1101/2024.05.24.595676.
26. Guillaume Holley and Páll Melsted. Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs. *Genome Biology*, 21(1):1–20, 2020. doi:10.1186/s13059-020-02135-8.
27. James Holt and Leonard McMillan. Merging of multi-string bwts with applications. *Bioinformatics*, 30(24):3524–3531, August 2014. doi:10.1093/bioinformatics/btu584.
28. Guy Joseph Jacobson. *Succinct static data structures*. PhD thesis, Carnegie Mellon University, USA, 1988. AAI8918056.
29. Lauris Kaplinski, Maarja Lepamets, and Maido Remm. GenomeTester4: a toolkit for performing basic set operations - union, intersection and complement on k-mer lists. *GigaScience*, 4(1):s13742–015–0097–y, 12 2015. doi:10.1186/s13742-015-0097-y.
30. Mikhail Karasikov, Harun Mustafa, Daniel Danciu, Marc Zimmermann, Christopher Barber, Gunnar Rätsch, and André Kahles. Indexing all life's known biological sequences. *bioRxiv*, 2020.10.01.322164, 2024. doi:10.1101/2020.10.01.322164.
31. Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 05 2017. doi:10.1093/bioinformatics/btx304.
32. Heng Li. Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*, 28(14):1838–1844, 2012. doi:10.1093/bioinformatics/bts280.
33. Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013. doi:10.48550/arXiv.1303.3997.
34. Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009. doi:10.1093/bioinformatics/btp324.
35. Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010. doi:10.1093/bioinformatics/btp698.
36. Antoine Limasset, Guillaume Rizk, Rayan Chikhi, and Pierre Peterlongo. Fast and scalable minimal perfect hashing for massive key sets. In Costas S. Iliopoulos, Solon P. Pissis, Simon J. Puglisi, and Rajeev Raman, editors, *16th International Symposium on Experimental Algorithms, SEA 2017, June 21-23, 2017, London, UK*, volume 75 of *LIPIcs*, pages 25:1–25:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017. doi:10.4230/lipics.sea.2017.25.
37. Abdullah-Al Mamun, Soumitra Pal, and Sanguthevar Rajasekaran. Kcmbt: ak-mer counter based on multiple burst trees. *Bioinformatics*, 32(18):2783–2790, 2016.
38. Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

39. Camille Marchet, Christina Boucher, Simon J Puglisi, Paul Medvedev, Mikaël Salson, and Rayan Chikhi. Data structures based on k-mers for querying large collections of sequencing data sets. *Genome Research*, 31(1):1–12, 2021. doi:10.1101/gr.260604.119.
40. Camille Marchet, Zamin Iqbal, Daniel Gautheret, Mikaël Salson, and Rayan Chikhi. REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. *Bioinformatics*, 36(Supplement-1):i177–i185, 2020. doi:10.1093/bioinformatics/btaa487.
41. Camille Marchet, Maël Kerbiriou, and Antoine Limasset. Blight: efficient exact associative structure for k-mers. *Bioinformatics*, 37(18):2858–2865, 2021. doi:10.1093/bioinformatics/btab217.
42. Igor Martayan, Bastien Cazaux, Antoine Limasset, and Camille Marchet. Conway–bromage–lyndon (cbl): an exact, dynamic representation of k-mer sets. *Bioinformatics*, 40(Supplement_1):i48–i57, 06 2024. arXiv:https://academic.oup.com/bioinformatics/article-pdf/40/Supplement_1/i48/58354678/btae217.pdf, doi:10.1093/bioinformatics/btae217.
43. Martin D. Muggli, Bahar Alipanahi, and Christina Boucher. Building large updatable colored de Bruijn graphs via merging. *Bioinformatics*, 35(14):i51–i60, 2019. doi:10.1093/bioinformatics/btz350.
44. Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017. doi:10.1038/nmeth.4197.
45. Giulio Ermanno Pibiri. Sparse and skew hashing of K-mers. *Bioinformatics*, 38(Supplement_1):i185–i194, 2022. doi:10.1093/bioinformatics/btac245.
46. Amatur Rahman and Paul Medevedev. Representation of k-mer sets using spectrum-preserving string sets. *Journal of Computational Biology*, 28(4):381–394, 2021. PMID: 33290137. doi:10.1089/cmb.2020.0431.
47. Arang Rhie, Brian P Walenz, Sergey Koren, and Adam M Phillippy. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology*, 21:1–27, 2020.
48. Guillaume Rizk, Dominique Lavenier, and Rayan Chikhi. Dsk: k-mer counting with very low memory usage. *Bioinformatics*, 29(5):652–653, 2013.
49. Kamil Salikhov. Efficient algorithms and data structures for indexing dna sequence data. PhD thesis, Université Paris-Est, 2017.
50. Kamil Salikhov, Karel Břinda, Simone Pignotti, and Gregory Kucherov. ProPhex. https://github.com/prophyle/prophex, 2018. doi:10.5281/zenodo.1247432.
51. Sebastian Schmidt and Jarno N. Alanko. Eulertigs: minimum plain text representation of k-mer sets without repetitions in linear time. *Algorithms for Molecular Biology*, 18(1):5, 2023. doi:10.1186/s13015-023-00227-1.
52. Sebastian Schmidt, Shahbaz Khan, Jarno N. Alanko, Giulio E. Pibiri, and Alexandru I. Tomescu. Matchtigs: minimum plain text representation of k-mer sets. *Genome Biology*, 24(1):136, 2023. doi:10.1186/s13059-023-02968-z.
53. Ondřej Sladký, Pavel Veselý, and Karel Břinda. Masked superstrings as a unified framework for textual k-mer set representations. *bioRxiv*, 2023.02.01.526717, 2023. doi:10.1101/2023.02.01.526717.
54. Ondřej Sladký, Pavel Veselý, and Karel Břinda. FroM Superstring to Indexing: a space-efficient index for unconstrainedk-mer sets using the masked burrows-wheeler transform (MBWT). *bioRxiv*, November 2024. URL: http://dx.doi.org/10.1101/2024.10.30.621029, doi:10.1101/2024.10.30.621029.

55. Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: Astronomical or genomical? *PLoS Biology*, 13(7):e1002195, 2015. doi:10.1371/journal.pbio.1002195.

56. Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):1–12, 2014. doi:10.1186/gb-2014-15-3-r46.

A Proof of Uniqueness of Union Function

▶ **Theorem 11.** ***or*** *is the only comprehensive demasking function f with $f((1)) = 1$ such that for any two k-mer sets K and K' and any of their valid f-MS (f, S, M) and (f, S', M'), respectively, their concatenation $(f, S + S', M + M')$ is a valid f-MS representing the set $K \cup K'$.*

Proof. For a contradiction assume there is a comprehensive demasking function f different than **or** that satisfies the above. Consider the smallest n such that f behaves differently than **or** for a length-n input, meaning that there is $x \in \{0,1\}^n$ not equal to the all-zeros vector such that $f(x) \neq 1$. As f is comprehensive, it cannot happen that $f(x) = 1$ for all $x \in \{0,1\}^n$ by Definition 8, and moreover, $f((1)) = 1$, implying $n > 1$. Fix a k-mer, for simplicity $\mathtt{A}^k$ (although similar approach works for all k-mers). We take the first f-MS to be the k-mer with mask being $M_0 = x_0$ and $M_i = 0$ for the remaining $k-1$ positions $i > 0$. The second f-MS is $\mathtt{CA}\ldots\mathtt{A}$ where $\mathtt{A}$ appears $n+k-2$ times with the mask being: $M_0 = 0$, $M_i = x_i$ for $i = 1, \ldots, k$, and $M_i = 0$ for $i > k$. At least one of the represented sets contains the k-mer as $x \neq 0$ and n is the but the resulting f-MS is either invalid or does not contain the k-mer in the represented set as $f(x) \neq 1$, a contradiction. ◀

▶ **Theorem 12.** ***or*** *is the only comprehensive demasking function f with $f((1)) = 1$ such that for any sequence of f-MS, where individual superstrings are matchtigs, the concatenation of all the f-MS represents the union of represented k-mers.*

Proof. It is sufficient to find a construction of matchtigs such that we can construct an arbitrary sequence of ones and zeros at the occurrences of a given k-mer and the rest follows similarly as in Theorem 11.

We do this with k-mer `CG` and matchtigs `Cg` and `Gc`. Consider the counterexample sequence of occurring ones and zeros from Theorem 11. For every one in the sequence, we add the matchtig `Cg` and for each m consecutive zeros, we add $m+1$ times the matchtig `Gc`, since at the boundary of two `Gc` matchtigs an OFF occurrence of k-mer `CG` appears. At any other boundary, the k-mer `CG` does not appear, therefore the construction is correct. The rest of the proof follows a similar argument as in Theorem 11. ◀

Note, however, that the same does not hold if we want to represent simplitigs/SPSS solely. As an individual k-mer cannot appear more than once with an ON occurrence, any comprehensive function generalizes SPSS representations if it satisfies that if there is one ON occurrence of a k-mer, it returns 1, and if there is none, it returns 0.

B Non-Existence of a Preserved Comprehensive Function for Intersection.

▶ **Theorem 13.** *There is no comprehensive demasking function f with the property that the result of f-concatenation of two f-MS always represents the intersection of the originally represented k-mer sets.*

Proof. Let f be any comprehensive demasking function. Consider MS `A` and `C` with the appropriate mask, each representing a single 1-mer, which from coprehensiveness is always possible. Their concatenation is `AC`. For k-mer `A`, the output of occurrence function is the same in `AC` as in `A` and therefore `A` is represented in the concatenation. However, the intersection is empty and thus, f cannot be used to compute the intersection from the concatenation. ◀

Note that the proof cannot be generally extended to all demasking function as there exist non-comprehensive demasking functions acting as the intersection on the represented sets upon concatenation, for instance the constant zero function. However, since the constant zero function always represents the empty set, it is of no use in practice.

