Strobemers: an alternative to k-mers for sequence comparison

Kristoffer Sahlin[1,✉]

[1]Department of Mathematics, Science for Life Laboratory, Stockholm University, 106 91, Stockholm, Sweden.

K-mer-based methods are widely used in bioinformatics for various types of sequence comparison. However, a single mutation will mutate k consecutive k-mers and makes most k-mer based applications for sequence comparison sensitive to variable mutation rates. Many techniques have been studied to overcome this sensitivity, *e.g.*, spaced k-mers and k-mer permutation techniques, but these techniques do not handle indels well. For indels, pairs or groups of small k-mers are commonly used, but these methods first produce k-mer matches, and only in a second step, a pairing or grouping of k-mers is performed. Such techniques produce many redundant k-mer matches due to the size of k.

Here, we propose *strobemers* as an alternative to k-mers for sequence comparison. Intuitively, strobemers consists of linked minimizers. We show that under a certain minimizer selection technique, strobemers provide more evenly distributed sequence matches than k-mers and are less sensitive to different mutation rates and distributions. Strobemers also give a higher total coverage of matches across sequences. Strobemers are a useful alternative to k-mers for performing sequence comparisons as commonly used in read alignment, clustering, classification, and error-correction.

k-mers | minimizers | sequence matching | data structures
Correspondence: ***ksahlin@math.su.se***

Introduction

The decreased sequencing costs have led to a dramatic increase in sequencing data over the past two decades, which has prompted development of computational methods for sequence comparison. A popular sequence comparison paradigm is k-mer based analysis, where k-mers are substrings of length k of, *e.g.*, genomic, transcriptomic, or protein sequences. K-mer-based methods have been applied for sequence comparison for error correction (1), genome assembly (2, 3), metagenomic (4) and chromosome (5) sequence classification, sequence clustering (6), database search (7, 8), structural variation detection (9, 10), transcriptome analysis (11, 12), and many other applications. Because of this widespread use, many data structures and techniques for efficiently storing, querying, and counting k-mers have been proposed (see (13) for a review).

While k-mers has proven to be practical in several sequence comparison problems, they are sensitive to mutations. A mutation will mutate k consecutive k-mers across a string. As the mutation rate increases, the number of matching k-mers quickly reduces. In (14), the distribution of mutated k-mers were studied in detail. The authors provided closed-form expressions for the mean and variance estimates on the number of mutated k-mers under a random mutation model. For sequence comparison methods, while the number of k-mer matches between sequences is of interest, it is often more informative to know how they are distributed across the matching region. K-mer matches, because of their consecutive nature, cluster tightly in shared sequence regions, while matches may be absent in regions with higher mutation rates. Spaced k-mers (or spaced seeds) have been studied in several sequence comparison contexts to overcome the k-mers' sensitivity to mutations (15, 16). The advantage of spaced k-mers is that matches of spaced k-mers at different positions are less correlated with each other than k-mer matches. In fact, k-mers are the worst seed pattern for the problem of similarity search (17). Another innovative idea has been to permute the letters in a string before comparison (18, 19). The main idea is that, under several different permutations in a region of a string, at least one permutation will, with statistical certainty, have pushed any mutation(s) towards the end of the strings under comparison, which allows for constant-time query of the prefix of the permuted strings. However, both spaced k-mers and permutation techniques are only practical for substitutions. An insertion or deletion (*indel*) will shift the sequence and, similarly to k-mers, result in a long stretch of dissimilar k-mers. For certain applications such as genome assembly, selecting several sizes of k for inference has also shown to help sequence comparison (20), but it significantly increases runtime and complexity of analysis. There are also methods to collapse repetitive regions before k-mer based comparison (21), which reduces the processing time of repetitive hits. However, such techniques is usually employed for reference-based analysis and do not apply to general sequence comparison problems.

As third-generation sequencing techniques appeared with sequencing errors mostly consisting of insertions and deletions, many of the previously developed sequence comparison techniques for short-read data became unsuitable. For the third generation sequencing, MinHash (22) and minimizers (23, 24) proved to be useful data structures for such sequence comparison as minimizers can be preserved in a window affected by an indel. In addition, they also reduce the size of the index by subsampling the data. This has made MinHash and minimizers a popular technique for subsampling k-mers employed for sequence comparison in a range of applications such as metagenome distance estimation (25) and alignment (26, 27), clustering (28) error-correction (29), and assembly (30) of long-read sequencing data. Because of the widespread practical use of minimizers, several methods

DRAFT

have been proposed for sampling them with as low density as possible (31), more evenly (32, 33).

