Sequence aligners can guarantee accuracy in almost $O(m \log n)$ time: a rigorous average-case analysis of the seed-chain-extend heuristic

Jim Shaw[1]* and Yun William Yu[1,2]

[1] Department of Mathematics, University of Toronto
[2] Computer and Mathematical Sciences, University of Toronto at Scarborough
*Corresponding author: jshaw@math.toronto.edu

Running title: Rigorous analysis of seed-chain-extend alignment

Abstract. Seed-chain-extend with k-mer seeds is a powerful heuristic technique for sequence alignment employed by modern sequence aligners. While effective in practice for both runtime and accuracy, theoretical guarantees on the resulting alignment do not exist for seed-chain-extend. In this work, we give the first rigorous bounds for the efficacy of seed-chain-extend with k-mers *in expectation*. Assume we are given a random nucleotide sequence of length $\sim n$ that is indexed (or seeded) and a mutated substring of length $\sim m \leq n$ with mutation rate $\theta < 0.206$. We prove that we can find a $k = \Theta(\log n)$ for the k-mer size such that the expected runtime of seed-chain-extend under optimal linear gap cost chaining and quadratic time gap extension is $O(mn^{f(\theta)} \log n)$ where $f(\theta) < 2.43 \cdot \theta$ holds as a loose bound. The alignment also turns out to be good; we prove that more than $1 - O(\frac{1}{\sqrt{m}})$ fraction of the homologous bases are *recoverable* under an optimal chain. We also show that our bounds work when k-mers are *sketched*, i.e. only a subset of all k-mers is selected, and that sketching reduces chaining time without increasing alignment time or decreasing accuracy too much, justifying the effectiveness of sketching as a practical speedup in sequence alignment. We verify our results in simulation and on real noisy long-read data and show that our theoretical runtimes can predict real runtimes accurately. We conjecture that our bounds can be improved further, and in particular, $f(\theta)$ can be further reduced.

Table of Contents

Sequence aligners can guarantee accuracy in almost $O(m \log n)$ time: a rigorous average-case analysis of the seed-chain-extend heuristic . . . 1

Jim Shaw and Yun William Yu*

1 Introduction . . . 4
- 1.1 Our contribution . . . 4

2 Results . . . 5
- 2.1 Assumptions and models . . . 5
- 2.2 Theoretical results . . . 7
- 2.3 Simulated genome alignment experiments . . . 8
- 2.4 Real nanopore read alignment experiments . . . 9

3 Discussion . . . 13
- 3.1 Improving bounds for the current random model . . . 14
- 3.2 Generalizing the random model . . . 14

4 Methods . . . 15
- 4.1 Homology and recoverability . . . 15
- 4.2 Fundamental tools and bounds . . . 16
- 4.3 Bounding sums of k-mer random variables . . . 17
- 4.4 Proof of non-sketched main result . . . 18
- 4.5 Sketching and local k-mer selection . . . 19

5 Software availability . . . 21

6 Competing interest statement . . . 21

7 Acknowledgements . . . 21

A Mathematical conventions and reminders . . . 26

B Missing proofs and definitions from "Homology and recoverability" . . . 26
- B.1 Recoverability and definition of $Align(\mathcal{C})$ . . . 26
- B.2 Proof of Lemma 1 . . . 27

C Missing proofs from "Fundamental tools and bounds" and "Bounding sums of k-mer random variables" . . . 28
- C.1 Proof of Theorem 3 . . . 28
- C.2 Proof of Corollary 1 . . . 30
- C.3 Proof of Lemma 2 . . . 30
- C.4 Proof of Lemma 3 . . . 31

D Missing proofs from "Proof of non-sketched main result" . . . 32
- D.1 Proof of Theorem 6 . . . 32
- D.2 Proof of Lemma 6 . . . 33
- D.3 Proof of Lemma 7 . . . 33
- D.4 Proof of Corollary 3 . . . 36
- D.5 Proof of Lemma 8 . . . 38
- D.6 Proof of Lemma 9 . . . 41

E Missing proofs from "Sketching and local k-mer selection" . . . 42
- E.1 Proof of Lemma 11 . . . 42
- E.2 Proof of Theorem 8 . . . 42
- E.3 Sketched concentration bounds and gap sizes . . . 42
- E.4 Proving Lemma 13 . . . 44

E.5 Re-proving sketched bounds 46
F Additional figures and tables 48

1 Introduction

Since the earliest years of bioinformatics, one primitive task has been sequence *alignment* (Altschul et al., 1990; Smith and Waterman, 1981), which plays a major role in genomic sequencing and phylogenetics. Intuitively, alignment matches together similar parts of two strings, inserting gaps where needed. More formally, alignment is defined by a set of allowed edit operations (e.g. single-character substitutions, insertions, deletions) with associated costs; an alignment is a sequence of those operations transforming one string into another string (or for local alignment, transforming one string into a substring of the other), and the alignment score is the summed cost of those operations (Berger et al., 2021).

Unfortunately, the best guaranteed algorithms for computing alignment (Needleman and Wunsch, 1970; Smith and Waterman, 1981) are quadratic in time-complexity; worse, this bound appears to be tight (Backurs and Indyk, 2018). There exist optimal algorithms with parameterized runtimes based on sequence similarity (Marco-Sola et al., 2021) that are faster in practice, but these still have worst-case quadratic time complexity as a function of the sequence length. Thus, to deal with large volumes of sequencing data (Marçais et al., 2019; Berger et al., 2016), sequence alignment programs employ heuristics (Altschul et al., 1990; Li, 2018; Marçais et al., 2018; Kiełbasa et al., 2011; Langmead and Salzberg, 2012; Lipman and Pearson, 1985) without performance guarantees (Medvedev, 2022a) for computational efficiency. All these heuristic algorithms are by necessity fast, achieving empirically sub-quadratic runtimes on real problems. They fail for adversarial examples, but aligners perform well in practice because the sequences being aligned are similar and not pathological examples (Ivanov et al., 2022; Myers, 1986; Ukkonen, 1983; Koerkamp and Ivanov, 2022).

Phylogenetics often makes use of comparative genomics, where aligning together multiple whole genomes allows annotating the ways in which two species have diverged over evolutionary time (Koonin et al., 2000). In genomic sequencing, the alignment task manifests because the sequencing machines are only able to read a small portion of a chromosome at a time, producing short snippets known as *reads* (Alser et al., 2021). It is then incumbent on read-mapping software to determine the likely origin location of that read in the genome, and for already sequenced species, this is usually performed by aligning the read to a known reference genome (Nurk et al., 2022).

Historically, different types of heuristics were used for the two tasks, because aligning a very small substring to a longer string is easier than aligning comparable-size strings. Indeed, the seed-and-extend heuristic, as seen in BWA and Bowtie 2, is preferred for aligning NGS short-reads to genomes instead of aligning genomes to genomes. However, as 3rd-generation long-read sequencing becomes more prominent, the two tasks become more similar and the same heuristics can be used for both (Sahlin et al., 2022)—in this manuscript, we address the *seed-chain-extend* heuristic (distinct from seed-and-extend) used in modern software for both read mapping and whole genome alignment.

1.1 Our contribution

Our goal in this manuscript is to close the gap between theory and practice, rigorously justifying the heuristics used in some of the most widely used alignment software. To this end, we turn to the methods of average-case analysis (Szpankowski, 2001), which gives us a way of breaking through the quadratic barrier of alignment (Medvedev, 2022b). Given a random string, we define a substitution model giving rise to a distribution on pairs of inputs and then average our analysis on pairs of strings over this distribution.

Recently, Ganesh and Sy (Ganesh and Sy, 2020) also used probabilistic analysis to show that a heuristic algorithm based on banded alignment can run in $O(n \log n)$ time for two length n

sequences and is optimal with high probability. However, their banded alignment heuristic has not yet seen any usage in practical software, and their analysis is invalid in the case of read mapping as it only pertains to two nearly equal-length strings. Thus we turn instead to the analysis of an empirically battle-tested heuristic for sequence alignment: seed-chain-extend. Seed-chain-extend is a well-established technique for comparative genomics (Abouelhoda and Ohlebusch, 2005; Myers and Miller, 1995), and recently the addition of *sketching* (or subsampling) has made it popular for long-read aligners (Ren and Chaisson, 2021; Chaisson and Tesler, 2012; Sović et al., 2016) including minimap2 (Li, 2018), the primary algorithm our model of seed-chain-extend is based on.

We provide, to the best of our knowledge, the first average-case bounds on runtime and optimality for the *sketched k-mer seed-chain-extend* alignment heuristic under a pairwise mutation model. Our optimality result shows that for large enough k-mer size k, the alignment is mostly constrained to be near the correct diagonal of the alignment matrix and that runtime is close to linear when the mutation or error rate is reasonably small. We also show that subsampling $\Theta(\frac{1}{\log n})$ of k-mers asymptotically reduces our bounds on chaining time but not for extension time. Our results give a theoretical justification for both the empirical accuracy and sub-quadratic runtime of seed-chain-extend. A simplified version of our theorem follows.

Simplified Theorem 1 (Informal main result; no sketching). Suppose we are given a uniformly random DNA string of length n and a mutated substring of length m where each base is substituted with probability θ. If $\theta < 0.206$ and the longer string is already seeded, then we can choose $k = \Theta(\log n)$ such that the expected runtime of k-mer seed-chain-extend is $O(mn^{f(\theta)} \log(n)) = O(mn^{2.43\cdot\theta} \log(n))$, and in expectation $\geq 1 - O(\frac{1}{\sqrt{m}})$ fraction of the homologous bases can be recovered from this alignment.

We will state our models and definitions precisely in Sections 2.1 and 4.4. Our main result, Theorem 1, precisely defines the function $f(\theta) < 2.43 \cdot \theta$ in the exponent, but $2.43 \cdot \theta$ is a convenient upper bound. We can quickly see from this bound that for modest mutation rates, $n^{2.43\cdot\theta}$ is not too large.

2 Results

2.1 Assumptions and models

Mutation model Let $S = x_1 x_2 \ldots x_{n+k-1}$ be a random uniform string with $n + k - 1$ i.i.d letters on an alphabet of size σ for some $k = \Theta(\log n)$. Let $S' = y_{p+1} y_{p+2} \ldots y_{p+m+k-1}$ be a substring of S of length $m + k - 1$ starting at a fixed position $p + 1$ with each character independently mutated to a different letter with probability θ. Although notationally a little confusing at first, the $k - 1$ term ensures S, S' contain exactly n and m k-mers respectively—k-mers are in many ways the natural unit of measurement, rather than individual characters. We model only point substitutions here and not indels. Independent substitution models have been considered in theoretical work (Blanca et al., 2022; Shaw and Yu, 2022; Ganesh and Sy, 2020), and importantly, also demonstrated to be useful empirically (Chaisson and Tesler, 2012; Ondov et al., 2016; Sarmashghi et al., 2019). Also, while genomes can be repetitive, on the level of k-mers a random model has been shown to be reasonable (Fofanov et al., 2004). We discuss other possible random models in the Discussion section.

Modelling seed-chain-extend A brief overview of seed-chain-extend based alignment is given as follows: first a subset of k-mers in both S, S' are taken as *seeds* and exact seed matches between

S, S' called *anchors* are obtained. We only use k-mer seeds in this study, although other types of seeds are possible (Keich et al., 2004; Kiełbasa et al., 2011). An optimal increasing subsequence of possibly overlapping anchors based on some score is then collected into a *chain*, where increasing is defined with the standard precedence relationship (Jain et al., 2022) between k-mer anchors (See Figure 5a and **Chaining** below). The chain is *extended* into a full alignment by aligning between anchor gaps in the chain.

Model overview: our model of seed-chain-extend is primarily inspired by minimap2 with a few key differences. It captures the following steps: seeding the *query* S', matching the k-mers to obtain anchors, sorting the anchors, chaining, and extending. We assume the *reference* S has already been seeded and only the query needs seeding, as in the case of read alignment. For comparing two similar length genomes, the seeding time of either genome is comparable, so the asymptotics will be the same for comparative genomics.

Runtimes: non-sketched seeding runtime with a hash table is $O(m)$, whereas sketched seeding runtime is $O(mk)$ (discussed in Section 4.5). Letting N be a random variable for the number of anchors, matching is $O(N + m)$ by iterating through a hash table, sorting is $O(N \log N)$, and (optimal) chaining is $O(N \log N)$ (see Section 2.1). Extension is the only step with unknown time complexity. It will turn out that extension and $N \log N$ are the dominating asymptotic terms, so our goal is to bound these terms in expectation. Empirically, it has been shown that chaining and alignment usually take the most time (Kalikar et al., 2022) for read mapping.

Chaining: a chain is a sequence of tuples $\mathcal{C} = ((i_1, j_1), \ldots, (i_u, j_u))$ where i_ℓ and j_ℓ are the starting positions of the anchoring k-mers on S and S' respectively, under the convention that $S' = y_{p+1} \ldots$ so the k-mer labelled $(p+1)$ on S' is actually the first k-mer. The precedence relation $i_\ell > i_{\ell-1}$ and $j_\ell > j_{\ell-1}$ must hold for all ℓ, and k-mers can overlap. Our chaining score is the L1 or linear gap cost (Abouelhoda and Ohlebusch, 2005; Li et al., 2020) of the form $u - \zeta[(i_u - i_1) + (j_u - j_1)]$ for $\zeta > 0$, which penalizes long chains from distant spurious anchors and is necessary for our proofs when $n > m$. The score is sometimes defined equivalently as $u - \sum_{\ell=2}^{u} \zeta[(i_\ell - i_{\ell-1}) + (j_\ell - j_{\ell-1})]$. In the language of (Abouelhoda and Ohlebusch, 2005), we let our anchor fragments have length 1, so the k-mers can overlap. While minimap2 v2.22's default chaining score is different and uses a heuristic banded chaining approach, it does use a linear gap cost (without overlaps) in certain situations, e.g. mapping long contigs (Li, 2021).

Extension: we use quadratic time extension between gaps based on any alignment score (e.g. edit distance, affine gap costs (Durbin et al., 1998)) as our optimality criterion only depends on the chain (Section 4.1). We do not extend past the ends of the chain and *do not* use banded alignment (Chao et al., 1992) in this step, unlike minimap2 (Li, 2018).

Extension and chaining runtimes Given sorted anchors, let T_{Chain} be the time spent finding an optimal chain. T_{Chain} depends on the objective function (Mäkinen and Sahlin, 2020; Jain et al., 2022; Abouelhoda and Ohlebusch, 2005; Otto et al., 2011). Since our gap costs are linear, $T_{Chain} = O(N \log N)$ where N is the number of anchors (Abouelhoda and Ohlebusch, 2005). For extension time T_{Ext}, let $(G_1, G'_1), \ldots, (G_{u-1}, G'_{u-1})$ be the size of the gaps for an optimal chain. G_ℓ indicates the length of the substring between the k-mers $i_\ell, i_{\ell+1}$ on S and G'_ℓ similarly for S'; G_ℓ, G'_ℓ can be zero. The extension time is $T_{Ext} = \sum_{\ell=1}^{u-1} O(G_\ell G'_\ell)$. Using the fact that $\sum G_\ell \leq n$, $G'_\ell \leq \sum G'_\ell \leq m$, one can get that $T_{Ext} = \sum_{\ell=1}^{u-1} O(G_\ell m) = O(nm)$. We will show that the expected runtime is better than this upper bound, but it serves as a useful worst case. Since S, S' are random strings, T_{Chain}, T_{Ext}, N, and the alignment itself all become random variables. Our goal will be to bound $\mathbb{E}[T_{Chain}]$ and $\mathbb{E}[T_{Ext}]$.

2.2 Theoretical results

First, a few definitions are in order. Recall that $|S| = n+k-1 \sim n$, and $|S'| = m+k-1 \sim m$ are our string lengths. We define $\sigma > 1$ to be the size of the alphabet and $0 < \theta < 1$ to be the probability a base mutates. Our theory holds for any alphabet size, but we use $\sigma = 4$ for specifying numerical constants. Let $\log = \log_\sigma$ with base σ and ln be the natural logarithm. Let $k = C \log n$ for a fixed $C > 0$ so that key quantities can be expressed interchangeably as $\sigma^k = n^C$ and $(1-\theta)^k = n^{-C\alpha}$ where $\alpha = -\log_\sigma(1-\theta) > 0$. We define the actual goodness of the chain in terms of *recoverability* in Section 4.1; it measures the fraction of homologous bases under our mutation model that could potentially be recovered by extending through an optimal chain and *only depends on the chain*, not the actual alignment.