C Details of Proof-of-Concept Experiment

We conducted a simple experiment with genomes of *C. elegans* (`NC_003279.8`, 100M base pairs) and *C. briggsae* (`NC_013489.2`, 108M base pairs), performing the union, intersection, and symmetric difference on their sets of k-mers for $k \in \{17, 19, 21, 31\}$. We evaluated the FMSI program [54][1] and compared it with CBL [42][2]. The experiments were run on an AMD EPYC 7302 (3 GHz) server with 251 GB RAM and SSD storage.

Our proposed pipeline for these performing these operations with FMSI, as depicted in Figure 2, consists of five steps: First, we compute a textual representation of the k-mer sets interpreted as **or**-MS. In our experiments, this was done using KmerCamel's global greedy algorithm [53]. Second, we recast the mask to the desired demasking function, specifically we keep **or** for union and change to **oon** for intersection and to **xor** for symmetric difference. In the case of **or**-MS computed by KmerCamel, mask recasting is actually not needed as the output already minimizes the number of 1s in the mask. Then we index the f-MS using FMSI [54] (this can be done even before mask recasting). The last two steps are concatenating the two indexes by index merging and compacting the resulting index if needed. Note that once indexed, we can ask membership queries on the resulting k-mer sets.

We evaluated the superstring length of each computed f-MS, and the memory requirements to perform queries on the indexed individual and concatenated f-MS, both before and after compaction. The results for intersection of the two roundworm genomes for $k = 21$ are depicted in Figure 2. Note also that the fact that compaction significantly reduces the MS length highly depends on the particular use case, namely on the proportion of represented k-mers in the result. For union and symmetric difference for the same data, the compaction lead to negligible length reduction.

---

[1] Obtained from `https://github.com/OndrejSladky/fmsi`, version at commit '`778f2f0`'.

[2] Obtained from `https://github.com/imartayan/CBL`, version at commit '`328bcc6`'. The index was computed on canonical k-mers, to handle reverse complements, that is, we ran '`cbl build -c`'. CBL was compiled using '`RUSTFLAGS="-C target-cpu=native" K={kmer-size} PREFIX_BITS=28 \\ cargo +nightly build -release -examples -target-dir target.k_{kmer-size}`'.

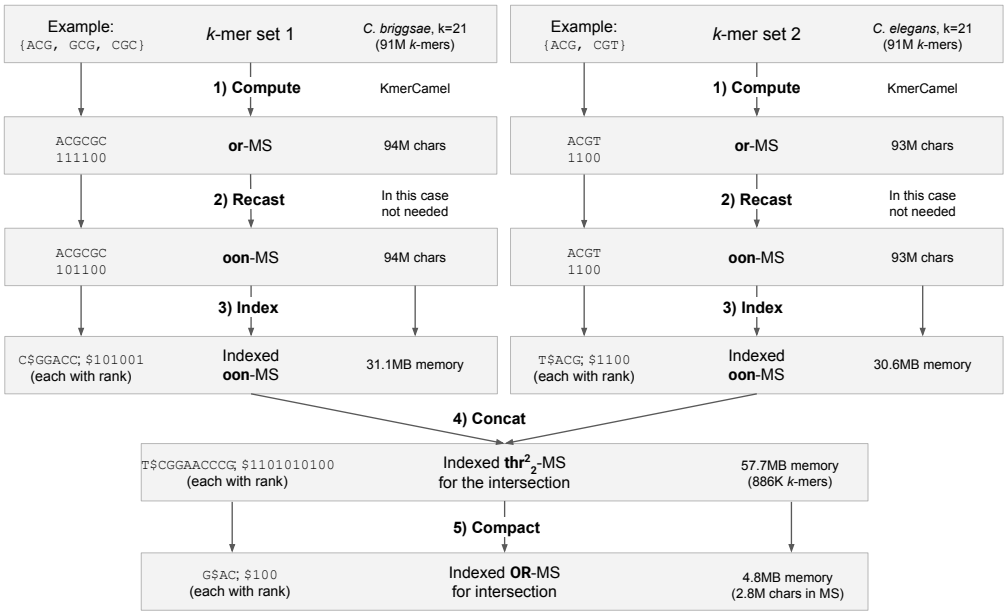

Figure 2 Illustration of the k-mer set operations workflow using f-masked superstrings: an example of intersection of two sets using FMSI [54]. The workflow also contains an illustrative example on a set of 3-mers as well as experimental results on *C. briggsae* and *C. elegans* genomes with $k = 21$; namely, we show the number of masked superstring characters after each change, the number of distinct k-mers in each set, and for each indexed representation, the memory required to perform queries on the underlying set.

For results of FMSI with symmetric difference, union, and other values of k, see the supplementary repository (`https://github.com/OndrejSladky/f-masked-superstrings-supplement`).