Due to the error rates of long-read sequencing, minimizers are often chosen much shorter (about 13-15nt) than k-mer lengths considered unique in, e.g., the human genome (>20nt). With this length of minimizers, they produce many candidate sequence matches. Therefore, it would be useful to combine the robustness of minimizers to indels and mutation errors with larger k-mer sizes that would offer more unique matches. One approach is to use a small k-mer size and identify pairs (34) or groups (35) of them clustered tightly together. Such methods are robust to any mutation type and have shown to, e.g., improve overlap detection between long reads (36). However, they still match single k-mers individually and group them based on statistics after individual k-mer hits have been found. To remove the redundancy in matches, we suggest that it is beneficial to couple the k-mers in the initial matching step and perform a single constant-time lookup of coupled k-mers. Coupled k-mers have been explored in, e.g., (29, 37) where *paired minimizers* are generated and stored as a single hash. A paired minimizer-match signals that the region is similar between sequences. Due to the gap between the minimizers, such a structure is not as sensitive to indels or substitutions as k-mers. Paired minimizers were shown to be useful for both genome assembly (37) and error correction of long cDNA reads (29) where the reads are similar only in some regions due to alternative splicing. However, in both (37) and (29), they consider pairs of minimizers where each minimizer is produced independently, and paired up after the minimizer generation. Here, we show that we can substantially improve on paired minimizers for sequence matching by producing minimizers chosen jointly depending on previous windows, and we generalize paired minimizers link more than two.

We propose a method to extract subsequences from a string and we call those subsequences *strobemers*. The name is inspired by strobe sequencing technology (an early Pacific Biosciences sequencing protocol), which would produce multiple subreads from a single contiguous fragment of DNA where the subreads are separated by 'dark' nucleotides whose identity is unknown, illustrated in (38). Strobemers introduced here are, however, produced computationally. Intuitively, strobemers are groups of linked short k-mers (*strobes*) from adjacent windows. The k-mers can be chosen as minimizers either independently within windows (*e.g.*, as in (29, 37)), which we call *minstrobes*, or dependent on previous k-mers called *randstrobes*. We show that strobemers, and particularly randstrobes, improve sequence matching by providing more evenly distributed matches than k-mers, are less sensitive to different mutation rates and give a higher total coverage of matches across strings. We also show on human chromosomes that strobemers can offer higher uniqueness, and therefore confidence in a match, than k-mers due to the spacing of the strobes. Strobemers generalizes paired minimizers and we show that choosing minimizers dependent on previous k-mers (randstrobes) significantly improves over independently generated paired minimizers as in (29, 37). Strobemers are easy to both construct and query, making it a compelling alternative to k-mers for sequence comparison. Furthermore, strobemers can, similarly to k-mers, be subsampled using MinHash, minimizers, syncmers (33) or any other thinning protocol.

Methods

Definitions. We refer to a *subsequence* of a string as a set of ordered letters that can be derived from a string by deleting some or no letters without changing the order of the remaining letters. A substring is a subsequence where all the letters are consecutive. We use i to index the position in a string s and let $s[i:i+k]$ denote a *k-mer* substring at position i in s. If $s[i:i+k]$ is identical to a k-mer $t[i':i'+k]$ in string t, we say that the k-mers *match*, and that the match occurs at position i in s (and j in t). Similarly, let $f(i,s)$ be any function to extract a subsequence with first letter at position i from s. If $f(i,s)$ is identical to $f(i',t)$, we say that the subsequences match, and that the match occurs at position i in s (and j in t). For example, for k-mers we have $f(i,s) = s[i:i+k]$. We let $|\cdot|$ denote the length of strings.
We use h to denote a hash function $h : \sum^* \rightarrow \mathbb{Z}$ mapping strings to random numbers. Given two integers w and k such that $1 \leq k \leq w \leq |s|$, the minimizer at position i in s is the substring of s of length k starting in the interval of $[i, i+w-k+1)$ that minimizes h.

Aim. We will introduce strobemers by describing the problem they aim to solve. Consider two strings s and t that are identical up to m mutations. We desire a function f to produce a set of subsequences from s and t that have two characteristics: (1) as few possible placements of the m mutations result in no matches between s and t, and (2) the subsequences of length k should be as unique as k-mers on s and t. Characteristic 1 and 2 relates to the sensitivity and precision of searching for matches and we will discuss this in an example below. For practical purposes we also require that at most one subsequence is produced per position to mimic how k-mers are derived in a string (and limit the amount of data we store for each string). Certainly, we could produce all possible subsequences at each position to minimize criteria 1, but this is not feasible.

A motivational example. Consider two strings of 100 nucleotides with $m = 3$ mutations between them. This could occur, *e.g.*, in splice alignment to an exon, or in sequence clustering. If we use a k-mer of size 30 to find matches, and the two strings differ at position 25, 50 and 75, there will be no matching k-mers. Similarly, this holds for mutations at positions 15, 40, and 65, and many other combinations. As described, we want as few possible placements of errors leading to the region being unmatched.
Using spaced k-mers (15) or permutations of the string (19) would help the mutations were substitutions. We could consider lowering k, but this would generate more matches to other stings as well. To achieve the same uniqueness as longer k, we could consider coupled k-mers (34) of say 15nt