Theorem 1 (Main result; no sketching). *Under our model of seed-chain-extend in Section 2.1, assume $\theta < 0.206$, let $\alpha = -\log_4(1-\theta)$, and pick any $C > \frac{2}{1-2\alpha}$. If $m = \Omega(n^{2C\alpha+\epsilon}) < n$ for some arbitrary $\epsilon > 0$, letting $k = C \log n$ and $\zeta = \frac{1}{6g(n)}$ where $g(n) = C\frac{50}{8}\log(n)\ln(n)n^{C\alpha}$, the expected running time is $O(mn^{C\alpha}\log(n))$ for extension and $O(mn^{-C\alpha}\log m)$ for (optimal) chaining. $C\alpha$ can always be chosen to be $< \frac{1}{2}$ and the expected recoverability of any optimal chain is $\geq 1 - O(\frac{1}{\sqrt{m}})$.*

Equivalently, $C\alpha = \frac{2\alpha}{1-2\alpha} + \delta$ for any $\delta > 0$. $\frac{2\alpha}{1-2\alpha}$ is plotted in Supplementary Figure 7; it is convex and thus leads to the bound $C\alpha < 2.43 \cdot \theta$ stated in Section 1.1. E.g., if $|S| = |S'|$ and $\theta = 0.05$, which is approximately the level of divergence seen between human and chimpanzee genomes (Britten, 2002), $\frac{-2\log_4(0.95)}{1+2\log_4(0.95)} < 0.08$, so for $C\alpha = 0.08$ the running time is $O(n^{1.08}\log(n))$. Even for relatively large values of n, say the entire human genome which has size of approximately 3 billion, $n^{1.08}$ is essentially linear as $(3,000,000,000)^{0.08} \sim 5.73$.

Our proof follows in three main steps. First, we bound the first and second moments of the random variable N, which denotes the number of anchors, implying that chaining is fast. Second, we use concentration inequalities for sums of *dependent* random variables and exploit the structure of chaining to show that with high probability, an optimal chain does not deviate much from the chain with only "homologous anchors" (Figure 5a). The failure probabilities will be on the order of $\Theta(\frac{1}{n})$, allowing us to also bound everything in expectation. Lastly, we bound the expected runtime of extension through gaps between homologous anchors.

Asymptotically, the chaining runtime is smaller than extension runtime. However, the implied constants can be much smaller for the extension term (Kalikar et al., 2022), so it is practically useful to reduce the runtime of chaining via sketching. Let $0 < 1/c \leq 1$ be the expected fraction of selected k-mers for a sketching method. We use the *open syncmer* k-mer seeding method (Edgar, 2021) which has a useful mathematical property (Theorem 7), giving:

Theorem 2 (Sketched result). *In addition to the hypotheses outlined in Theorem 1, let $c = O(\log n) < k$ and $\zeta = \frac{1}{6g'(n)}$ instead where $g'(n) = C\frac{200}{8}\log(n)\ln(n)(\frac{3}{2}n^{C\alpha}+2)$. For open syncmer sketched seed-chain-extend, the expected running time is $O\big(\min(\frac{1}{c^2}mn^{C\alpha}\log^3(n), c \cdot mn^{C\alpha}\log n)\big)$ for extension and $O(\frac{1}{c}mn^{-C\alpha}\log m)$ for chaining. The expected recoverability of any optimal chain is $\geq 1 - O(\frac{1}{\sqrt{m}})$.*

This shows that the asymptotic upper bound on extension runtime is the same even if we let c grow with n like $c = \Theta(\log n) < k$, leading to the following conclusion: *sketching can reduce chaining time without increasing extension time much.* Other seeding schemes used, e.g. minimizers (Sirén et al., 2021; Rautiainen and Marschall, 2020; Li, 2018; Colquhoun et al., 2021) or FracMinHash (Irber et al., 2022) behave differently, but our techniques provide intuition and in some cases can be extended (See Section 4.5).

2.3 Simulated genome alignment experiments

Although the model of seed-chain-extend we consider in the theory is based off of real aligners (e.g. minimap2), real aligners implement many additional tricks to make things work better. As such, to empirically validate our theory, we implemented a basic version of sketched seed-chain-extend and tested it on simulated random sequences with independent point substitutions. For the chaining step, we implemented an AVL tree based max-range-query method (Mäkinen et al., 2015; Li et al., 2020). For the extension step, we used a standard dynamic programming (DP) algorithm implemented in rust-bio (Köster, 2016).

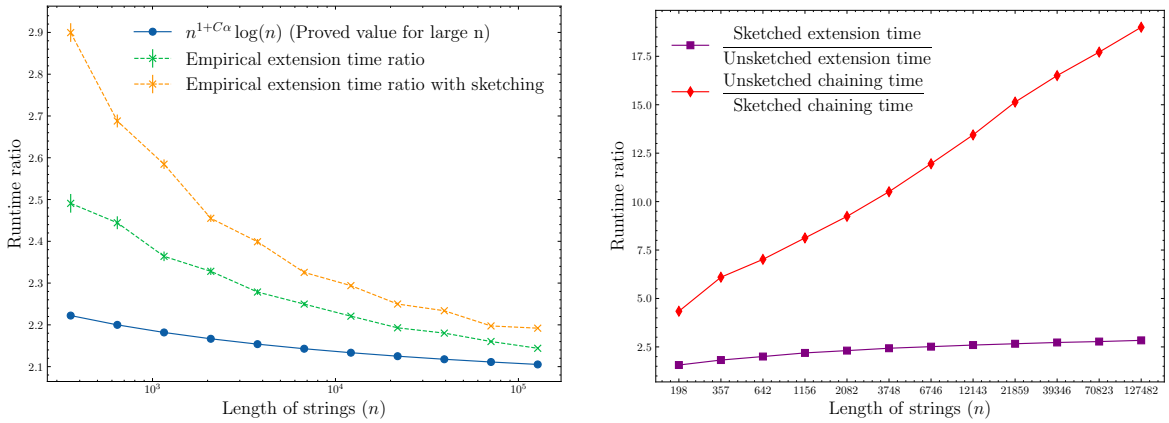
We let $k = C \log n$ where $C = \frac{2}{1-2\alpha}$, and do seed-chain-extend on two length n sequences in this section, but we do an additional experiment for substrings of length $m = n^{0.7} + 100$ in Supplementary Figure 10. As k cannot be fractional, we let $n_k = 4^{\frac{k}{C}}$ for varying integer values of k. We will set $c = k - 7 = C\log(n) - 7$ which *grows with* n. The constant 7 is arbitrary but chosen sufficiently large that so s-mers in the open syncmer method are not so small as to make the method degenerate. We found that recoverability was always quite high and that breaks did not occur very often in actual simulations; we show this in Supplementary Figure 9. Thus the primary focus of the empirical results will be runtime. We used 50,000 iterations for every data point.

Accuracy of asymptotic extension runtime predictions We first empirically investigate our upper bound on expected extension runtime, which is $O(n^{1+C\alpha}\log(n))$ for both sketched and non-sketched extension when $c = \Theta(\log n)$. Assuming that the runtime is simply $\lambda n_k^{1+C\alpha}\log(n_k)$ for some fixed constant λ, we can predict the ratio of the runtimes as $\frac{n_{k+1}^{1+C\alpha}\log(n_{k+1})}{n_k^{1+C\alpha}\log(n_k)}$. Of course, this is incorrect for small n, as smaller terms may dominate runtime, but we expect it to be reasonably accurate for large n. We plot the empirical and predicted ratios of extension runtimes in Figure 1 (a).

Figure 1 (a) shows that our upper bound looks reasonable for both sketched and non-sketched cases for $\theta = 0.10$. In Supplementary Figure 8, we show the same plots for $\theta = 0.05$. The empirical results never cross the predicted ratio, but this is not too surprising as the predicted ratio is only approximate. Importantly, the empirical extension runtime ratios slope downwards, which is in agreement with the prediction.

Sketching with $\Theta(\frac{1}{\log n})$ density gives favourable runtime tradeoffs One of our key results in Theorem 2 is that the upper bound on asymptotic runtime of extension does not depend on the density as long as $1/c < k$, but chaining speed-up scales multiplicatively with c. In Figure 1 (b), we plot the multiplicative speed-up and slow-down of chaining and extension runtimes. This figure shows why sketching is so effective in practice with respect to runtimes; the maximal extension slow-down plateaus to ≤ 3 times in this case, whereas the chaining speed up is linear as the string grows exponentially. Therefore it is worth sketching aggressively to reduce runtime if chaining is slow. One should however still be careful of the sensitivity loss due to sketching in practice (Shaw and Yu, 2022).

In practice, because extension is heavily optimized (Marco-Sola et al., 2021; Farrar, 2007; Suzuki and Kasahara, 2018) and repetitive k-mers lead to more anchors, chaining can be a bottleneck even though it is asymptotically faster in runtime. In fact, we tried using WFA (Marco-Sola et al., 2021), a highly optimized algorithm for extension instead of the generic DP algorithm and found it was ~ 60 times faster than chaining without sketching for $n \sim 2,800,000, k = 23, \theta = 0.05$ in our implementation.



(a) Runtime analysis of extension. The y-axis shows the runtime ratios (with 95% confidence intervals) for iteration i divided by runtime for iteration $i-1$ where the sequence length n_i is plotted on the x-axis. As n grows, the ratio seems to asymptotically approach our conjectured bound.

(b) Multiplicative speed-up for sketched versus non-sketched alignment when sketching with density $\frac{1}{c} = \frac{1}{k-7}$ where $k = C\log n$. The slow-down for extension flattens out, whereas the chaining speed-up is almost linear on the log scaled x-axis.

Fig. 1: Runtime results on seed-chain-extend between for two sequences of length n and $k = C\log n = \frac{2}{1-2\alpha}\log(n) = 9, 10, \ldots, 19$ where $\theta = 0.10$ and $\alpha = -\log_4(1-\theta)$. In (a), both sketched/non-sketched extension ratios approach the predicted $n^{1+\frac{2}{1-2\alpha}\alpha}\log(n)$ runtime ratio. In (b), chaining speed up dominates extension slow down as predicted by our theory.

2.4 Real nanopore read alignment experiments

To test the applicability and generalizability of our theorems to real, non-simulated data, we performed an experiment by aligning real long-read data from Oxford Nanopore Technologies (ONT) using seed-chain-extend. We use the sketched seed-chain-extend aligner in the simulated experiments with a slight modification (see the next section). Notably, these reads are smaller than the genome length. Importantly, our two main assumptions with respect to uniform random strings and point substitutions are violated in this setup: real genomes are not uniformly random strings, and ONT reads contain a significant amount of *indel errors* (Delahaye and Nicolas, 2021).

Our data set consists of five reference genomes and corresponding long-read data sets. The references consist of a viral genome (SARS-CoV-2), a bacterial genome (E. coli), a fungi genome (M. oryzae), an insect genome (D. melanogaster), and a human genome (H. sapiens). For each genome, we aligned a corresponding set of publicly available ONT long-reads which may have different length and error distributions. The data sets can be found in Supplementary Table 1.

We are interested in testing the predicted results as a function of m (the read length) and n (the genome size). We let $k = C\log n = \frac{2}{1-2\alpha}\log n$ as before. Because α and C in our results depend on the error rate θ, we first aligned each read set to the genomes with minimap2 and retained only reads with gap-compressed identity $> 94.5\%$ and $< 95.5\%$ so that $\theta \approx 0.05$ is fixed. In addition, we only used reads for which minimap2 aligned $> 90\%$ of the read, as to avoid benchmarking on unalignable reads.

Practical linear-gap cost consideration We use almost the same seed-chain-extend algorithm as in Section 2.3 with fixed density 1/7 sketching (i.e. not scaling with n), except we apply one change: we found that the predicted linear-gap cost $\zeta(n)$ given by our theorems, while asymptotically

appropriate in theory for large n, was too small for reasonable n. Therefore, we multiplied ζ by 1000.

This discrepancy is because the bounds we use in our proofs do not have optimized constants. The exact (non-asymptotic) value of ζ is only used in proving the recoverability result, and our change does not modify ζ's asymptotic behavior, which is used in other parts of our proofs. Thus our runtime results are still technically valid, while our experiments below will also suggest that the recoverability result is also still valid.

Evaluating model assumptions We first introduced the following filters to our basic seed-chain-extend alignment algorithm for computational reasons:

1. We filtered reads where the number of anchors is > 10 times the number of bases. This occurs due to extremely repetitive k-mers, and can cause the chaining step to stall.
2. We did not consider reads where the chains had gaps of length > 10000 bp, which could occur due to structural variations or failure of chaining and can cause extension to stall.

The two filters also measure how badly our model assumptions, with respect to independent substitutions and uniform random strings, are violated. We report the number of reads that fail filters (1) and (2) in Table 1. Only a small fraction of reads fail the 10000 bp gap filter, but the repetitive k-mer filter is violated more frequently as the genomes become more complex. Still, $> 80\%$ of the human reads and $> 97\%$ of the D. melanogaster reads are not deemed too repetitive. Mapping to repetitive regions is an active area of research (Jain et al., 2020) which is out of the scope of this article, but effective heuristics such as repetitive k-mer masking (Li, 2018; Sahlin, 2022) exist.

	# reads with >10 kb gap in chain	# reads with number of anchors > 10 times number of bases	Total # of reads
SARS-CoV-2	0 (0%)	0 (0%)	1692
E. coli	0 (0%)	0 (0%)	8223
M. oryzae	21 (0.39%)	3 (0.06%)	5424
D. melanogaster	25 (0.56%)	116 (2.61%)	4439
H. sapiens	11 (0.07%)	3004 (19.48%)	15419

Table 1: Counting the number of reads which did not pass our gap filter and repetitive k-mer filter. These two statistics measure how badly our model breaks down with respect to repetitiveness, large variation, and chaining failure. The total number of reads only counts a portion of the original reads that pass our filters, i.e. 90% of the read is aligned by minimap2 and the sequence divergence is $5\% \pm 0.5$.

Extension and chaining runtimes are well predicted by theory In Figure 2, we plot the times of chaining and extension and perform a robust linear regression using the Siegel estimator (Siegel, 1982) as a function of the read length m. For this plot specifically, we only timed alignments where the read was $> 90\%$ aligned by our seed-chain-extend implementation. For each plot, the genome size n is held fixed. For fixed n, all runtimes are well-approximated by a linear

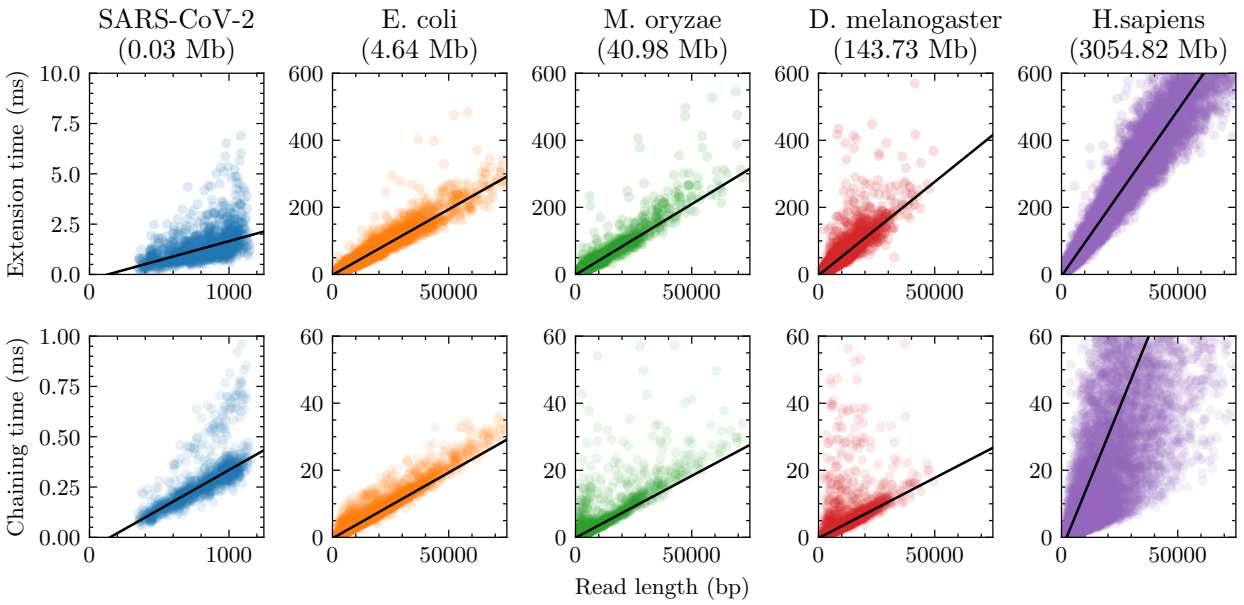


Fig. 2: Extension and chaining time for real ONT reads with approximately 5% sequence divergence. k-mer size was chosen to be $k = C \log n$ for reference length n and $C = \frac{2}{1-2\alpha} \approx 2.16$ with error parameter $\theta = 0.05$. Exact data sets are described in Supplementary Table 1. Note the scaled axes for the SARS-CoV-2 reads, which were much smaller than the other data sets. The well-fit linear regression lines indicate essentially linear runtime in read length with fixed reference length n (i.e. fixed $k = C \log n$ and constant C), although larger human chaining time variance is due to repetitive k-mers.

function in m as predicted by our theory, although chaining times for the human genome have higher variance due to the presence of repetitive k-mers.