DRAFT

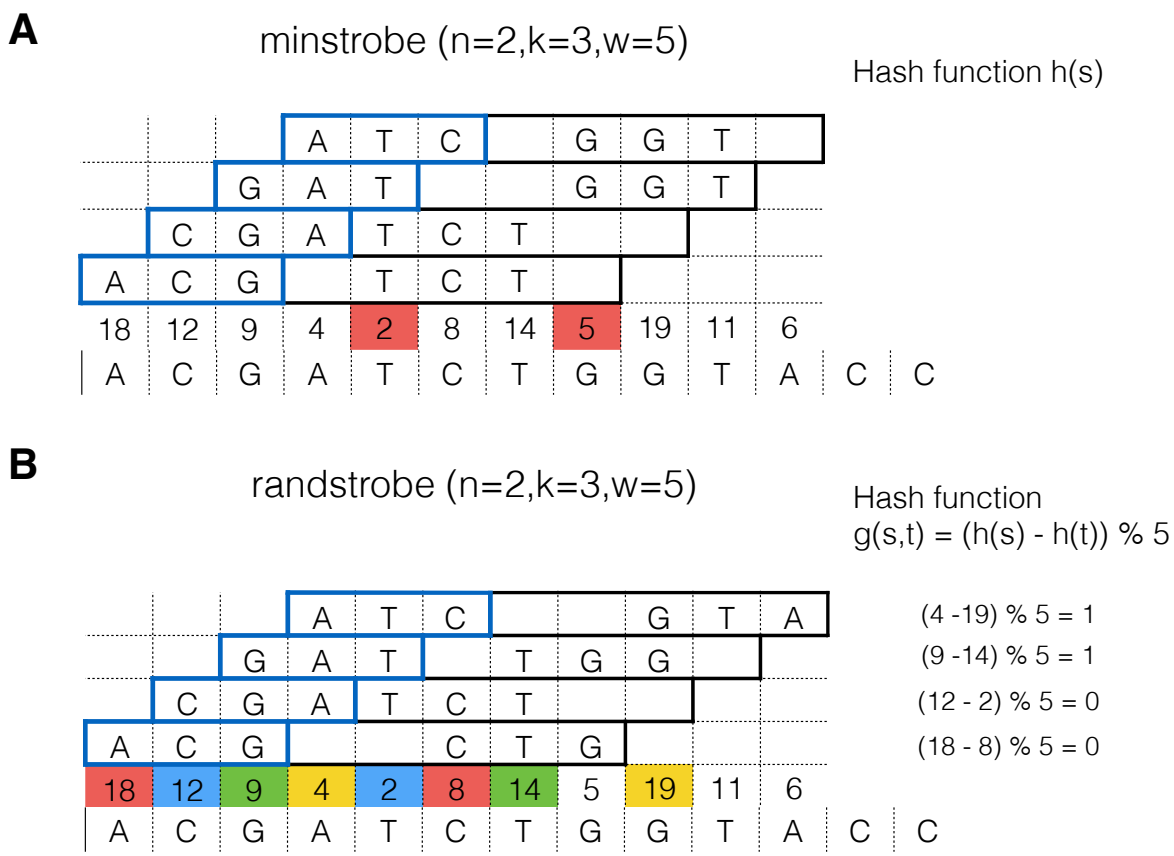

Fig. 1. An illustration of four minstrobes (A) and randstrobes (B) with $n = 2, k = 3, w = 5$ generated from a DNA string of 13 letters. The values on top of letters are hash values computed for each 3-mer under some random hash function h. The start k-mer k_1 is highlighted in blue, and the adjacent window of five nucleotides is in boldfaced boxes. In the minstrobe protocol, minimizers (red squares) are selected independently in each window, which gives rise to many similar minimizer pairs. In the randstrobe protocol, minimizers in w_1 are selected dependent on previous k-mer k_1, here illustrated with an example hash function $g(s,t) = (h(s) + h(t)) \mod 5$. Example hash values are provided fore each of the four randstrobes. Two coupled values that minimizes g are showed in the same color. Because of joint hash function, randstrobes are more randomly (but deterministically) distributed across the sequence.

per pair, with some gap in between them. Note that, however, to avoid many matches to other sequences, the k-mers would need to be coupled before searching for matches, not simply look for clustered k-mers. Furthermore, if the coupled k-mers have a fixed distance from each other, we have just created a specific type of spaced k-mers, which are only robust to substitutions. We, therefore, could consider coupled minimizers (29) to select a random gap size for us, but in a deterministic manner.

This brings us to the strobemers. In the scenario above, we could pick a k-mer of size 15 at a position we want to sample and couple it with a minimizer of length 15 derived from a window downstream from the k-mer. Together, they have sequence length 30 and are therefore robust to false matches. They can also match across the mutations, where the mutations could be both substitutions and indels. If we increase the mutation density on our string, eventually, our two k-mers of length 15nt will also fail to produce any matches. Therefore, we could consider triplets of a kmer and two minizizers of length, *e.g.*, 10nt. Finally, as we will show in this study, we could further reduce the sampled minimizers' dependency, and therefore the matches, by a joint hash function (over k-mers) to compute minimizers.

Strobemers. Consider a string s. A strobemer of order n in s is a subsequence of s of length k at some position $i \in [1, |s| - k + 1]$. Specifically, a strobemer at position i in s is composed of a set of k-mers $k_1, ...k_n$ of equal length, that we call *strobes*, where $k_1 = s[i, i + k]$ and $k_2, ...k_n$ are minimizers chosen from adjacent windows of size $w > |k_j|$, with k_j starting in interval $[i + |k_1| + (j - 1)w, i + |k_1| + jw]$. We will from now on parametrize a strobemer as $(n, |k_1|, w)$ denoting the order, the length of the strobe, and the window size, respectively.