In Figure 3, we plot the slopes of the regressions for extension time obtained from the first row of Figure 2, which measures the dependence on n given fixed m. This dependence is reasonably fit by a $\log n$ function ($R^2 = 0.766$), so $O(m \log n)$ is not a bad approximation for the runtime in practice. However, our theory leads to a $n^{C\alpha} \log n$ dependence, which comes from the big O extension runtime. $C\alpha \sim 0.08$ when $\theta = 0.05$ in our setup, so by using a $n^{0.08} \log n$ function instead, we end up getting a better fit ($R^2 = 0.928$), as predicted by our theory.

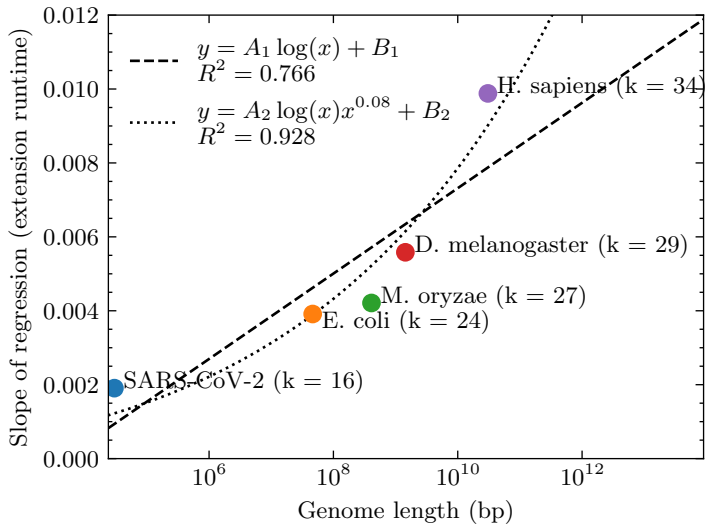


Fig. 3: Data points represent the slopes of the linear regressions in Figure 2 for extension, with the value of k (which is $k = C \log n$ for a constant $C = \frac{2}{1-2\alpha}$) labelled. This dependence of the genome size, n (x-axis), is decently approximated by a naive $A_1 \log(n) + B_1$ fit where A_1 and B_1 are paraeters. However, our theory states that the dependence should be $\log(n) n^{C\alpha}$ with $C\alpha \sim 0.08$ when $\theta = 0.05$. Fitting $A_2 \log(n) n^{0.08} + B_2$ gives a better R^2 value (0.928 vs 0.766) with the same number of parameters (2 parameters for both fits), indicating the goodness of our theoretical predictions.

Recoverability upper bound by aligned fraction Recoverability is not measurable on real, non-simulated reads because true sequence homology is not known. However, we can upper bound recoverability by simply measuring the length of the chain relative to the read length, as any homologous bases outside of the chain (i.e. not within the first and last anchor) can not be recovered under our model. We call this the aligned fraction, which our theory also predicts should be $> 1 - O(\frac{1}{\sqrt{m}})$ in expectation (with implied constants depending on θ and n). Practically, the aligned fraction is easily interpretable, and the ability to predict the aligned fraction is useful for ensuring that reads with a given error rate and k-mer size are long enough to align to a reference.

We also observed that the aligned fraction was the main cause of recoverability loss in the simulated experiments (see Supplementary Figure 9), and it turns out that the main $O(\frac{1}{\sqrt{m}})$ term in the theoretical recoverability result exactly comes from bounding the aligned fraction in the proof (Supplementary Lemma 7), so this quantity is interesting theoretically as well.

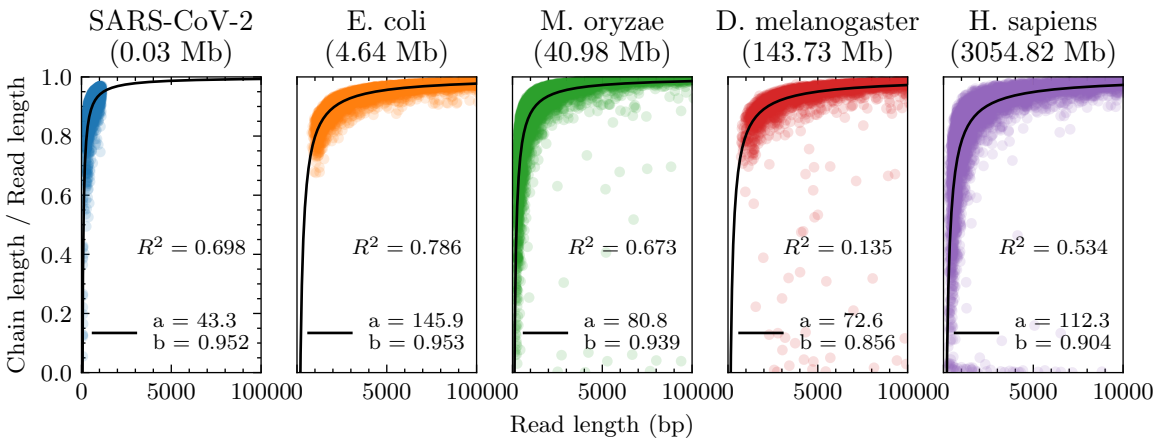


Fig. 4: The aligned fraction of the aligned reads, which is chain length divided by read length. We fit a function of the form $1 - \frac{a}{m^b}$ where m is the read length (x-axis) and display the resulting R^2 values. Aligned fraction is an upper bound for recoverability, and our upper bound on recoverability (and also aligned fraction) is $1 - O(\frac{1}{\sqrt{m}})$, which the data suggests may be too conservative.

In Figure 4, we plot the aligned fraction for each data set. We perform a least-squares fit of the function $1 - \frac{a}{m^b}$ where a and b are parameters. We exclude points with < 0.25 aligned fraction if the read length is > 1000 when fitting because these points appear to be outliers caused by poor chaining and can bias the least-squares fit. The resulting fits for all plots are reasonable, with $R^2 > 0.5$ except for the D. melanogaster dataset. The seemingly small value of R^2 can be explained by the fact that most reads (passing our filters) are well-aligned, so the data is almost constant and ≈ 1. Constant data will have $R^2 = 0$, so we can not do much better in this case. Nevertheless, $b > 0.85$ on all data sets, suggesting that our $\sqrt{m}$ term may be too conservative. However, the exact parameter values in the fit are highly sensitive to the fitting methodology, so the exact values displayed are only suggestive and not to be overinterpreted.

3 Discussion

In this work, we are able to *rigorously* justify the empirical results seen by the seed-chain-extend heuristic through average case analysis under a simple mutation model. We showed the alignment is both accurate and fast: $\geq 1 - O(\frac{1}{\sqrt{m}})$ fraction of the sequence homology is recoverable from the chain while only running in $O(mn^{C\alpha}\log(n))$ time, where for even a moderate mutation rate $\theta = 0.05$, $C\alpha < 0.08$. Interestingly, a recent aligner (Koerkamp and Ivanov, 2022) empirically achieved this predicted $n^{1.08}$ runtime for $\theta = 0.05$ on two length n sequences using similar but distinct techniques. In addition, we also proved that one can sketch to arbitrary densities $1/c$ where $c < k = O(\log n)$ while asymptotically decreasing runtime and without asymptotically decreasing recoverability, justifying the effectiveness of sketching. Because our set-up is modeled by techniques used by practical software such as minimap2, our results provide a theoretical backing for why modern sequence alignment software actually performs so well in practice.

We verified our theoretical results in a synthetic experiment, which showed that our main theoretical predictions were valid for our random model. Specifically, our strongly sub-quadratic bounds predict runtime well, sketching can decrease chaining time without increasing extension time too much, and recoverability is high. Importantly, our results generalize past our uniform string and point substitution assumptions; we show that the runtimes for aligning real nanopore

reads to references are still well-approximated by our theory (see Figure 3). However, on more complex, repetitive genomes, our runtime predictions for chaining break down due to real k-mers being much more repetitive than under a uniform random model.

In terms of further work, we propose two possible directions. The first is to optimize the bounds obtained for the current random model proposed in this work, which assumes uniformly random strings and point substitutions. The second is to generalize the random model to model complex genomes more accurately, e.g. repetitive k-mers.

3.1 Improving bounds for the current random model

For runtime, it seems unlikely that the expected runtime is truly quasilinear due to the fundamental quantity $(1-\theta)^k = (1-\theta)^{C \log n} = n^{-C\alpha}$, the likelihood of a k-mer match, decreasing faster than logarithmically in n when $k = \Theta(\log n)$; this turns out to be the cause of the $n^{C\alpha}$ term in the runtime (see Methods). In spite of this, we believe there can be significant improvements for our bounds in Theorems 1 and 2. The most significant aspect which we believe can be improved is the restriction on the constant C. C is required to be $> \frac{2}{1-2\alpha}$, which seems unsatisfactory because this leads to high k-mer sizes, e.g. $k = 34$ for $\theta = 0.05$ on the human genome. This is much larger than the k-mer sizes used by noisy long-read aligners in practice (minimap2 defaults to $k = 15$).

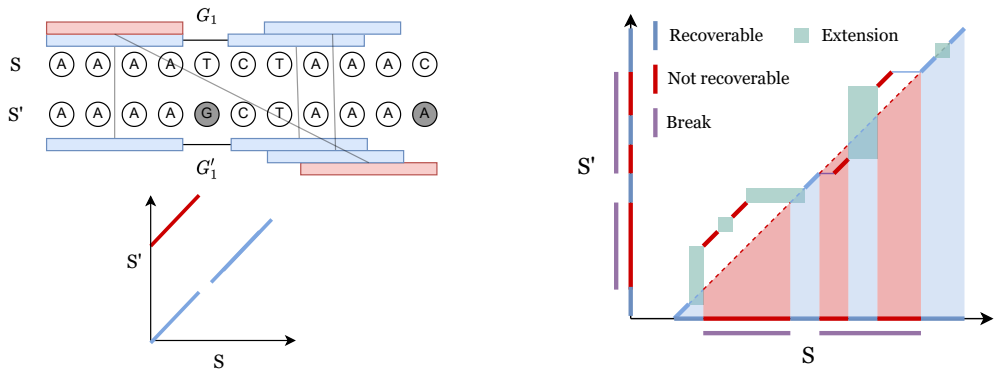
The restricted values of C is due to our analysis of spurious anchors relying on a relatively weak variance bound (Lemmas 3 and 4). To use the bound effectively, C must be quite large. The variance bound is also the cause for the $1 - O(\frac{1}{\sqrt{m}})$ bound for recoverability, which we also believe can be tightened based on both the simulated and real experiments. To surpass these bounds, a deeper understanding of spurious anchor subsequences must be developed. Spurious subsequences of anchors is a very similar problem to common subsequences in random strings for extremely large alphabets, a topic that has had much theoretical attention (Chvátal and Sankoff, 1975; Kiwi et al., 2005; Lember and Matzinger, 2009; Navarro, 2001).

3.2 Generalizing the random model

Our current random model does not model complex genomes properly. It is well known that genomes can be extremely repetitive (Koning et al., 2011), which violates the uniform random string assumption. This is why the expected runtime of chaining deviates from our predictions in Figure 2 for the human genome. One idea for modelling repeats is to use a k-mer Markov model for the random strings instead of an independent uniform string. However, the analysis of such models is highly non-trivial (Reinert et al., 2000), and our techniques for analyzing k-mers in this work may not apply. Regardless of how repeats are modelled, any sort of non-trivial analysis would still require a better handling of spurious anchor subsequences, which is exactly the issue that needs to be tackled in order to improve the bounds for our current model.

Another direction is to incorporate indels into the random model. The technical reason we opted for modelling point substitutions instead of indels is that indels complicate the statistics of k-mer matching. For example, an indel of the nucleotide A may not "change" the 4-mers in the sequence AAAAAA, and sequence homology becomes ambiguous. A formulation of an "indel channel" as a model of sequence mutation was used by Ganesh and Sy (2020) for analysis of edit distance, but the resulting analysis for this model was more complex than in the independent substitution case.

Generalizing our model to small indels seems less pressing as our theory was successful for real nanopore data, which has many small indel errors, showing that the substitution model already generalizes well to small indels. However, modelling large structural variations could be an interesting and relevant problem.



(a) k-mer anchors under mutations and their corresponding alignment matrix. Blue anchors are "homologous anchors" while red are "spurious anchors".

(b) Recoverable bases correspond to the bases on the homologous diagonal where a k-mer lies or is accessible by the green DP matrix between gaps. Breaks cover non-recoverable sections.

Fig. 5: Seed-chain-extend visualized on an alignment matrix. Anchors are short diagonal matches in the matrix. Extension corresponds to performing dynamic programming (DP) on the sub-matrix between anchors. $Align(\mathcal{C})$ consists of anchors and DP matrices in (b).

4 Methods

4.1 Homology and recoverability

Let the interval $[a..b]$ be the discrete interval $\{a, a+1, \ldots, b\}$ and $S[a..b]$ be the substring indexed by $[a..b]$. S' is a mutated version of the substring $S[p+1..p+m]$ of length m starting at $p+1$, so the optimal alignment should map S' to the homologous indices $[p+1..p+m]$. We now define *recoverability* as the number of homologous bases one can possibly recover by seed-chain-extend; a visual example is shown in Figure 5b.

Definition 1. *Let an alignment matrix be given as* $[1..|S|] \times [p+1..p+|S'|]$. *Define the homologous diagonal* $D_H = \{(p+1, p+1), \ldots, (p+|S'|, p+|S'|)\}$. *Given a chain* $\mathcal{C}$, *let* $Align(\mathcal{C})$ *be the set of possible alignments obtained by extending through gaps and k-mer matching, where* $(x, y) \in Align(\mathcal{C})$ *indicates* x *and* y *can be aligned. Then the set of recoverable bases is* $Align(\mathcal{C}) \cap D_H$ *and* $R(\mathcal{C}) = \frac{|Align(\mathcal{C}) \cap D_H|}{|S'|}$.

We define $Align(\mathcal{C})$ carefully in the appendix. Note that an optimal chain with respect to our linear gap cost objective $u - \zeta[(i_u - i_1) + (j_u - j_1)]$ does not directly optimize for recoverability. We provably find the chain with the optimal linear gap cost, and then we will argue that it leads to good recoverability.

Traditionally, alignment optimality is defined using a generalized edit distance (Vingron and Waterman, 1994). However, these distances are only used as a proxy for detecting sequence homology (Batzoglou, 2005; Thorne et al., 1991; Lunter et al., 2005). We know the true underlying sequence ancestry in our model, so defining optimality with respect to sequence homology suits the actual goal of sequence alignment. The reason for the name "recoverability" is that extension could *potentially* recover all recoverable bases, but this depends on the extension algorithm and is not guaranteed (Lunter et al., 2008).

Under our model, the trivial $O(1)$ alignment that aligns S' back to the originating substring of S without indels is the most homologous. This may seem to make our results superfluous; however,

remember that the algorithm does not know where S' "begins" if $m < n$. Also, while we do not attempt the case with indels, seed-chain-extend is still valid when indels are present whereas the trivial solution does not allow for indels.

We will lower bound $\mathbb{E}[R(\mathcal{C})]$. To do this, we will work with a more natural object called a *break*.

Definition 2 (inspired by Ganesh and Sy (2020)). *We call matching bases and anchors of the form (x, x) homologous and spurious otherwise. Given a chain $((i_1, j_1), \ldots, (i_u, j_u))$ and a maximal interval $[p..q]$ such that $(i_p, j_p), ..., (i_q, j_q)$ are all spurious anchors, define the break B as $B = [\min(i_p, j_p).. \max(i_q, j_q) + k - 1]$. Let the length or size of a break be $L(B) = \max(i_q, j_q) - \min(i_p, j_p) + k$ and $L(\mathcal{C}) = \sum_B L(B)$ be the total length over all breaks.*

Lemma 1. *Given any chain $\mathcal{C} = ((i_1, j_1), \ldots, (i_u, j_u))$, we have that $(j_u - j_1 - L(\mathcal{C}))/|S'| \leq R(\mathcal{C})$.*

We prove Lemma 1 in the appendix. The concept of a break is illustrated in Figure 5b. Breaks cover non-recoverable regions, so subtracting the breaks from the span of the anchors lower bounds the recoverability.

4.2 Fundamental tools and bounds

In this section, we describe some fundamental tools for dealing with pairs of random mutating strings. We first need to give a careful probabilistic exploration of random k-mer anchors on $S = x_1 x_2 \ldots$ and $S' = y_{p+1} y_{p+2} \ldots$. This requires a bit of work due to the dependence between the random strings S, S'. For the rest of the paper, missing proofs can be found in the appendix.

Definition 3. *Let $M(i, j)$ be random variables such that $M(i, j) = 1$ if $x_i = y_j$ and 0 otherwise. Let $A(i, j)$ be a random variable $\prod_{\ell=0}^{k-1} M(i+\ell, j+\ell)$. $A(i, j)$ is an indicator random variable for the presence of a k-mer anchor at positions (i, j). We will also refer to $A(i, j)$ variables as "anchors".*

We would like for $M(i, j), M(i+1, j+1), \ldots$ to be independent, so finding the probability of k-mer matches is easy. However, in our model, it is not actually obvious *a priori* that $M(i, j), M(i+1, j+1)$ are independent when $j \neq i$. Consider $M(1, 2)$ and $M(2, 3)$. The former is a function of x_1, y_2, and the latter x_2, y_3. However, x_2 and y_2 are dependent in our model, so more work is needed. A convenient graphical representation of independence for the M random variables is the *match graph*, which we define below.

Definition 4. *The match graph of a set of random variables of the form $\mathcal{M} = \{M(i_1, j_1), M(i_2, j_2), \ldots\}$ is a graph $G(\mathcal{M}) = (V, E)$ where the vertices V are the letters $x_1, \ldots, x_{n+k-1}, y_{p+1}, \ldots, y_{p+m+k-1}$, and the edges*

$$E = \{(x_i, y_i) : i \in [p + 1..p + m + k - 1]\} \cup \{(x_h, y_l) : M(h, l) \in \mathcal{M}\}.$$

In particular, the match graph is bipartite for the sets $\{x_1, \ldots, x_{n+k-1}\}$ and $\{y_{p+1}, \ldots, y_{p+m+k-1}\}$.

A match graph is shown in Supplementary Figure 1. The main reason for defining the match graph is the following theorem, which allows us to graphically determine independence of the A variables.

Theorem 3. *The random variables $\mathcal{M} = \{M(i_1, j_1), M(i_2, j_2), \ldots\}$ where $i_\ell \neq j_\ell$ for all $M(i_\ell, j_\ell) \in \mathcal{M}$ are independent if the induced match graph has no cycles.*

Corollary 1. *If $i \neq j$, $\Pr(A(i, j) = 1) = \frac{1}{\sigma^k}$. Otherwise, $\Pr(A(i, i) = 1) = (1 - \theta)^k$.*

We will denote the random variables $\sum_{i=p+1}^{p+m} A(i,i) = \sum_i A(i,i) = N_H(n,m) = N_H$ and $\sum_{i\neq j} A(i,j) = N_S(n,m) = N_S$ respectively as the number of homologous anchors and spurious anchors; we will drop the dependence on n, m to simplify notation. These are key random variables that we wish to bound later on. The below theorem follows directly from Corollary 1 by linearity of expectation.

Theorem 4. *$\mathbb{E}[\sum_i A(i,i)] = \mathbb{E}[N_H] = m(1-\theta)^k$, and $\mathbb{E}[\sum_{i,j,i\neq j} A(i,j)] = \mathbb{E}[N_S] = m(n-1)\frac{1}{\sigma^k}$. In particular, $\mathbb{E}[N] = \mathbb{E}[N_H + N_S] = m(1-\theta)^k + m(n-1)\frac{1}{\sigma^k}$.*

4.3 Bounding sums of k-mer random variables

We proceed to bound the *distribution* of the random variables N_S and N_H, which are sums of *dependent* random variables. We first bound N_S by computing the second moments and using variance based bounds. To do this, we need to examine the independence structure of the $A(i,j)$ random variables.

Lemma 2. *For $A(i,j)$ and $A(h,l)$, if both of the following conditions hold:*

1. *$|i-h| \geq k$ or $|j-l| \geq k$ and*
2. *$|i-l| \geq k$ or $|j-h| \geq k$,*

then the induced match graph on the M variables for $A(i,j)$ and $A(h,l)$ has no cycles.

Intuitively, the first condition states that two anchors do not overlap too much, e.g. $A(1,1)$ and $A(2,2)$ are not independent when $k = 3$. The intuition behind the second condition can be illustrated by the following situation where $k = 1$ and $\theta \sim 0$: consider the anchors $A(1,5), A(5,1)$. Since it is likely that $x_1 = y_1$ and $x_5 = y_5$, if $x_1 = y_5$ then $x_5 = y_1$ with high probability, so $x_1 = y_5$ is not independent of $x_5 = y_1$.

Corollary 2. *If $A(i,j)$ and $A(h,l)$ satisfy Lemma 2, they are independent.*

Proof. By condition (1) in Lemma 2, $A(i,j)$ and $A(h,l)$ do not share any M variables i.e. $A(i,j) = M(i,j)M(i+1,j+1)\ldots$ and similarly for $A(h,l)$, but no product shares a variable with the others. Since the match graph has no cycles under these conditions, by Theorem 3, all M variables are independent so $\Pr(A(i,j)A(h,l) = 1) = \frac{1}{\sigma^{2k}} = \Pr(A(i,j) = 1)\Pr(A(h,l) = 1)$ as desired.

Lemma 3. *$\mathbb{E}[N_S^2] \leq 8k^2 mn\frac{1}{\sigma^k} + \mathbb{E}[N_S]^2$. Thus the variance $Var(N_S)$ can be upper bounded by $8k^2\frac{mn}{\sigma^k}$. Furthermore, $\mathbb{E}[N_H^2] \leq 2mk(1-\theta)^k + m^2(1-\theta)^{2k}$ and $\mathbb{E}[N_H N_S] \leq 4k\frac{mn}{\sigma^k} + m^2 n(1-\theta)^k\frac{1}{\sigma^k}$.*

Now we can use the variance bound and Chebyshev's inequality to get the result below. Note the bound uses $k = C\log n$; we will prefer this form for quantities directly used for proving the main result. The label F1 in the theorem refers to the particular event space for which the bound always holds. We will label each proposition that holds with high probability with the event space that we are operating in. We will continue this convention for the rest of the paper as this will be useful when computing our final bounds.

Lemma 4 (F1). *With probability $\geq 1-\frac{1}{n}$, the number of spurious anchors is $\leq n^{2-C} + \sqrt{8m}C\log(n)n^{1-C/2}$ where we used $\frac{m(n-1)}{\sigma^k} < n^{2-C}$. That is,*

$$\Pr\Big(N_S \geq n^{2-C} + \sqrt{8m}C\log(n)n^{1-C/2}\Big) \leq \frac{1}{n}.$$

If $C > 3$, then for large n, $N_S = 0$ with high probability, and our analysis would be easy. But, we want C as small as possible. It turns out we can make $C \sim 2$ for reasonable θ, significantly tightening our bounds.

For N_H we can get a stronger exponential bound because of its independence structure. $A(i,i)$s, which we call homologous anchors, are only dependent in a small neighbourhood around i of size k because k-mers on non-overlapping substrings are independent. This is called $k-$*dependence* (not to be confused with *k-independence*) and is used in (Blanca et al., 2022) to show N_H is asymptotically normal. Concentration bounds can also be translated in the $k-$dependent scenario (Janson, 2004).

Theorem 5 (Dependent Chernoff-Hoeffding bound - Corollary 2.4 from Janson (Janson, 2004) reworded and simplified). *Suppose we have $X = \sum_{a\in\mathcal{A}} Bernoulli_a(q)$ for some $0 < q < 1$. A proper cover of $\mathcal{A}$ is a family of subsets $\{\mathcal{A}_i\}_{i\in I}$ such that all random variables in $\mathcal{A}_i \subset \mathcal{A}$ are independent and $\bigcup_{i\in I} \mathcal{A}_i = \mathcal{A}$. Let $\chi(\mathcal{A})$ be the minimum size of the cover, $|I|$, over all possible proper covers. Then for $t \geq 0$,*

$$\Pr(X \leq \mathbb{E}X - t) \leq \exp\left(-\frac{8t^2}{25|\mathcal{A}|\chi(\mathcal{A})q}\right).$$

Lemma 5.

$$\Pr\Big(N_H \leq m(1-\theta)^k - t\Big) \leq \exp\Big(-\frac{8t^2}{25mk(1-\theta)^k}\Big)$$

Proof. We simply use Theorem 5 with $q = (1-\theta)^k$. By k-dependence, we can easily see that $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ where $\mathcal{A}_j = \{A(j,j), A(j+k, j+k), A(j+2k, j+2k)\ldots\}$ is a partition satisfying the independence condition, and we will have at most k sets. Thus $\chi(\mathcal{A}) \leq k$, and we're done.

4.4 Proof of non-sketched main result

To prove the main result on seed-chain-extend without sketching, we will need to bound three quantities in expectation: the runtime of chaining, the recoverability of chaining, and the runtime of extension.

We first bound $O(\mathbb{E}[N \log N])$, the expected runtime of chaining (and also anchor sorting). Note that $\mathbb{E}[N] \leq \frac{1}{n^{C-2}} + mn^{-C\alpha}$. We would like $\mathbb{E}[N\log N] = O(\mathbb{E}[N]\log\mathbb{E}[N]) = O(mn^{-C\alpha}\log m)$ to hold. Unfortunately, Jensen's inequality only gives $\mathbb{E}[N\log N] \geq \mathbb{E}[N]\log\mathbb{E}[N]$ because $x\log x$ is convex. However, with a bit more work:

Theorem 6. *Assume $m = \Omega(n^{2C\alpha+\epsilon})$ for some $\epsilon > 0$, and $C > \frac{1}{1-\alpha}$. Letting N be the total number of k-mer anchors, $\mathbb{E}[N\log N] = O(\mathbb{E}[N]\log\mathbb{E}[N]) = O(mn^{-C\alpha}\log m)$. Thus the runtime of chaining is $O(mn^{-C\alpha}\log m)$.*

Now we bound the expected recoverability of the chain. Given S and S', let a *homologous gap* of size $\ell + k - 1$ bases be an interval of ℓ consecutive k-mers for which no homologous anchors exist (i.e. the k-mers are mutated). In the context of a chain, a homologous gap will refer to a gap flanked by two homologous anchors. Technically, if ℓ consecutive k-mers are uncovered but are flanked by two homologous anchors, this gives $\ell - k + 1$ uncovered bases. We will ignore these factors of k as we will show that they are asymptotically small. It turns out homologous gaps grow relatively slowly in n with high probability.

Lemma 6 (F2). *With probability $\geq 1 - 1/n$, no homologous gap has size greater than*

$$g(n) = \frac{50k}{8(1-\theta)^k}\ln(n) = \frac{C\cdot 50}{8}\log(n)\ln(n)\cdot n^{C\alpha}$$

plus a small $C \log n$ term we will ignore because it is small asymptotically.

In a chain, gaps may also be flanked by one or two spurious anchors. We call these *non-homologous gaps*. We first bound break lengths, which will imply good recoverability, and bound non-homologous gaps later on.

Lemma 7 (F1 + F2). *Take any $C > \min(3, \frac{2}{1-2\alpha})$ and let $\zeta = \frac{1}{6g(n)}$ where $g(n) = C\frac{50}{8}\log(n)\ln(n)n^{C\alpha}$. Assume $m = \Omega(n^{2C\alpha+\epsilon})$ for some $\epsilon > 0$. Then for large enough n, there are no breaks of length $\geq m^{1/2}$ with probability greater than $1 - 2/n$ in an optimal chain.*

Corollary 3. *Under the same assumptions as in Lemma 7, the expected recoverability of any optimal chain is $\geq 1 - O(\frac{1}{\sqrt{m}})$.*

The idea behind proving the above propositions is to work in a space of events F1 $\cap$ F2 where "bad events" do not occur and any optimal chain has good recoverability. Because this space of bad events is small, they do not contribute to our expected value too much. This finishes the recoverability result for the main theorem.

The last step is to bound extension running time, and this comes down to separately bounding the size of the homologous and non-homologous gaps in any optimal chain. We bound the runtime of extension through homologous gaps by directly calculating the expectation through *all possible* homologous gaps. We then show that the runtime through non-homologous gaps is small and does not contribute to the asymptotic term.

Lemma 8. *Let T^H_{Ext} be the time of extension over only the homologous gaps of any optimal chain. $\mathbb{E}[T^H_{Ext}] = O(mn^{C\alpha}\log n)$.*

Lemma 9. *Let T^S_{Ext} be the runtime of extension through only the non-homologous gaps of an optimal chain. Under the same assumptions as in Lemma 7, $\mathbb{E}[T^S_{Ext}] = O(m)$.*

Proof (Theorem 1). The expected runtime of chaining follows from Theorem 6. The recoverability result follows from Corollary 3. The expected runtime of extension is $\mathbb{E}[T_{Ext}] = \mathbb{E}[T^H_{Ext}] + \mathbb{E}[T^S_{Ext}]$, and $\mathbb{E}[T^H_{Ext}], \mathbb{E}[T^S_{Ext}]$ are both $O(mn^{C\alpha}\log(n))$ by Lemmas 8 and 9. This completes the proof as long as we satisfy the assumptions of Lemma 7 and Theorem 6 on C, α. To satisfy the assumptions, we require $C > \frac{1}{1-\alpha}$, $C > \min(3, \frac{2}{1-2\alpha})$, and $C\alpha < 1/2$ otherwise $m = \Omega(n^{2C\alpha+\epsilon}) > n$ for large enough n. It's not hard to check that the limiting condition is $\alpha < 1/6$, so we require $-\log_4(1-\theta) < 1/6$. This works out to be $\theta < 1 - 4^{-\frac{1}{6}} < 0.2063$. We can also remove the minimum condition on C because $\alpha < 1/6$ implies $\frac{2}{1-2\alpha} < 3$.

4.5 Sketching and local k-mer selection

Now consider not selecting all of the k-mers in a string, but only a subset of them during the initial seeding step. This allows one to chain only a subset of the k-mers, potentially providing runtime savings.

We use the *open syncmer* method (Edgar, 2021). Given a string, we take all k-mers of the string and break the k-mer into s-mers with $s < k$. There are $k - s + 1$ s-mers in the k-mer. We select or *seed* the k-mer if the smallest s-mer (subject to some ordering, which we choose as uniform random) is in the $\lceil \frac{k-s+1}{2} \rceil$-th 1-indexed position; we call a selected k-mer an *open syncmer*. Given repeated smallest s-mers, we take the rightmost one to be the smallest. Finding the smallest s-mer among the $k - s + 1$ s-mers in a k-mer takes $k - s + 1$ iterations, so finding all open syncmer seeds in S' takes $O((k - s + 1)m) = O(mk) = O(m \log n)$ time.