We select $k_2, ...k_n$ based on some hash function. We will consider two different hash functions for producing them, which give significantly different results. First we denote as *minstrobe*, a strobemer where strobes $k_2, \ldots, k_n$ are independently selected as minimizers in their respective windows under a random asymmetric hash function h.

Second, we denote as *randstrobe*, a strobemer where strobe k_j is selected as minimizer dependent on the previous $k_1, \ldots k_{j-1}$ strobes. Specifically, in this study we chose the random asymmetric hash function $h(k_1, \ldots, k_{j-1}, k) = k_1 \oplus \ldots \oplus k_{j-1} \oplus k$, where $\oplus$ denotes string concatenation. Here, $h(k_1, \ldots, k_{j-1}, k)$ concatenates the previous selected strobes $k_1, \ldots, k_{j-1}$. Thus, the k-mers in the randstrobe are

produced iteratively from $i = 1, .., n$ and yields a more randomly distributed set of strobes. Fig. 1 gives an illustration of minstrobes and randstrobes.

There are two important practical differences between minstrobes and randstrobes. Firstly, for two different starting strobes k_1 and k_1', k-mers $k_2, .., k_n$ will most frequently be the same under the minstrobe generation due to independent minimizers, and most frequently differ in a randstrobe. Secondly, generating minstrobes is in practice as fast as producing minimizers while generating randstrobes, under the function we consider here, is not. We elaborate on this in the section on time complexity.

Construction. We aim to produce a minstrobe or randstrobes of a string s in a similar manner to how k-mers are produced, *i.e.*, one strobemer per position $i \in [1, |s| - k + 1]$. This would mean that we extract the same amount of kmers and strobemers from a string s. Note however that the number of unique k-mers and strobmers may differ. We construct strobemers as follows. We denote the total length of the string considered to construct the n-strobe as $W = |k_1| + nw_i$, and the total subsequence length as $k = \sum_{j=1}^{n} k_j$. If $W \leq |s| - i$ we use the predefined window size w and compute the strobemers under the respective hash functions described above. If $W > |s| - i$, we can narrow the window sizes until k_1 to k_n are all adjacent to each other producing a substring of length k. Any way to narrow the windows can be considered. Here, we choose to shorten each window w to $\frac{\lfloor |s| - i \rfloor}{n}$.

Time complexity. If we ignore the time complexity of the hash function, the time complexity of generating minimizers is $O(|s|w)$ for a window size w. However, as (26) noted, computing minimizers is in practise close to $O(|s|)$ if we use a queue to cache previous minimizer values in the window. The expensive step is when a previous minimizer is discarded from the queue and a new minimizer needs to be computed for the window.

Similar to computing minimizers, both minstrobes and randstrobes have the same worst-case time complexity of $O(|s|W)$. However, the independence of hash values in the minstrobes protocol makes it $O(|s|)$ in practice by using a queue in the same manner as computing minimizers independently. The randstrobe protocol does not have this independence under the hashing scheme we consider in this study, which means that all hash values have to be recomputed at each position. This means its practical time complexity is therefore $O(|s|W)$. We did not study different schemes to produce randmers but note that this is subject for future study. For example, it is possible that, by exploring symmetry properties (39), we may come up with faster methods to generate them.

Results

Here we will compare sequence matching performance for k-mers, minstrobes and randstrobes (order 2 and 3). We first compare how effective the different protocols are at finding matches under different error rates. We then compare how effective they are at providing the same uniqueness (or confidence) in a match. Naturally, the size of k is central in these two aspects. We are interested in comparing sizes of subsequences that are similar between the protocols. Specifically, if the size of the k-mer is 30, we want to compare the k-mers to strobemers parameterized, e.g., by $(2, 15, \cdot)$ and $(3, 10, \cdot)$ as all the extracted subsequences have a length of 30 on the strings.

evaluation metrics. We say that a match between two sequences s_1 and s_2 occur at position i and i' in the two strings respectively, if the the k-mer (strobemer) extracted from position i in s and i' in t produce the same k-mer (strobemer). Furthermore, we say that this match *covers* positions $[i, i + k]$ for k-mers, and $[i, i + k_1]$ for strobemers in s. We adapt similar terminology as in (14) and denote a maximal interval of consecutive positions without matches between s and t as an *island*.