The expected fraction of selected k-mers over a string with i.i.d uniform letters is called the *density*, and it is $\frac{1}{k-s+1}$ for the open syncmer method (up to a small error term $O(\frac{(k-s+1)^2}{\sigma^s})$ which we will ignore; see (Zheng et al., 2020) or (Shaw and Yu, 2022)). We will let c be the reciprocal of the density, so $c = (k-s+1)$.

The original open syncmer definition in (Edgar, 2021) had a parameter t where a k-mer was selected if the smallest s-mer was in the $t-$th position; we proved in (Shaw and Yu, 2022) that the optimal t is $\lceil \frac{k-s+1}{2} \rceil$ with respect to maximizing the number of conserved bases from k-mer matching. The reason we choose open syncmers is primarily to the following fact which was shown in (Edgar, 2021):

Theorem 7. *Define $t = \lceil \frac{c}{2} \rceil = \lceil \frac{k-s+1}{2} \rceil$. If $k-s+1$ is odd, two consecutive open syncmers must have starting positions $\geq t$ bases apart. If even, they must have starting positions $\geq t-1$ bases apart.*

Theorem 7 follows by examining the smallest s-mer in a k-mer and noticing that in the next overlapping k-mer, the locations for the new smallest s-mer are restricted. This theorem is the reason we use open syncmers and is crucial to our proofs. The spacing property makes selected open syncmers a polar set (Zheng et al., 2021); other methods also give rise to polar sets (Frith et al., 2020, 2022) but open syncmers seem to perform well empirically (Shaw and Yu, 2022; Frith et al., 2022; Dutta et al., 2022) and are easy to describe. For the rest of the section, we will assume $c = k-s+1$ is odd, so $\lceil \frac{c}{2} \rceil = \frac{c+1}{2}$.

Let $A(i,j)$ be the random variables as defined before. Let $J(i)$ be the indicator random variable for if the ith k-mer is selected on S as an open syncmer, and $J'(j)$ similarly for the j-th k-mer on S'. We now wish to calculate $\mathbb{E}[J(i)J'(j)A(i,j)]$, the probability that a k-mer match exists and the k-mer is an open syncmer.

Lemma 10. *If $i = j$, then $\mathbb{E}[J(i)J'(j)A(i,j)] = \frac{(1-\theta)^k}{c}$. Otherwise, $\mathbb{E}[J(i)J'(j)A(i,j)] = \frac{1}{c\sigma^k}$.*

Proof. If $A(i,j) = 1$, then the i-th k-mer and $j-$th k-mer are the same. If a k-mer is selected as an open syncmer on S, it must also be selected on S', so $\mathbb{E}[J(i)J'(j)A(i,j)|A(i,j) = 1] = \mathbb{E}[J(i)] = \frac{1}{c}$. $A(i,j)$ and $J(i)$ are independent because we assume the random ordering for the s-mers is independent of the random mutations. Using the law of total expectation and Theorem 4 gives the result for both cases.

Definition 5. *We will replace all random variables involving anchors and matches with a superscript $*$ to indicate sketched seeds, e.g. $A(i,j)^* = A(i,j)J(i)J'(j)$, M^*, N_S^*, N_H^*, etc.*

Corollary 4. *The expected total number of anchors after applying open syncmer seeds with density $\frac{1}{c}$ is*

$$\mathbb{E}[\sum_{i,j} A(i,j)^*] = \frac{1}{c}\left(n(1-\theta)^k + m(n-1)\frac{1}{\sigma^k}\right).$$

The above follows directly from Lemma 10. As expected, subsampling to a $\frac{1}{c}$ fraction of the k-mers gives $\frac{1}{c}$ expected hits. Note the important property of *context independence* used in the proofs: if $A(i,j) = 1$, then $J(i) = J'(j)$. This property is not satisfied if one were sampling k-mers randomly or using minimizers (Roberts et al., 2004; Shaw and Yu, 2022). We now deduce the sketched moment bounds on N^*.

Lemma 11. *The variance $Var(N_S^*)$ can be upper bounded by $\frac{1}{c}8k^2mn^{1-C}$. Furthermore, $\mathbb{E}[N_H^{*2}] \leq \frac{1}{c}2mk(1-\theta)^k + \frac{1}{c^2}m^2(1-\theta)^{2k}$ and $\mathbb{E}[N_H^*N_S^*] \leq \frac{1}{c}4k\frac{mn}{\sigma^k} + \frac{1}{c^2}m^2n(1-\theta)^k\frac{1}{\sigma^k}$.*

Lemma 12 (F1*). $\Pr\left(N_S^* \geq \frac{1}{c}mn^{1-C} + \sqrt{\frac{8}{c}}km^{1/2}n^{1-C/2}\right) \leq \frac{1}{n}$

The sketched moment bounds allow us to start re-analyzing the crucial propositions in Section 4.4 in the context of sketching. The first main result is that sketching reduces the chaining time as one would expect.

Theorem 8. *Under the same assumptions as in Theorem 6, letting N^* be the total number of sketched k-mer anchors, the expected chaining time is $O(\mathbb{E}[N^* \log N^*]) = O(\frac{1}{c}mn^{-C\alpha}\log m)$.*

The second result we wish to highlight is the new bound on extension runtime through homologous gaps.

Lemma 13. *The expected runtime of sketched extension through the homologous gaps in an optimal chain is $\mathbb{E}[T_{Ext}^{H*}] = O\left(\min\left(\frac{1}{c^2}mn^{C\alpha}\log^3(n), c \cdot mn^{C\alpha}\log n\right)\right)$. If $c = \Theta(\log n)$ then $\mathbb{E}[T_{Ext}^{H*}] = O(mn^{C\alpha}\log n)$.*

It will turn out that in the sketched case, extension over homologous gaps also dominates runtime. Since we sketch with density $\frac{1}{c}$, this result states that we can sketch with *decreasing density* and have the *same* asymptotic extension time as without sketching, which is surprising. The crucial fact we use in the proof of Lemma 13 is Theorem 7, which shows that open syncmer seeding weakens the k-dependence of the seeds since they are now at least $(c+1)/2$ bases apart. This tightens the bound in Theorem 5 and also shows that the sketched maximum homologous gap size is $\Theta(g(n))$, with $g(n)$ as in Lemma 6. If we used other methods such as closed syncmers (Edgar, 2021) or FracMinHash (Irber et al., 2022; Hera et al., 2022), then we would recover the $O(c \cdot mn^{C\alpha}\log n)$ result in Lemma 13 but not the $O(\frac{1}{c^2}mn^{C\alpha}\log^3(n))$ result. Thus, we have saved a log factor by using open syncmers when we let $c = \Theta(\log n)$. The bulk of Section E in the appendix is dedicated to proving Lemma 13.

The key quantities used for the rest of the proofs are N_S^* and $g'(n)$, the new maximum homologous gap size. We will prove $g'(n) = \Theta(g(n))$ with $g(n)$ as in Lemma 6, so the only asymptotic difference in these bounds is a factor of $\frac{1}{\sqrt{c}}$ in N_S^*. Since we want N_S^* to be small, this does not affect downstream analysis. It follows that the proofs of the rest of the lemmas in Section 4.4 can be replicated almost verbatim. We give a sketch of these proofs in the appendix.

5 Software availability

Scripts for regenerating the simulated experiments are available at `https://github.com/bluenote-1577/basic_seed_chainer/`. The aligner used for the real nanopore experiments is available at `https://github.com/bluenote-1577/sce-aligner/` along with scripts for regenerating figures.

6 Competing interest statement

The authors declare no competing interests.

7 Acknowledgements

J.S. was supported by an NSERC CGS-D scholarship. This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2022-03074.

Bibliography

Abouelhoda MI and Ohlebusch E. 2005. Chaining algorithms for multiple genome comparison. *Journal of Discrete Algorithms* **3**: 321–341.

Alon N and Spencer JH. 2015. *The Probabilistic Method*. John Wiley & Sons.

Alser M, Rotman J, Deshpande D, Taraszka K, Shi H, Baykal PI, Yang HT, Xue V, Knyazev S, Singer BD, et al.. 2021. Technology dictates algorithms: recent developments in read alignment. *Genome Biology* **22**: 249.

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Backurs A and Indyk P. 2018. Edit Distance Cannot Be Computed in Strongly Subquadratic Time (Unless SETH is False). *SIAM Journal on Computing* **47**: 1087–1097.

Batzoglou S. 2005. The many faces of sequence alignment. *Briefings in Bioinformatics* **6**: 6–22.

Berger B, Daniels NM, and Yu YW. 2016. Computational Biology in the 21st Century: Scaling with Compressive Algorithms. *Communications of the ACM* **59**: 72–80.

Berger B, Waterman MS, and Yu YW. 2021. Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE transactions on information theory* **67**: 3287–3294.

Blanca A, Harris RS, Koslicki D, and Medvedev P. 2022. The Statistics of k-mers from a Sequence Undergoing a Simple Mutation Process Without Spurious Matches. *Journal of Computational Biology* **29**: 155–168.

Britten RJ. 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences* **99**: 13633–13635.

Chaisson MJ and Tesler G. 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**: 238.

Chao KM, Pearson WR, and Miller W. 1992. Aligning two sequences within a specified diagonal band. *Bioinformatics* **8**: 481–487.

Chvátal V and Sankoff D. 1975. Longest Common Subsequences of Two Random Sequences. *Journal of Applied Probability* **12**: 306–315.

Colquhoun RM, Hall MB, Lima L, Roberts LW, Malone KM, Hunt M, Letcher B, Hawkey J, George S, Pankhurst L, et al.. 2021. Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. *Genome Biology* **22**: 267.

Delahaye C and Nicolas J. 2021. Sequencing dna with nanopores: Troubles and biases. *PLoS One* **16**: e0257521.

Durbin R, Eddy SR, Krogh A, and Mitchison G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Dutta A, Pellow D, and Shamir R. 2022. Parameterized syncmer schemes improve long-read mapping. *PLOS Computational Biology* **18**: e1010638.

Edgar R. 2021. Syncmers are more sensitive than minimizers for selecting conserved k-mers in biological sequences. *PeerJ* **9**: e10805.

Farrar M. 2007. Striped Smith–Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* **23**: 156–161.

Fofanov Y, Luo Y, Katili C, Wang J, Belosludtsev Y, Powdrill T, Belapurkar C, Fofanov V, Li TB, Chumakov S, et al.. 2004. How independent are the appearances of n-mers in different genomes? *Bioinformatics* **20**: 2421–2428.

Frith MC, Noé L, and Kucherov G. 2020. Minimally-overlapping words for sequence similarity search. *Bioinformatics (Oxford, England)* .

Frith MC, Shaw J, and Spouge JL. 2022. How to optimally sample a sequence for rapid analysis. *bioRxiv* .

Ganesh A and Sy A. 2020. Near-Linear Time Edit Distance for Indel Channels. *arXiv:2007.03040* .

Hera MR, Pierce-Ward NT, and Koslicki D. 2022. Debiasing fracminhash and deriving confidence intervals for mutation rates across a wide range of evolutionary distances. *bioRxiv* .

Irber LC, Brooks PT, Reiter TE, Pierce-Ward NT, Hera MR, Koslicki D, and Brown CT. 2022. Lightweight compositional analysis of metagenomes with fracminhash and minimum metagenome covers. *bioRxiv* .

Ivanov P, Bichsel B, and Vechev M. 2022. Fast and optimal sequence-to-graph alignment guided by seeds. In *International Conference on Research in Computational Molecular Biology*, pp. 306–325. Springer.

Jain C, Gibney D, and Thankachan SV. 2022. Co-linear chaining with overlaps and gap costs. In *International Conference on Research in Computational Molecular Biology*, pp. 246–262. Springer.

Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, and Phillippy AM. 2020. Weighted minimizer sampling improves long read mapping. *Bioinformatics* **36**: i111–i118.

Janson S. 2004. Large deviations for sums of partly dependent random variables: Large Deviations for Dependent Random Variables. *Random Structures & Algorithms* **24**: 234–248.

Kalikar S, Jain C, Vasimuddin M, and Misra S. 2022. Accelerating minimap2 for long-read sequencing applications on modern CPUs. *Nature Computational Science* **2**: 78–83.

Keich U, Li M, Ma B, and Tromp J. 2004. On spaced seeds for similarity search. *Discrete Applied Mathematics* **138**: 253–263.

Kiełbasa SM, Wan R, Sato K, Horton P, and Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Research* **21**: 487–493.

Kiwi M, Loebl M, and Matoušek J. 2005. Expected length of the longest common subsequence for large alphabets. *Advances in Mathematics* **197**: 480–498.

Koerkamp RG and Ivanov P. 2022. Exact global alignment using a* with seed heuristic and match pruning. *bioRxiv* .

Koning APJd, Gu W, Castoe TA, Batzer MA, and Pollock DD. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLOS Genetics* **7**: e1002384.

Koonin EV, Aravind L, and Kondrashov AS. 2000. The Impact of Comparative Genomics on Our Understanding of Evolution. *Cell* **101**: 573–576.

Köster J. 2016. Rust-Bio: a fast and safe bioinformatics library. *Bioinformatics* **32**: 444–446.

Langmead B and Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Lember J and Matzinger H. 2009. Standard deviation of the longest common subsequence. *The Annals of Probability* **37**.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**: 4572–4574.

Li H, Feng X, and Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. *Genome Biology* **21**: 265.

Lipman DJ and Pearson WR. 1985. Rapid and Sensitive Protein Similarity Searches. *Science* **227**: 1435–1441.

Lunter G, Drummond AJ, Miklós I, and Hein J. 2005. Statistical Alignment: Recent Progress, New Applications, and Challenges. In *Statistical Methods in Molecular Evolution*, pp. 375–405. Springer-Verlag, New York.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, and Hein J. 2008. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Research* **18**: 298–309.

Mäkinen V and Sahlin K. 2020. Chaining with overlaps revisited. *arXiv* .

Marco-Sola S, Moure JC, Moreto M, and Espinosa A. 2021. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**: 456–463.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, and Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology* **14**: e1005944.

Marçais G, Solomon B, Patro R, and Kingsford C. 2019. Sketching and Sublinear Data Structures in Genomics. *Annual Review of Biomedical Data Science* **2**: 93–118.

Medvedev P. 2022a. The limitations of the theoretical analysis of applied algorithms. *arXiv* .

Medvedev P. 2022b. Theoretical analysis of edit distance algorithms: an applied perspective. *arXiv* .

Myers EW. 1986. AnO(ND) difference algorithm and its variations. *Algorithmica* **1**: 251–266.

Myers G and Miller W. 1995. Chaining multiple-alignment fragments in sub-quadratic time. In *Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms*, SODA '95, pp. 38–47. Society for Industrial and Applied Mathematics, USA.

Mäkinen V, Belazzougui D, Cunial F, and Tomescu AI. 2015. *Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing*. Cambridge University Press, Cambridge.

Navarro G. 2001. A guided tour to approximate string matching. *ACM Computing Surveys* **33**: 31–88.

Needleman SB and Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**: 443–453.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al.. 2022. The complete sequence of a human genome. *Science* **376**: 44–53.

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, and Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**: 132.

Otto C, Hoffmann S, Gorodkin J, and Stadler PF. 2011. Fast local fragment chaining using sum-of-pair gap costs. *Algorithms for Molecular Biology* **6**: 4.

Rautiainen M and Marschall T. 2020. Graphaligner: rapid and versatile sequence-to-graph alignment. *Genome biology* **21**: 1–28.

Reinert G, Schbath S, and Waterman MS. 2000. Probabilistic and Statistical Properties of Words: An Overview. *Journal of Computational Biology* **7**: 1–46.

Ren J and Chaisson MJP. 2021. lra: A long read aligner for sequences and contigs. *PLOS Computational Biology* **17**: e1009078.

Roberts M, Hayes W, Hunt BR, Mount SM, and Yorke JA. 2004. Reducing storage requirements for biological sequence comparison. *Bioinformatics* **20**: 3363–3369.

Sahlin K. 2022. Strobealign: flexible seed size enables ultra-fast and accurate read alignment. *Genome Biology* **23**: 1–27.

Sahlin K, Baudeau T, Cazaux B, and Marchet C. 2022. A survey of mapping algorithms in the long-reads era. *bioRxiv* .

Sarmashghi S, Bohmann K, P Gilbert MT, Bafna V, and Mirarab S. 2019. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology* **20**: 34.

Shaw J and Yu YW. 2022. Theory of local k-mer selection with applications to long-read alignment. *Bioinformatics* **38**: 4659–4669.

Siegel AF. 1982. Robust regression using repeated medians. *Biometrika* **69**: 242–244.

Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang PC, Carroll A, et al.. 2021. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* .

Smith T and Waterman M. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**: 195–197.

Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, and Nagarajan N. 2016. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature Communications* **7**: 11307.

Suzuki H and Kasahara M. 2018. Introducing difference recurrence relations for faster semi-global alignment of long sequences. *BMC Bioinformatics* **19**: 45.

Szpankowski W. 2001. *Average Case Analysis of Algorithms on Sequences: Szpankowski/Average.* John Wiley & Sons, Inc., Hoboken, NJ, USA.

Thorne JL, Kishino H, and Felsenstein J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution* **33**: 114–124.

Ukkonen E. 1983. On approximate string matching. In *Foundations of Computation Theory* (ed. M Karpinski), Lecture Notes in Computer Science, pp. 487–495. Springer, Berlin, Heidelberg.

Vingron M and Waterman MS. 1994. Sequence alignment and penalty choice: Review of concepts, case studies and implications. *Journal of Molecular Biology* **235**: 1–12.

Zheng H, Kingsford C, and Marçais G. 2020. Improved design and analysis of practical minimizers. *Bioinformatics* **36**: i119–i127.

Zheng H, Kingsford C, and Marçais G. 2021. Sequence-specific minimizers via polar sets. *Bioinformatics* **37**: i187–i195.

A Mathematical conventions and reminders

We list a few reminders and conventions that will be useful for the technical parts of our work. $\gg$ will mean asymptotically dominates, i.e. $f(n) \gg g(n)$ if and only if $\lim_{n\to\infty} g(n)/f(n) = 0$. $\log(n) = \log_\sigma(n)$ and $\ln(n)$ uses the e as the base, where σ is the alphabet size. $|S| \sim n$ and $|S'| \sim m$, where S, S' are our pairs of strings and $\sim$ ignores small factors of k present. This is because $k = C \log n$ for a constant $C > 0$ and $m = \Omega(n^{2C\alpha+\epsilon})$ for the assumptions of our main theorems, and these terms dominate $k = O(\log n)$. S' is a mutated substring of S, where the substring starts at position p. $(1-\theta)^k = n^{-C\alpha}$ where $\alpha = -\log_\sigma(1-\theta) > 0$ is a function of $0 < \theta < 1$, the mutation rate, and σ is our alphabet size which is > 1. Our results are general, but we will use $\sigma = 4$ for specific numerical results. When we use the assumption $C > \frac{2}{1-2\alpha}$, C is also greater than 2, a useful fact we will use repeatedly.

A chain $\mathcal{C}$ is a sequence of anchors (exact matches of k-mers), represented by tuples $((i_1, j_1), \ldots, (i_u, j_u))$ where i_ℓ is the starting position of the ℓth k-mer on S and j_ℓ is the starting position of the ℓth k-mer on S'. Anchors can overlap. The cost for a chain $\mathcal{C}$ is $u - \zeta[(i_u - i_1) + (j_u - j_1)]$ where $\zeta = \zeta(n) > 0$ will eventually be considered as a (decreasing) function of n.

B Missing proofs and definitions from "Homology and recoverability"

We note that technically, the alignment matrix in Definition 1 is slightly different than the standard dynamic programming matrix (Durbin et al., 1998) which can be thought of a directed graph with a path representing an alignment. Our representation does not include the graphical information; it only captures information pertaining to the possibility of matching bases, which is sufficient for us.

B.1 Recoverability and definition of $Align(\mathcal{C})$

Figure 5b will serve as a helpful guide for our definitions. Let $\mathcal{C}$ be a chain of anchors $((i_1, j_1), \ldots, (i_u, j_u))$. Below we define carefully define recoverability and set up our proof of Lemma 1.

A chain gives a set of k-mer matches and a set of possible alignments by extending through bases, constraining the full alignment matrix between S and S'. We can formalize this as follows. Given two consecutive anchors (i_ℓ, j_ℓ) and $(i_{\ell+1}, j_{\ell+1})$, extending allows for possible matches between the gaps given by the following set of possible matches

$$Ext(\ell) = [i_\ell + k..i_{\ell+1} - 1] \times [j_\ell + k..j_{\ell+1} - 1].$$

The factor of $+k$ and -1 is because i_ℓ is the start of the ℓth anchor, and extension starts after the end of the first k-mer and goes until 1 base before the start of the second k-mer. $Ext(\ell)$ corresponds to the green boxes in Figure 5b. The set $Ext(\ell)$ is empty if $i_\ell + k > i_{\ell+1} - 1$ and similarly for j, i.e. there are no gaps between possibly overlapping anchors. $Align(\mathcal{C})$ also takes into account the k-mer matches given by $\mathcal{C}$. Thus we define $Align(\mathcal{C})$ formally as

$$Align(\mathcal{C}) = \left(\bigcup_{\ell=1}^{u} \{(i_\ell, j_\ell), \ldots, (i_\ell + k - 1, j_\ell + k - 1)\} \right) \cup \left(\bigcup_{\ell=1}^{u-1} Ext(\ell) \right)$$

where the first term corresponds to the k-mer matches in the alignment matrix.

B.2 Proof of Lemma 1

Since we care about the intersection of $Align(\mathcal{C})$ with the homologous diagonal, we can manipulate the expression for $Align(\mathcal{C})$ to get something more tractable. Let $\text{Diag}[a..b] = \{(x,x) : x \in [a..b]\}$, $\mathcal{C}_H$ be the set of homologous anchors in $\mathcal{C}$, and $\mathcal{C}_S$ be the set of spurious anchors. The homologous anchors give rise to the homologous matches in the alignment matrix

$$M_H = \bigcup_{(i_\ell, i_\ell) \in \mathcal{C}_H} \text{Diag}[i_\ell..i_\ell + k - 1].$$

The spurious anchors give no matches on the homologous diagonal, so it contributes nothing to the $Align(\mathcal{C}) \cap D_H$ term. Therefore,

$$Align(\mathcal{C}) \cap D_H = M_H \cup \big(\bigcup_{\ell=1}^{u-1} Ext(\ell) \cap D_H \big).$$

We define

$$NR = \bigcup_{(i_\ell, j_\ell) \in \mathcal{C}_S} \text{Diag}[\min(i_\ell, j_\ell).. \max(i_\ell, j_\ell) + k - 1] \cap D_H.$$

NR represents the parts of the diagonal in Figure 5b that are not recoverable or accessible by extension through anchors (although it may technically intersect $Align(\mathcal{C})$ in our mathematical definition). The following identity then holds after rewriting M_H:

$$M_H \cup NR = \bigcup_{\ell=1}^{u} \text{Diag}[\min(i_\ell, j_\ell).. \max(i_\ell, j_\ell) + k - 1] \cap D_H$$

because the min-max condition is redundant for homologous anchors where $i_\ell = j_\ell$. We can fill in the gaps between $M_H \cup NR$ along the diagonal with $Ext(\ell)$ after noticing that we can rewrite

$$Ext(\ell) \cap D_H = \text{Diag}[\max(i_\ell, j_\ell) + k.. \min(i_{\ell+1}, j_{\ell+1}) - 1]$$

using the fact that $(x,x) \in Ext(\ell) \iff i_\ell + k \leq x \leq i_{\ell+1} - 1$ and $j_\ell + k \leq x \leq j_{\ell+1} - 1$. Finally,

$$(Align(\mathcal{C}) \cap D_H) \cup NR = \Bigg(M_H \cup N_R \cup \bigcup_{\ell=1}^{u-1} Ext(\ell) \cap D_H \Bigg)$$

$$= \Bigg(\bigcup_{\ell=1}^{u} \text{Diag}[\min(i_\ell, j_\ell).. \max(i_\ell, j_\ell) + k - 1] \cap D_H] \Bigg) \cup \Bigg(\bigcup_{\ell=1}^{u-1} \text{Diag}[\max(i_\ell, j_\ell) + k.. \min(i_{\ell+1}, j_{\ell+1}) - 1] \Bigg)$$

$$= \text{Diag}[\min(i_1, j_1).. \max(i_u, j_u) + k - 1] \cap D_H.$$

This result follows because the union over all sets covers the entire diagonal between the first and last anchor. After bounding the diagonal to lie within D_H and removing NR from both sides of the equation, we obtain the following result.

Supplementary Lemma 1.

$$|Align(\mathcal{C}) \cap D_H| \geq k + \min\big(p + m, \max(i_u, j_u)\big) - \max\big(p + 1, \min(i_1, j_1)\big) - |NR|.$$

Proof. The following inequality holds by simple set theoretic arguments:

$$|(Align(\mathcal{C}) \cap D_H)| \geq |(Align(\mathcal{C}) \cap D_H) \cup NR| - |NR|.$$

Now notice that the term $\text{Diag}[\min(i_1, j_1).. \max(i_u, j_u) + k - 1]$ may technically lie outside the diagonal $[p+1..m] \times [p+1..m]$, so to constrain it to lie on the diagonal after intersecting with D_H, we must make $\max(i_u, j_u) = p + m$ if $i_u > p + m$ and $\min(i_1, j_1) = p + 1$ if $i_1 < p + 1$. Thus the cardinality of the set $\text{Diag}[\min(i_1, j_1).. \max(i_u, j_u) + k - 1] \cap D_H$ is $k + \min(p + m, \max(i_u, j_u)) - \max(p + 1, \min(i_1, j_1))$, which finishes the proof.

Lemma 1 *Given any chain $\mathcal{C} = ((i_1, j_1), \ldots, (i_u, j_u))$, we have that $(j_u - j_1 - L(\mathcal{C}))/|S'| \leq R(\mathcal{C})$.*

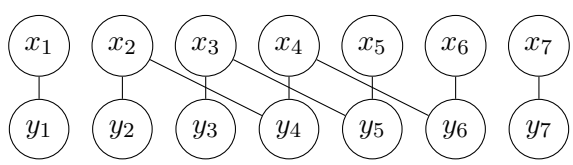
Proof. Using the supplementary lemma above, we note that $j_u \leq p + m$ and $j_1 \geq p + 1$ since the anchors must lie between $[p + 1..p + m]$, so after ignoring the k term we get

$$|S'| \cdot R(\mathcal{C}) = |(Align(\mathcal{C}) \cap D_H)| \geq j_u - j_1 - |NR|.$$

We now prove that the breaks "cover" NR, so $|NR| \leq L$ where L is the total length of the breaks, proving the result. Let $\pi(NR)$ be the projection of the set onto one of the coordinate axes (it doesn't matter which one). Let $x \in \pi(NR)$. Then for some spurious anchor (i, j), $\min(i, j) \leq x \leq \max(i, j) + k - 1$. But every spurious anchor is contained in exactly one break, since breaks partition the set of spurious anchors. This break $B = [a..b + k - 1] \supset [\min(i, j).. \max(i, j) + k - 1]$ since the break takes the minimum and maximum coordinates over *all* spurious anchors in the break. Thus $x \in B$ for some break B and the set of breaks $\bigcup_{\text{breaks } B \text{ in } \mathcal{C}} B \supset \pi(NR)$. $L(\mathcal{C}) \geq |\pi(NR)| = |NR|$, and we're done.

C Missing proofs from "Fundamental tools and bounds" and "Bounding sums of k-mer random variables"

C.1 Proof of Theorem 3



Supplementary Figure 1: A match graph $G(\mathcal{M})$ where $\mathcal{M} = \{M(2, 4), M(3, 5), M(4, 6)\}$ and $|S'| = |S|$.

Intuitively, a match graph encodes dependencies between random letters after conditioning on $\mathcal{M} = \{M(i_1, j_1), M(i_2, j_2), \ldots\}$. The main idea is that for such a graph, the connected components should be mutually independent of everything not in the component because all of the dependencies lie only in that connected component. The match graph is an example of a *dependency graph*; see Chapter 5 in (Alon and Spencer, 2015).

Supplementary Lemma 2. *Consider a set of random variables of the form $\{M(i_1, j_1), \ldots\} = \mathcal{M}$. If two vertices x, y are in separate connected components in $G(\mathcal{M})$ then they are conditionally independent of $\mathcal{M}$.*

Proof. Consider the connected components $G_1, \ldots, G_q$ of $G(\mathcal{M})$ and the edges within the connected components corresponding to the random variables in $\mathcal{M}$; call this partition of edges $\{E_1, \ldots, E_q\}$ where $\bigcup_{i=1}^{q} E_i = \mathcal{M}$.

First, we claim that $G_1, \ldots, G_q$, considered as random letters in S and S', are mutually independent (without considering additional $\mathcal{M}$ variables). Indeed, each G_i can be considered as just a set of pairs of (x_ℓ, y_ℓ) random variables because the edges (x_ℓ, y_ℓ) always exist in the original graph and so the connected component must leave no x_i or y_i unpaired. All pairs of (x_ℓ, y_ℓ) random variables are mutually independent by definition of our original mutation model. Thus all G_is are mutually independent.

Now we consider the E_ℓs. The random variables in E_ℓ are functions of G_ℓ, the letters in the connected component. It follows that the E_ℓs are functions of mutually independent G_ℓs, and are themselves mutually independent. Now let $x \in G_x$ and $y \in G_y$ be two letters in different connected components. x, E_x are both functions of G_x and similarly for y, E_y, and G_y. Thus $\Pr(x, y, E_1, \ldots, E_q) = \Pr(x, E_x)\Pr(y, E_y)\prod_{\ell \neq x,y}\Pr(E_\ell)$ which shows conditional independence of x, y with respect to $\{E_1, \ldots, E_\ell\} = \mathcal{M}$.

Theorem 3 *The random variables $\mathcal{M} = \{M(i_1, j_1), M(i_2, j_2), \ldots\}$ where $i_\ell \neq j_\ell$ for all $M(i_\ell, j_\ell) \in \mathcal{M}$ are independent if the induced match graph has no cycles.*

Proof. We shorten $M(i_\ell, j_\ell) = M_\ell$. We will proceed by induction on the size of $\mathcal{M}$. The case $|\mathcal{M}| = 0$ is vacuously true. If $|\mathcal{M}| = q$, We want to show that $\Pr(M_1, \ldots, M_q) = \Pr(M_1, \ldots, M_{q-1})\Pr(M_q)$; the first term is $\frac{1}{\sigma^{q-1}}$ because $M_1, \ldots, M_{q-1}$ must also be cycle free and by the induction assumption. Let $\mathcal{M}^- = \{M_1, \ldots, M_{q-1}\}$. Let $M_q = M(\alpha, \beta)$. We need to calculate $\Pr(M_q = 1 \mid \mathcal{M}^-)$ and show that it is equal to $\frac{1}{\sigma}$.

By definition, we have that

$$\Pr\big(M_q = 1 \mid \mathcal{M}^-\big) = \sum_{a \in \Sigma} \Pr\big(x_\alpha = a, y_\beta = a \mid \mathcal{M}^-\big).$$

Since $G(\mathcal{M})$ has no cycles, x_α and y_β lie on two different connected components of $G(\mathcal{M}^-)$ after removing the edge induced by $M(\alpha, \beta)$ (we use the $\alpha \neq \beta$ assumption here). Therefore x_α, y_β are conditionally independent by the above Supplementary Lemma 2, so we can write the above sum as

$$= \sum_{a \in \Sigma} \Pr\big(x_\alpha = a \mid \mathcal{M}^-\big)\Pr\big(y_\beta = a \mid \mathcal{M}^-\big).$$

We claim that both terms in the sum are $\frac{1}{\sigma}$. To show this, let the permutation C_σ be a cyclic permutation of order σ on letters in Σ (e.g. C_4 sends $A \mapsto C \mapsto T \mapsto G$) and apply it to every letter on both strings when we write $C_\sigma(S, S')$. It's not hard to see that this is a measure (or probability) preserving transformation for $\Pr(\cdot \mid \mathcal{M}^-)$ because it preserves matches between S and S' and letters are distributed uniformly.