To evaluate the ability to preserve matches under different error rates, we compare (i) the number of matches, (ii) the total fraction of covered positions across the strings, and (iii) the distribution of islands. We need to make some clarifications on these evaluation metrics. First, our experiments on simulated data are designed with parameters so that the event of observing a false match under any protocol has a negligible probability. This means that our simulated experiments only measure the raw ability to identify correct matches. Secondly, to compute the total coverage of matches across the strings, we let the full string k contributes to k-mer coverage, while only the first strobe k_1 contributes to the strobemer protocols. This is highly unfavorable for the strobemer-protocols as it only provides a lower bound on coverage. While we could store the other strobes' positions and get exact coverage, this is not the intended use of the data structure and such information is lost (under a non-reversible hash function). Therefore, we simply omit them from the coverage computation. Consequently, strobemer protocols will not produce perfect coverage even for identical strings as the last segment between the strings cannot be added to the coverage.

strobemer vs k-mer matching. We compare how effective the different protocols are at preserving matches for different error rates. We start with a controlled scenario, where mutations are distributed with a fixed distance. In our second experiment, we use a random mutation distribution. We perform the fixed-distance mutation experiment to illustrate clearly the potential advantage of strobemer protocols.

First, we provide a small simulation to illustrate a scenario similar to the motivational example described earlier. We simulate a string s of 100 random nucleotides and a string t derived from simulating mutations every 15th position in s. Insertions, deletions, and substitutions are chosen at random with equal probability. We simulate s and t five times to illustrate the variability in matches for the strobemers between simulations. The distribution of matches under two different parametrizations for minstrobes and randstrobes are shown in Fig. 2. We note that we do not obtain any matches for k-mers of 15nt or larger in this scenario. Minstrobes, while

DRAFT

SIM-C

		m (%)	c (%)	g
k-mer	30	0	0	1000
minstrobe	(2,15,50)	4.7	19.3	84.6
	(3,10,25)	**9.0**	27.2	37.7
randstrobe	(2,15,50)	4.7	50.0	16.7
	(3,10,25)	**9.0**	**55.2**	**10.7**

Table 1. Statistics of the number of matches (m) as a percentage of the length of the original sequence of 10,000nt, the percentage of covered bases by matches (c) and the average island size (g) for the SIM-C dataset which has evenly spaced mutations with distance 20nt. The second column shows the parameters to the protocols.

SIM-R

		0.01			0.05			0.1		
		m	c	g	m	c	g	m	c	g
k-mer	30	**74.3**	95.9	38.4	**22.3**	54.7	75.3	**4.7**	18.2	206.2
minstrobe	(2,15,50)	71.4	89.2*	26.6	18.9	40.9*	58.5	3.6	11.1*	185.8
	(3,10,25)	69.5	86.0*	20.1	16.7	35.2*	46.3	2.9	8.7*	160.4
randstrobe	(2,15,50)	71.4	**96.4***	20.4	18.8	**58.2***	38.4	3.6	**19.7***	100.7
	(3,10,25)	69.4	95.4*	**14.8**	16.7	53.0*	**28.9**	2.9	15.8*	**82.1**

Table 2. Statistics of the number of matches (m), the percentage of covered bases by matches (c) and the average island size (g) under mutations rates of 0.01, 0.05, 0.1. *In the strobemer protocols, only k_1 contributs to calculated covered positions, which is a lower bound of the actual covered positions. For the k-mers, this value is exact.

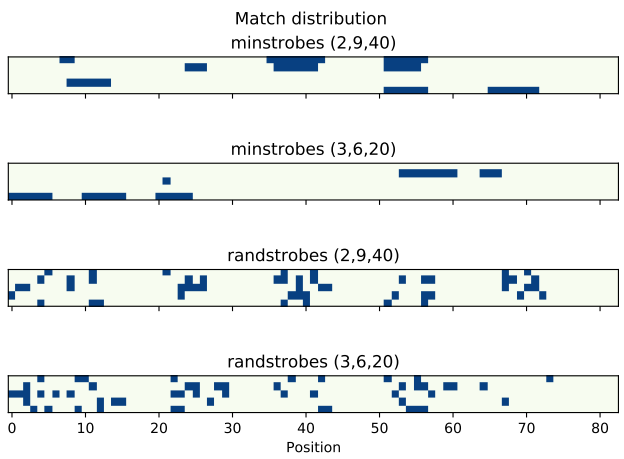

Fig. 2. An example of strobemer matches for minstrobes and randstrobes with two different parametrizations each (separate panels). Each panel shows matches between a string s of 100nt and a string t derived from simulating mutations every 15th position in s. Indels and substitutions are chosen at random with equal probability. The matches are plotted with respect to the positions in s on the 81 possible matching positions (x-axis). Each row in a panel corresponds to a separate simulation.

more effective than k-mers, fail to produce matches in simulation 3 for the $(2, 9, 40)$ parametrization and in simulation 1 and 4 for the $(3, 6, 20)$ parametrization. We observe that randstobes produce matches in all five experiments under both parametrizations and provide a more random match distribution across the string than minstrobes.

To better quantify the performance in this scenario, we increase the size of our controlled experiment. We simulate a string of length 10,000nt and construct a second string by generating an insertion, deletion, or substitution (equal probability) every 20 nucleotides. We then simulate k-mers with size 30, and strobemers with parameters (2,15,50) and (3,10,25) so that they all have the same subsequence length, and compare the number of matches, coverage, and average island size under the different protocols (table 1). We repeat the experiment 1000 times to alleviate sample variance. We refer to this as the SIM-C experiment (for simulation controlled). This experiment highlights the advantage of strobemers in such a scenario and further shows the difference in performance between minstrobes and randstrobes. Randstrobe matches cover the largest fraction of the sequences, and they also have the smallest average island size (table 1).