Let A be the set of strings (S, S') for which $x_\alpha = a$, and we can see that $\Pr(x_\alpha = a \mid \mathcal{M}^-) = \Pr(A \mid \mathcal{M}^-)$. Now applying C_σ σ times and using measure preservation gets that

$$\Pr\big(A \mid \mathcal{M}^-\big) = \Pr\big(C_\sigma(A) \mid \mathcal{M}^-\big) = \Pr\big(C_\sigma^2(A) \mid \mathcal{M}^-\big) = \ldots$$

Finally, all $C_\sigma^n(A), n = 1, \ldots, \sigma - 1$ are clearly disjoint and partitions the space of strings. Thus, we see that $\Pr(x_\alpha = a \mid \mathcal{M}^-) = \frac{1}{\sigma}$ as desired. The argument works exactly the same for $\Pr(y_\beta = a \mid \mathcal{M}^-)$, so

$$\Pr\big(M_q = 1 \mid \mathcal{M}^-\big) = \sum_{a \in \Sigma} \Pr\big(x_\alpha = a \mid \mathcal{M}^-\big)\Pr(y_\beta = a \mid \mathcal{M}) = \sigma \cdot \frac{1}{\sigma^2} = \frac{1}{\sigma} = \Pr(M_q = 1)$$

and our induction step is complete.

C.2 Proof of Corollary 1

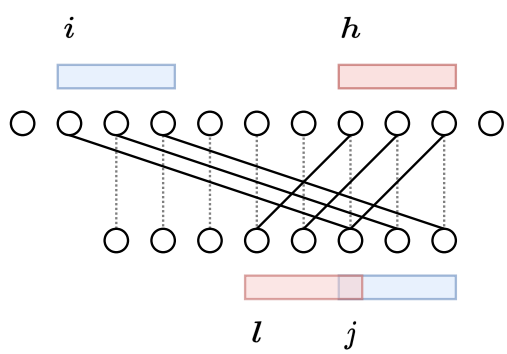
Corollary 1 *If $i \neq j$, $\Pr(A(i,j)=1) = \frac{1}{\sigma^k}$. Otherwise, $\Pr(A(i,i)=1) = (1-\theta)^k$.*

Proof. If the match graph induced by $A(i,j)$, i.e. induced by $M(i,j), M(i+1,j+1), \ldots, M(i+k-1, j+k-1)$ has no cycles, using Theorem 3 gives the result. This is easy to see from drawing out the match graph which has edges (x,y) for all x,y and $(x_i,y_j),\ldots,(x_{i+k-1},y_{j+k-1})$ but we give a rigorous argument below. See Supplementary Figure 1 for an example of such a match graph.

First, suppose $j > i$ and suppose $x_{i+\ell}$ is the vertex with a cycle and with the smallest possible $\ell \in [0..k-1]$. We can assume the cycle touches some x because the match graph is bipartite. Since $x_{i+\ell}$ has degree at most two and therefore equal to two, we can traverse the edge $(x_{i+\ell}, y_{i+\ell})$ as the first edge in the cycle. If $y_{i+\ell}$ has degree one, then $x_{i+\ell}$ has no cycle so suppose the degree is two. Then we have an edge $(y_{i+\ell}, x_{i+\alpha})$ which is in the cycle. However, this implies $x_{i+\alpha}$ also has a cycle and $\alpha < \ell$ because $j > i$; the edges from x_i to y_j have increasing positions. This creates a contradiction as ℓ was the *smallest* index with a cycle, so no cycles exist. Now if we have $j < i$, we repeat the same argument but with ℓ the largest in $[0..k-1]$ and the same contradiction arises.

The second result follows easily since clearly $M(i,i), M(i+1,i+1), \ldots$ are functions of sets of mutually independent random variables, and $\Pr(M(i,i)=1) = (1-\theta)$.

C.3 Proof of Lemma 2



Supplementary Figure 2: Match graph of the M variables induced by two anchors and pictorial representation of the setting of Lemma 2. The two anchors here generate no cycles in the match graph because $|i-h| \geq k$, and $|i-l| \geq k$, even though $|j-l| < k$ where $k=3$.

Lemma 2 *For $A(i,j)$ and $A(h,l)$, if both of the following conditions hold:*

1. *$|i-h| \geq k$ or $|j-l| \geq k$ and*
2. *$|i-l| \geq k$ or $|j-h| \geq k$,*

then the induced match graph on the M variables for $A(i,j)$ and $A(h,l)$ has no cycles.

The intuition for the conditions of Lemma 2 is that the first condition prevents both anchors from overlapping too much, while the second condition prevents the condition where, for example, two variables $M(1,5), M(5,1)$ can cause a cycle to form in the match graph by $x_1 \to y_5 \to x_5 \to y_5 \to x_1$.

Proof. If $|i-h| \geq k$ or $|j-l| \geq k$, then without loss of generality, assume $|i-h| \geq k$. Then we can find a set of $\{x_i, \ldots, x_{i+k-1}\} = X_i$ and $X_h = \{x_h, \ldots, x_{h+k-1}\} = X_h$ disjoint. See Supplementary Figure 2 for a pictorial representation. It will turn out that if $|j-l| \geq k$, then the xs become ys and the argument doesn't change.

Now we claim two facts.

1. The degree of every $x \in X_i$ or X_h must be at most two: each A variable induces one new edge for the xs covered by some k-mer, but these are disjoint sets of k-mers.
2. If a cycle exists, the cycle must touch some $x \in X_i$ and also a $x' \in X_h$: if it did not, a cycle must exist for some $x \in X_i$ or X_h (the graph is bipartite, so must include a x) and would not use induced edges on the match graph for one of $A(i,j)$ or $A(h,l)$. This implies the match graph for just one of $A(i,j)$ or $A(h,l)$ would have a cycle, which is impossible by the proof of Corollary 1.

To complete the proof, we proceed with an argument similar to the proof of Corollary 1. We need four different cases corresponding to either $|i-l| \geq k$ or $|h-j| \geq k$ and our previous assumption of $|i-h| \geq k$ or $|j-l| \geq k$. We prove only one case but it is easy to translate the argument to the other three.

Let's assume $|i-l| \geq k$. By the second fact above, if a cycle were present, we can assume it touches $x_\ell \in X_i$. If $j > i$, let x_ℓ be the leftmost $x_\ell \in X_i$ with a cycle, and if $j < i$, let x_ℓ be the rightmost. Because the degree of x_ℓ is at most two (by the first fact above) and thus must be exactly two, we can assume the cycle starts with the edge (x_ℓ, y_ℓ). Now, y_ℓ has degree at most two since $|i-l| \geq k$, i.e. the k-mer starting at l on S' doesn't overlap y_i, therefore the only other edges adjacent to y_ℓ from $x_\alpha \in X_i$. Therefore x_α is either to the left of x_ℓ if $j > i$ or to the right if $j < i$ and must also have a cycle. However, because x was the leftmost or rightmost such x, this is impossible, so we are done.

If $|j-l| \geq k$, we use ys instead of xs and we switch "rightmost" with "leftmost". If we have $|j-h| \geq k$, we switch X_i or Y_i with X_h or Y_h instead. Repeating the same argument verbatim gives the conclusion.

C.4 Proof of Lemma 3

Lemma 3 $\mathbb{E}[N_S^2] \leq 8k^2 mn \frac{1}{\sigma^k} + \mathbb{E}[N_S]^2$. *Thus the variance* $Var(N_S)$ *can be upper bounded by* $8k^2 \frac{mn}{\sigma^k}$. *Furthermore,* $\mathbb{E}[N_H^2] \leq 2mk(1-\theta)^k + m^2(1-\theta)^{2k}$ *and* $\mathbb{E}[N_H N_S] \leq 4k\frac{mn}{\sigma^k} + m^2 n(1-\theta)^k \frac{1}{\sigma^k}$.

Proof. We will frequently use $(m-1) < m$ and $(n-1) < n$ to simplify the bounds. With $N_S = \sum_{i \neq j} A(i,j)$, we get

$$N_S^2 = \sum_{h \neq l} \sum_{i \neq j} A(i,j)A(h,l).$$

Now we try to bound $\mathbb{E}[N_S^2]$, where we already know $\mathbb{E}[N_S] = m(n-1)\frac{1}{\sigma^k}$. Let $B_k(i,j) = \{(h,l) : |h-i| < k \text{ and } |l-j| < k\}$. We can separate the sum into three different parts,

$$S_1 = \sum_{(h,l) \notin B_k(i,j) \cup B_k(j,i)} \sum_{i \neq j} A(i,j)A(h,l)$$

$$S_2 = \sum_{(h,l) \in B_k(i,j)} \sum_{i \neq j} A(i,j)A(h,l)$$

$$S_3 = \sum_{(h,l) \in B_k(j,i) \setminus B_k(i,j)} \sum_{i \neq j} A(i,j)A(h,l)$$

and $N_S^2 = S_1 + S_2 + S_3$. Let us bound the expectation of each sum separately. Firstly, we get

$$\mathbb{E}[S_1] \leq m^2(n-1)^2 \frac{1}{\sigma^{2k}} = \mathbb{E}[N_S]^2$$

because $A(i,j)A(h,l)$ are independent when (h,l) is not in the set $B_k(i,j) \cup B(j,i)$ by Corollary 2, and there are at most $m^2(n-1)^2$ possible $A(h,l)A(i,j)$ tuples.

$$\mathbb{E}[S_2] \leq 4k^2 \frac{mn}{\sigma^k}$$

because $|B_k(i,j)| \leq 4k^2$ and $\mathbb{E}[A(i,j)A(h,l)] \leq \mathbb{E}[A(i,j)] = \frac{1}{\sigma^k}$. $\mathbb{E}[S_3] \leq 4k^2 \frac{mn}{\sigma^k}$ follows similarly. Thus,

$$Var(N_S) = \mathbb{E}[N_S^2] - \mathbb{E}[N_S]^2 \leq 8k^2 \frac{mn}{\sigma^k}.$$

The variance of N_H was calculated in (Blanca et al., 2022) for the variable $N_{mut} = n - N_H$, but we can use a simpler approximate bound. $\mathbb{E}[N_H^2] = \sum_{j=1}^n \sum_{i=1}^n \mathbb{E}[A(i,i)A(j,j)]$; notice that if $|i-j| < k$ then there is obviously dependence between $A(i,i)$ and $A(j,j)$ because the anchors have overlapping bases, but otherwise the variables are independent. Thus

$$\mathbb{E}[N_H^2] = \sum_{|j-i|<k} \sum_i \mathbb{E}[A(i,i)A(j,j)] + \sum_{|j-i|\geq k} \sum_i \mathbb{E}[A(i,i)]\mathbb{E}[A(j,j)] \leq 2km(1-\theta)^k + m^2(1-\theta)^{2k}$$

where we use the trivial bound $\mathbb{E}[A(i,i)A(j,j)] \leq \mathbb{E}[A(i,i)]$ in the first bound and independence in the second.

For the bound $\mathbb{E}[N_H N_S]$, we can write this as

$$\mathbb{E}[N_H N_S] = \sum_i \sum_{h \neq l} A(i,i)A(h,l).$$

The only indices where dependence may be an issue is when either $|h-i| < k$ or $|l-i| < k$. Thus for each pair h, l, there are at most $4k$ choices for i which may not be independent. We can use the same idea as the previous bound to show that

$$\mathbb{E}[N_H N_S] \leq 4kmn \frac{1}{\sigma^k} + m^2 n (1-\theta)^k \frac{1}{\sigma^k}.$$

D Missing proofs from "Proof of non-sketched main result"

D.1 Proof of Theorem 6

Theorem 6 *Assume $m = \Omega(n^{2C\alpha+\epsilon})$ for some $\epsilon > 0$, and $C > \frac{1}{1-\alpha}$. Letting N be the total number of k-mer anchors, $\mathbb{E}[N \log N] = O(\mathbb{E}[N] \log \mathbb{E}[N]) = O(mn^{-C\alpha} \log m)$. It follows that the runtime of chaining is $O(mn^{-C\alpha} \log m)$.*

Proof. For all $x > 0$, we have the inequality $\ln(x/u) \leq x/u - 1$ for any $u > 0$. Substituting this into $N \ln N$ we have

$$N \ln N = N \ln(N/u) + N \ln(u) \leq \frac{N^2}{u} - N + N \ln(u).$$

Now let $u = \mathbb{E}[N]$ and we get $\mathbb{E}[N \ln N] \leq \frac{\mathbb{E}[N^2]}{\mathbb{E}[N]} - \mathbb{E}[N] + \mathbb{E}[N] \ln \mathbb{E}[N]$. Thus if we are able to show $\mathbb{E}[N^2]/\mathbb{E}[N] = O(\mathbb{E}[N] \log \mathbb{E}[N]) = O(mn^{-C\alpha} \log m)$ then we are done. Since $N^2 = (N_S + N_H)^2 =$

$N_S^2 + 2N_S N_H + N_H^2$, we can just use the moment bounds in Lemma 3. Plugging in the bounds yields

$$\mathbb{E}[N^2] \leq (8k^2mn^{1-C} + m^2n^{2-2C}) + (2kmn^{-C\alpha} + m^2n^{-2C\alpha}) + (8kmn^{1-C} + 2m^2n^{1-C\alpha-C})$$

To simplify these terms, notice that $n^{-C\alpha} \gg n^{1-C}$ and $mn^{-C\alpha} \gg k^2$ where $\gg$ means asymptotically dominates. This is because $-C\alpha > 1 - C$ follows from our assumption $C > \frac{1}{1-\alpha}$. The second term is because $m = \Omega(n^{2C\alpha+\epsilon})$ so $mn^{-C\alpha} \gg n^{\epsilon} \gg k^2 = O(\log^2(n))$. This gives

$$\mathbb{E}[N^2] = O(m^2n^{-2C\alpha}).$$

Remembering that $\mathbb{E}[N] = \Theta(m(1-\theta)^k) = \Theta(mn^{-C\alpha})$, we get that $\mathbb{E}[N^2]/\mathbb{E}[N] = O(mn^{-C\alpha})$ as desired, so we are done.

D.2 Proof of Lemma 6

Supplementary Lemma 3. *For any interval consisting of ℓ k-mers on S', the probability that all ℓ homologous anchors are 0 is upper bounded by*

$$\exp\left(-\frac{8\ell(1-\theta)^k}{25k}\right).$$

Proof. The letters on an interval on S, S' are distributed identically as a length $\ell + k - 1$ version of S, S'. We can then use Lemma 5 for $t = m(1-\theta)^k$.

Lemma 6 (F2) *With probability $\geq 1 - 1/n$, no homologous gap has size greater than*

$$g(n) = \frac{50k}{8(1-\theta)^k}\ln(n) = \frac{C \cdot 50}{8}\log(n)\ln(n) \cdot n^{C\alpha}$$

plus a small $C \log n$ term we will ignore because it is small asymptotically.

Proof. Using the above supplementary lemma, let $\ell = (50/8)\frac{k}{(1-\theta)^k}\ln(n)$. Then the probability that there are no homologous anchors in a segment of ℓ k-mers is $\leq \frac{1}{n^2}$.

Let $HG_1, \ldots, HG_{m-\ell+1}$ be indicator random variables where $HG_i = 1$ if the next ℓ k-mers from position i have no homologous anchors and 0 otherwise. It follows that $\mathbb{E}[\sum_{i=1}^{m-\ell+1} HG_i] \leq \frac{m}{n^2} \leq 1/n$. Using Markov's inequality, we see that $\Pr(\sum HG_i \geq 1) \leq \frac{1}{n}$. Thus with probability $\geq 1 - 1/n$, no homologous gap is larger than $50/8 \cdot \frac{k}{(1-\theta)^k}\ln(n)$ k-mers as desired. ℓ k-mers corresponds to a homologous gap of size $\ell - k + 1$, so need to add $k - 1$ to get an upper bound on the homologous gap size, but we will ignore this in the analysis because it is small asymptotically.