In our second experiment, we simulate a string of length 10,000nt and construct a second string by generating insertions, deletions or substitutions with equal probability across the string with mutation rate $\mu \in 0.01, 0.05, 0.1$. This means that the positions for the mutations are randomly distributed. Each such simulation is replicated 1000 times to alleviate sample variation. We refer to this as the SIM-R experiment (for simulation random). We see that strobemers are still significantly better at distributing matches across the sequences compared to k-mers across all mutation rates with smaller islands for both minstrobes and randstrobes (table 2), and higher coverage for randstrobes with $n = 2$.

strobemer vs k-mer uniqueness. We also need to compare the confidence or uniqueness of a match. Strobemers offer more flexibility in matches by the ability to be preserved under mutations and can match regions that are not identical, contrary to k-mers. Therefore, it is reasonable to assume that, for the same size k of extracted subsequence, the strobemer protocols will have lower uniqueness (precision) than k-mers. We study this by computing the percentage of unique k-mers and strobemers on the five largest human chromosomes (Fig. 3). For a k-mer size of k, we parametrize the strobemer protocols with $(n, k/n, 50/(n-1))$ for $n = 2, 3$ in order to have the same subsequence lengths to occupy the same amount of memory and can be stored and queried under the same conditions.

Contrary to our intuition, the strobemer protocols also offer higher precision at sizes of $k \geq 24$ (Fig. 3). In our ex-

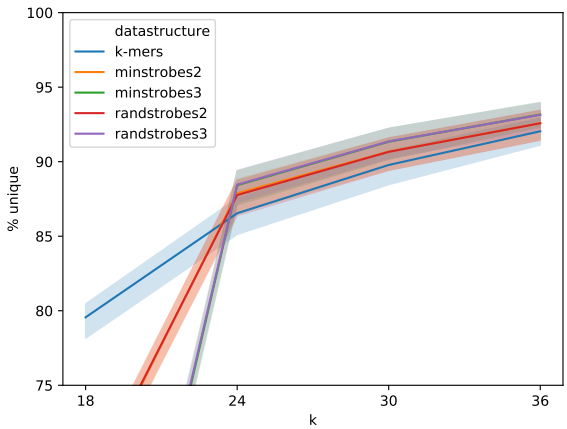

Fig. 3. The percent of uniqe k-mers, minstrobes and randstrobes (y-axis) on the five largest chromosomes (chr1-chr5) of the human genome for various sequence lengths of k (x-axis). For a given k in the plot, strobemers with $n = 2$ are computed with parameters $(2, k/2, 50)$ and strobemers with $n = 3$ are computed with parameters $(2, k/3, 25)$ so that the number of extracted nucleotides between the five methods are the same. Y-axes have been cut at 75% for illustration. The value for minstrobes and randstrobes with parameters $(3, 6, 25)$ are 43.2% and 43.4% respectively. The value for minstrobes and randstrobes with parameters $(2, 9, 50)$ are 67.9% and 68.0% respectively.

periments, strobemers of order 3 gives the highest percentage of unique mmatches for reasonably large subsequence lengths k. However, for $k = 18$, the strobemer protocols will be parametrized by $(2, 9, 50$ and $(3, 6, 25)$, which appear too small to guarantee reasonable uniqueness on the largest human chromosomes.

Discussion

We have studied strobemers, an alternative sampling protocol to k-mers. We have experimentally demonstrated that strobemers, particularly randstrobes, efficiently reduce gaps between matches in sequences under different mutation rates. We demonstrated that randstrobes are less sensitive to the distribution and density of mutations across all SIM-C and SIM-R experiments, providing smaller islands and a higher percentage of covered bases (Table table 1 and table 2).

While k-mers, in general, provide more matches under a random error distribution (Table table 2), the matches are compactly clustered between mutations, and the number of matches is not always helpful as the matches may cluster due to local repeats. Randstrobes, on the other hand, produce a more even distribution of matches, and higher match coverage across the sequence (Table table 1 and table 2), even though we underestimate the coverage of strobemer protocols in our experiments. Both of these features are useful for several algorithms that require chains of matches between two sequences (26, 28) to be considered candidates for alignment or clustering. Randstrobes may be particularly useful in these scenarios and may offer a better alternative than, *e.g.*, minimizer spans (29). Overall, the strobemers, particularly randstrobes, show a promising data structure for algorithms that rely on sequence comparison.

We also demonstrated on the largest human chromosomes that the strobemer protocols also have higher precision than k-mers, through a higher percentage of strobemers are unique on the chromosomes. We find this surprising, as the strobemers were designed to offer more flexibility in matches by allowing substitutions and indels between strobes. We hypothesize that the strobemers, due to their larger span compared to k-mers, offer more uniqueness in repetitive regions. This remains to be studied.

Currently, randstrobes are slower to generate than k-mers and minstrobes. In practice, generating randstrobes takes $O(|s|W)$ time compared to a runtime of $O(|s|)$ for k-mers, and an, in practice, runtime of $O(|s|)$ for minstrobes. By employing ideas like cyclic polynomial hash functions (39), we may come up with faster methods to generate them. In terms of memory requirements, for a given subsequence length k, k-mers and strobemers parametrized as $(n, k/n, \cdot)$ require the same memory to store as they can be treated as concatenated strings with the same length.

While our study provides an experimental evaluation of strobemers under some commonly used values of k and mutation rates, the statistics of strobemers remains to be explored. In (14), the authors derived the mean and variance of islands for k-mers and the number of mutated k-mers under given mutation rate. If we can derive analytic expressions for strobemers, it will help to optimize parameters of the strobemer protocols under various mutation rates, which will be useful for similarity comparison algorithms. Since strobemers are gapped sequences, it also motivates the study of match coverage and distribution of matches across regions (or positions) similarly to what has been done for gapped experimental protocols such as mate-pair or paired-end reads (40). For example, one could compute the match coverage across a position or smaller region to decide the confidence of the matches in the regions. Finally, similarly to k-mers, minstrobes and randstrobes can be subsampled as minimizers (24), syncmers (33), or any other thinning protocol that can be applied to k-mers. By stydying the mathematical properties of hashes and minimizers (31, 41), we may also find a effective subsampling techniques of strobemers.

Conclusions

We have presented strobemers as an alternative to k-mers for sequence comparison. Strobemers, particularly randstrobes, offer a more evenly distributed set of matches across sequences than k-mers, are less sensitive to the distribution of mutations across sequences, and can produce a higher match coverage under some parametrizations. We also showed that strobemers can offer higher uniqueness than k-mers at the same subsequence length for practical subsequence sizes. These features are useful for algorithms that perform sequence matching. Strobemers are also easy to both construct and query, making it a compelling alternative to k-mers. While we have empirically demonstrated the useful properties of strobemers, their statistical properties require further investigation.

ACKNOWLEDGEMENTS

DRAFT

The computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Bibliography

1. Leena Salmela, Riku Walve, Eric Rivals, and Esko Ukkonen. Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, 33(6):799–806, 06 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw321.
2. P A Pevzner. 1-tuple dna sequencing: computer analysis. *J Biomol Struct Dyn*, 7(1):63–73, Aug 1989. ISSN 0739-1102 (Print); 0739-1102 (Linking). doi: 10.1080/07391102.1989.10507752.
3. Rayan Chikhi and Paul Medvedev. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 30(1):31–37, 06 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt310.
4. Derrick E. Wood and Steven L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014. doi: 10.1186/gb-2014-15-3-r46.
5. Samarth Rangavittal, Natasha Stopa, Marta Tomaszkiewicz, Kristoffer Sahlin, Kateryna D. Makova, and Paul Medvedev. Discovery: a classifier for identifying y chromosome sequences in male assemblies. *BMC Genomics*, 20(1):641, 2019. doi: 10.1186/s12864-019-5996-3.
6. Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1):2542, 2018. doi: 10.1038/s41467-018-04964-5.
7. Brad Solomon and Carl Kingsford. Fast search of thousands of short-read sequencing experiments. *Nat Biotechnol*, 34(3):300–302, Mar 2016. ISSN 1546-1696 (Electronic); 1087-0156 (Print); 1087-0156 (Linking). doi: 10.1038/nbt.3442.
8. Robert S Harris and Paul Medvedev. Improved representation of sequence bloom trees. *Bioinformatics*, 36(3):721–727, 08 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz662.
9. Ryan P Abo, Matthew Ducar, Elizabeth P Garcia, Aaron R Thorner, Vanesa Rojas-Rudilla, Ling Lin, Lynette M Sholl, William C Hahn, Matthew Meyerson, Neal I Lindeman, Paul Van Hummelen, and Laura E MacConaill. Breakmer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic Acids Res*, 43(3):e19, Feb 2015. ISSN 1362-4962 (Electronic); 0305-1048 (Print); 0305-1048 (Linking). doi: 10.1093/nar/gku1211.
10. Daniel S. Standage, C. Titus Brown, and Fereydoun Hormozdiari. Kevlar: A mapping-free framework for accurate discovery of de novo variants. *iScience*, 18:28–36, 2021/01/25 2019. doi: 10.1016/j.isci.2019.07.032.
11. Rob Patro, Stephen M Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, 2014. doi: 10.1038/nbt.2862.
12. 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.
13. 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*, 2020. doi: 10.1101/gr.260604.119.
14. Antonio Blanca, Robert S. Harris, David Koslicki, and Paul Medvedev. The statistics of k-mers from a sequence undergoing a simple mutation process without spurious matches. *bioRxiv*, 2021. doi: 10.1101/2021.01.15.426881.
15. Bin Ma, John Tromp, and Ming Li. PatternHunter: faster and more sensitive homology search . *Bioinformatics*, 18(3):440–445, 03 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.3.440.
16. Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome Biology*, 20(1):257, 2019. doi: 10.1186/s13059-019-1891-0.
17. Uri Keich, Ming Li, Bin Ma, and John Tromp. On spaced seeds for similarity search. *Discrete Applied Mathematics*, 138(3):253 – 263, 2004. ISSN 0166-218X. doi: https://doi.org/10.1016/S0166-218X(03)00382-2.
18. Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, page 380â388, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134959. doi: 10.1145/509907.509965.
19. Roy Lederman. A random-permutations-based approach to fast read alignment. *BMC Bioinformatics*, 14(5):S8, 2013. doi: 10.1186/1471-2105-14-S5-S8.
20. Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, Alexey V Pyshkin, Alexander V Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A Alekseyev, and Pavel A Pevzner. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 19(5):455–477, May 2012. ISSN 1557-8666 (Electronic); 1066-5277 (Print); 1066-5277 (Linking). doi: 10.1089/cmb.2012.0021.
21. Bo Liu, Hongzhe Guo, Michael Brudno, and Yadong Wang. deBGA: read alignment with de Bruijn graph-based seed and extension. *Bioinformatics*, 32(21):3224–3232, 07 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw371.
22. Andrei Broder. On the resemblance and containment of documents. 06 1997. doi: 10.1109/SEQUEN.1997.666900.
23. Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. Winnowing: Local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, SIGMOD '03, page 76â85, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 158113634X. doi: 10.1145/872757.872770.
24. Michael Roberts, Wayne Hayes, Brian R. Hunt, Stephen M. Mount, and James A. Yorke. Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20(18):3363–3369, 07 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth408.
25. Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biology*, 17(1):132, 2016. doi: 10.1186/s13059-016-0997-x.
26. Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, Jul 2016. ISSN 1367-4811 (Electronic); 1367-4803 (Print); 1367-4803 (Linking). doi: 10.1093/bioinformatics/btw152.
27. Chirag Jain, Alexander Dilthey, Sergey Koren, Srinivas Aluru, and Adam M Phillippy. A fast approximate algorithm for mapping long reads to large reference databases. *J Comput Biol*, 25(7):766–779, Jul 2018. ISSN 1557-8666 (Electronic); 1066-5277 (Print); 1066-5277 (Linking). doi: 10.1089/cmb.2018.0036.
28. Kristoffer Sahlin and Paul Medvedev. De novo clustering of long-read transcriptome data using a greedy, quality value-based algorithm. *Journal of Computational Biology*, 27(4):472–484, 2020. doi: 10.1089/cmb.2019.0299. PMID: 32181688.
29. Kristoffer Sahlin, Botond Sipos, Phillip L. James, and Paul Medvedev. Error correction enables use of oxford nanopore technology for reference-free transcriptome analysis. *Nature Communications*, 12(1):2, 2021. doi: 10.1038/s41467-020-20340-8.
30. Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630, 2015. doi: 10.1038/nbt.3238.
31. Guillaume Marçais, Dan DeBlasio, and Carl Kingsford. Asymptotically optimal minimizers schemes. *Bioinformatics*, 34(13):i13–i22, 06 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty258.
32. Chirag Jain, Arang Rhie, Haowen Zhang, Claudia Chu, Brian P Walenz, Sergey Koren, and Adam M Phillippy. Weighted minimizer sampling improves long read mapping. *Bioinformatics*, 36(Supplement_1):i111–i118, 07 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa435.
33. Robert C. Edgar. Syncmers are more sensitive than minimizers for selecting conserved k-mers in biological sequences. *bioRxiv*, 2020. doi: 10.1101/2020.09.29.319095.
34. S F Altschul, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997. ISSN 0305-1048 (Print); 1362-4962 (Electronic); 0305-1048 (Linking). doi: 10.1093/nar/25.17.3389.
35. Laurent Noé and Gregory Kucherov. Improved hit criteria for dna local alignment. *BMC Bioinformatics*, 5(1):149, 2004. doi: 10.1186/1471-2105-5-149.
36. Nan Du, Jiao Chen, and Yanni Sun. Improving the sensitivity of long read overlap detection using grouped short k-mer matches. *BMC Genomics*, 20(2):190, 2019. doi: 10.1186/s12864-019-5475-x.
37. Chen-Shan Chin and Asif Khalak. Human genome assembly in 100 minutes. *bioRxiv*, 2019. doi: 10.1101/705616.
38. Anna Ritz, Ali Bashir, and Benjamin J. Raphael. Structural variation analysis with strobe reads. *Bioinformatics*, 26(10):1291–1298, 04 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq153.
39. Hamid Mohamadi, Justin Chu, Benjamin P. Vandervalk, and Inanc Birol. ntHash: recursive nucleotide hashing. *Bioinformatics*, 32(22):3492–3494, 07 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw397.
40. Kristoffer Sahlin, Mattias Frånberg, and Lars Arvestad. Structural variation detection with read pair information: An improved null hypothesis reduces bias. *Journal of Computational Biology*, 24(6):581–589, 2017. doi: 10.1089/cmb.2016.0124. PMID: 27681236.
41. Yaron Orenstein, David Pellow, Guillaume MarÃ§ais, Ron Shamir, and Carl Kingsford. Designing small universal k-mer hitting sets for improved analysis of high-throughput sequencing. *PLOS Computational Biology*, 13(10):1–15, 10 2017. doi: 10.1371/journal.pcbi.1005777.

DRAFT