D.3 Proof of Lemma 7

Lemma 7 (F1 + F2) *Take any $C > \min(3, \frac{2}{1-2\alpha})$ and let $\zeta = \frac{1}{6g(n)}$ where $g(n) = C\frac{50}{8}\log(n)\ln(n)n^{C\alpha}$. Assume $m = \Omega(n^{2C\alpha+\epsilon})$ for some $\epsilon > 0$. Then for large enough n, there are no breaks of length $\geq m^{1/2}$ with probability greater than $1 - 2/n$ in an optimal chain.*

If we assume $C > 3$, then Lemma 4 shows that with probability $\geq 1 - 1/n$ and large enough n, no spurious anchors exist at all. Of course, no breaks can occur so we are already done in this case. The rest of the section is for tackling the case $C \leq 3$. We will prove a series of supplementary lemmas, and then prove Lemma 7.

We will now assume the hypotheses of Lemma 7 for the rest of this section. That is, $3 \geq C > \frac{2}{1-2\alpha}$, $m = \Omega(n^{2C\alpha+\epsilon})$ for some $\epsilon > 0$, and define $\zeta = \frac{1}{6g(n)}$ as in the statement of Lemma 7. We will not be too careful with small additive constants of order $O(\log n)$ due to indexing offsets from now on. For example, a length $m^{1/2}$ interval of bases contains technically $m^{1/2} - k + 1$ k-mers, but since we work with asymptotics we'll treat this as $\sim m^{1/2}$.

Supplementary Lemma 4 (F1 + F2). *With probability $\geq 1 - 2/n$, given an optimal chain $((i_1, j_1), \ldots, (i_u, j_u))$ at least one of the anchors is a homologous anchor for large enough n.*

Proof. We will show that no chain of only spurious anchors can be optimal, implying that any optimal chain must have at least one homologous anchor. By F2, there are no homologous gaps of length $\geq g(n)$ where $g(n) = \Theta(n^{C\alpha} \log^2(n))$, so we can lower bound the number of homologous anchors N_H under the space F2 by

$$N_H \geq \frac{m}{g(n)} = \Omega\left(\frac{m}{n^{C\alpha} \log^2(n)}\right).$$

Let $score(N_H)$ be the score of the chain with only homologous anchors. Then

$$score(N_H) \geq N_H - 2\zeta m$$

because the only homologous chain has a linear cost of at most $2\zeta m$. Now consider a chain without homologous anchors, i.e. only spurious anchors. Such a chain has maximum score

$$score(N_S) \leq N_S - 2N_S\zeta$$

because there are at most N_S such anchors, and the N_S in linear cost is assuming the smallest possible linear cost where there are no gaps between the anchors, i.e. $(i_u - i_1) + (j_u - j_1) \geq 2N_S$ if $u = N_S$. Using condition F1, $N_S \leq n^{2-C} + \sqrt{8}C \log(n) m^{1/2} n^{1-C/2}$. Since $\zeta = \frac{1}{6g(n)}$,

$$score(N_H) \geq N_H - 2\zeta m = \Omega\left(\frac{mn^{-C\alpha}}{\log^2(n)}\right)$$

holds and also

$$score(N_S) \leq N_S - 2N_S\zeta = O(m^{1/2} n^{1-C/2} \log(n)[1 - \zeta])$$

holds. We therefore we want the following asymptotic inequality to hold:

$$\frac{mn^{-C\alpha}}{\log^2(n)} \gg m^{1/2} n^{1-C/2} \log(n)[1 - \zeta].$$

This holds when $1 - C/2 + C\alpha < 0$ which is equivalent to our assumed condition $C > \frac{2}{1-2\alpha}$, and thus the chain of only homologous anchors has asymptotically greater score than any chain with only spurious anchors. Therefore any optimal chain with high probability must contain at least one homologous anchor for large enough n.

Supplementary Lemma 5 (F1 + F2). *Suppose a break is flanked by two homologous anchors in an optimal chain. That is, homologous anchors exist on both sides of the break in the chain. Then with probability* $\geq 1 - 2/n$ *and large enough* n*, this break has size* $< m^{1/2}$*.*

Proof. Suppose a break of length $\geq m^{1/2}$ is flanked by two homologous anchors. Then for the given chain $((i_1, j_1), \ldots, (i_u, j_u))$, the break occurs somewhere in the middle, say $((i_r, j_r), \ldots, (i_t, j_t))$ where $r > 1, t < u$. Let us construct a new chain by removing all spurious anchors $((i_r, j_r), \ldots, (i_t, j_t))$ within our chain that lie in this break, and then adding all homologous anchors that are present within the break to the chain. See Supplementary Figure 3.

This does not change the gap cost part of the chaining score which is still $\zeta[(i_u - i_1) + (j_u - j_1)]$, so if there are more homologous anchors after the switch, then this switch is more optimal. Assuming condition F2, the number of homologous anchors contained within a break is lower bounded by $m^{1/2}/g(n)$. Assuming the condition F1, the number of spurious anchors, and therefore the size of the break, is upper bounded by $N_S = \Theta(\log(n) m^{1/2} n^{1-C/2})$. Notice that

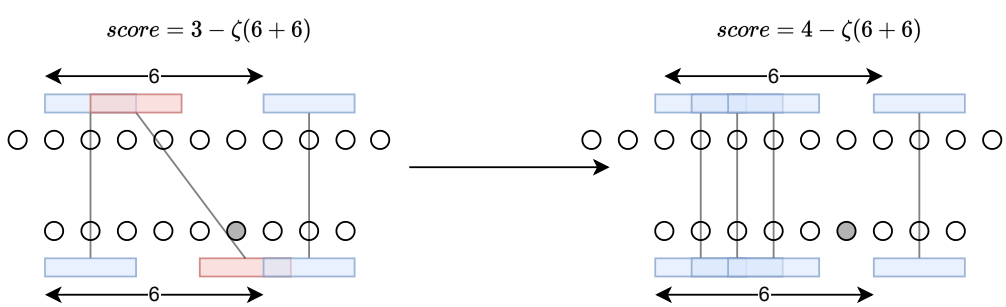
$$\frac{m^{1/2}}{g(n)} = \frac{m^{1/2}}{\Theta(\log^2(n) n^{C\alpha})} \gg \Theta(\log(n) m^{1/2} n^{1-C/2}) = N_S$$

again exactly when $1 - C/2 + C\alpha < 0$ which is equivalent to $C > \frac{2}{1-2\alpha}$. Thus under the event space F1 $\cap$ F2, switching to the homologous k-mers always improves our optimal chain for large enough n.

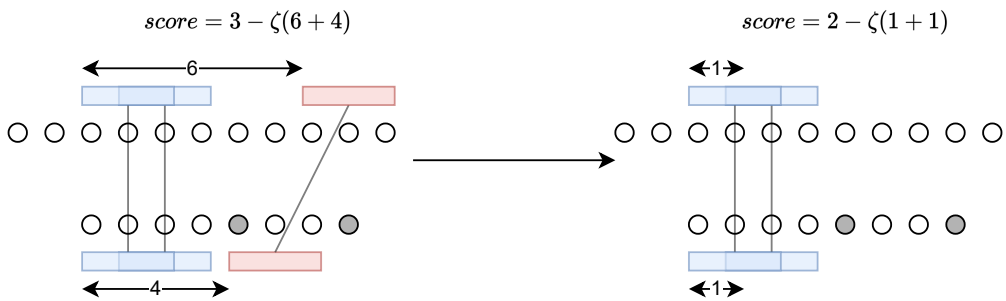
Supplementary Lemma 6 (F1 + F2). *Suppose a break is flanked by a single homologous anchor in an optimal chain. Then with probability* $\geq 1 - 2/n$ *and large enough* n*, this break has size* $< m^{1/2}$*.*

Proof. It follows easily from the definition of a break that a break flanked by one homologous anchor must occur at the start or end of the chain $((i_1, j_1), \ldots, (i_u, j_u))$. Let us assume that such a break occurs at the end of the chain at anchors $((i_r, j_r), \ldots, (i_u, j_u))$; the same argument works for the break occurring at the beginning of the chain.

We use a similar argument as above, but this time we must take into account the linear gap cost. Furthermore, instead of switching to the relative homologous anchors, we just remove the break from the chain and check if that improves the score. See Supplementary Figure 4.



Supplementary Figure 3: Circles are bases, grey circles are mutated bases, and boxes are k-mers. The red spurious anchor indicates a break. We can always remove a break flanked by two homologous anchors and then add in homologous anchors that may be present within the break. This idea is used in the proof of Supplementary Lemma 5 to show that for large enough breaks, such a procedure always improves the chaining score.



Supplementary Figure 4: Removing a break (the red anchor of k-mers) at the end may increase the chaining score if the break is long enough. We use this idea in the proof of Supplementary Lemma 6 to show that large breaks can not appear near the ends of chains.

The score of the old chain is $A-\zeta[(i_u-i_{r-1})+(j_u-j_{r-1})]+w$ where A is the score considering subchain of the anchors prior to r, and w is the number of anchors in the break. This is upper bounded by $A-\zeta(B-k)+N_S$ where $N_S=\Theta(\log(n)m^{1/2}n^{1-C/2})$ as before and B is the length of the break. This holds because the length of the break subtracted by k is $B-k=\max(i_u,j_u)-\min(i_r,j_r)$ which is less than $[(i_u-i_{r-1})+(j_u-j_{r-1})]$. Removing the break entirely is more effective when $\zeta B-k>N_S$, so when

$$(B-k)>\frac{N_S}{\zeta}=6N_Sg(n)=O(\log(n)m^{1/2}n^{1-C/2}\cdot n^{C\alpha}\log^2(n))$$

we can get a better chain by simply removing the break. By the previous discussion, $1-C/2+C\alpha<0$, so we just let $B=m^{1/2}$ and this completes the proof.

Proof (Lemma 7). Breaks are flanked by either one, two, or no homologous anchors. The last case can only occur if the entire chain is a break. By Supplementary Lemma 4, we always have at least one homologous anchor in any optimal chain (under F1 $\cap$ F2), so any break is flanked by at least one homologous anchor. Therefore by the previous supplementary lemmas, no breaks of size $m^{1/2}$ occur with probability $\geq 1-2/n$ after using the conditions F1, F2 with the appropriate assumptions.

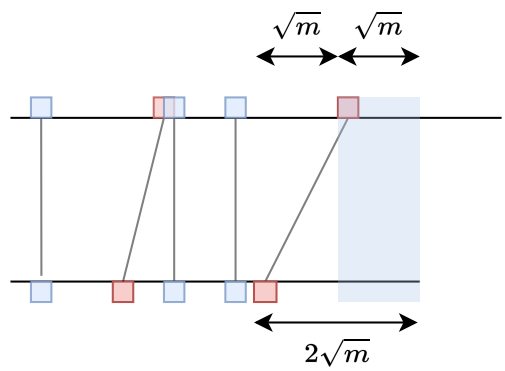
D.4 Proof of Corollary 3

Supplementary Lemma 7. *Under the assumptions of Lemma 7, given any optimal chain $(i_1,j_1),\ldots,(i_u,j_u)$, $j_u-j_1\geq m-4\sqrt{m}$ with probability $\geq 1-2/n$ for large enough n.*

Remark 1. The value $\frac{j_u-j_1}{m}$ is called the aligned fraction and is clearly an upper bound on recoverability. The above Supplementary Lemma shows that the expected value of the aligned fraction is $\geq 1-O(\frac{1}{\sqrt{m}})$.

Proof. Lemma 7 shows the max break size is $m^{1/2}$; we claim that $j_u\geq p+m-2\sqrt{m}$, i.e. j_u is less than $2\sqrt{m}$ away from the end of S'. Suppose otherwise. The argument proceeds in two cases. The idea is to essentially show that adding on all of the homologous anchors near the end of S' always increases the score (under the event space F1 $\cap$ F2).

If (i_u,j_u) is homologous, then $m^{1/2}\gg g(n)$: $g(n)=\Theta(n^{C\alpha}\log^2(n))$ is the maximum distance between homologous k-mers and $m=\Omega(n^{2C\alpha+\epsilon})$. Thus for large enough n, we can find another homologous k-mer near the edge of S'. Adding this homologous k-mer to the chain changes the



Supplementary Figure 5: Given an optimal chain (shown with k-mer anchors in red and blue), if the last k-mer on S' is $2\sqrt{m}$ distance away from the end, because the break size is $< \sqrt{m}$, there will remain at least $\sqrt{m}$ bases left (shaded in blue) near the end of S', S. We use this geometry in the proof of Supplementary Lemma 7 and argue that adding in every possible homologous k-mer in the shaded region gives a more optimal score.

score by at least $-2\zeta g(n)+1$ where the $-2\zeta g(n)$ is the linear cost. Since $\zeta = \frac{1}{6g(n)}$, this is positive, so the old chain was not optimal.

For the non-homologous case, refer to Supplementary Figure 5 for the geometry. Now if (i_u, j_u) is not homologous, then $i_u \in [j_u - \sqrt{m}, j_u + \sqrt{m}]$ because any spurious anchor is contained in a break, which has size $< \sqrt{m}$. Assume $j_u = p + m - L$ where $L > 2\sqrt{m}$ so j_u is L away from the end of S'.

We claim that adding in all homologous anchors in the interval $[\max(i_u, j_u) + 1..p + m]$ gives a more optimal chain. Indeed, since $\max(i_u, j_u) \leq j_u + \sqrt{m}$, $\max(i_u, j_u) < p + m - L + \sqrt{m}$, and thus there are at least $L - \sqrt{m}$ homologous positions to the right of $\max(i_u, j_u)$. Thus there are at least $(L - \sqrt{m})/g(n)$ additional homologous anchors.

The additional linear cost from adding on these homologous anchors near the end is at most $\zeta[(p+m) - (p+m-L)] = \zeta L$ on the side of S' by lengthening the chain from $p+m-L$ to $p+m$, and at most $\zeta[(p+m) - (p+m-L-\sqrt{m})] = \zeta(L+\sqrt{m})$ on the side of S for the same reason. Summing these two terms gives $-\zeta(2L+\sqrt{m})$ as the maximal linear cost penalty of adding these new anchors. However,

$$\#\text{ new homologous anchors } \geq (L-\sqrt{m})/g(m) = \zeta(6L-6\sqrt{m}) > \zeta(2L+8\sqrt{m}-6\sqrt{m}) > \zeta(2L+\sqrt{m})$$

after using the inequality $L > 2\sqrt{m}$. Therefore adding in these homologous k-mers makes the score bigger, contradicting the optimality of our chain. Thus $L \leq 2\sqrt{m}$, and the distance from j_u to the end of S' is at most $2\sqrt{m}$.

The argument works with directions flipped for j_1, so $j_u - j_1 \geq m - 4\sqrt{m}$ as desired.

Corollary 3 *Under the same assumptions as in Lemma 7, the expected recoverability of any optimal chain is* $\geq 1 - O(\frac{1}{\sqrt{m}})$ *for large enough* n.

Proof. Let $\mathcal{F} = \mathrm{F1} \cap \mathrm{F2}$, so all breaks have length $< m^{1/2}$; $\Pr(\mathcal{F}) \geq 1 - 2/n$. Letting $R(\mathcal{C}) = R$ be the recoverability as before and letting $|S'| = m + k - 1 \sim m$ because our final result uses big O notation anyways, we get from Lemma 1 relating recoverability to breaks that

$$\mathbb{E}[R \mid \mathcal{F}] \geq \frac{\mathbb{E}[(j_u - j_1) \mid \mathcal{F}]}{m} - \frac{\mathbb{E}[\sum_{B \in \text{ Breaks}} L(B) \mid \mathcal{F}]}{m} \geq 1 - \frac{4}{\sqrt{m}} - \frac{\mathbb{E}[\mathcal{C}_S m^{1/2} \mid \mathcal{F}]}{m}$$

where we define $\mathcal{C}_S$ to be the number of spurious anchors in an optimal chain. The number of breaks is upper bounded by $\mathcal{C}_S$, and we've used Lemma 7 and Supplementary Lemma 7 to bound $L(B)$ by $m^{1/2}$ and $j_u - j_1$ by $m - 4\sqrt{m}$ respectively. It is clear that $\mathcal{C}_S \leq N_S$ so

$$\mathbb{E}[R \mid \mathcal{F}] \geq 1 - \frac{4}{\sqrt{m}} - \frac{\mathbb{E}[N_S m^{1/2} \mid \mathcal{F}]}{m}.$$

Since $\mathbb{E}[N_S] \leq mn^{1-C}$, we have that $\mathbb{E}[N_S \mid \mathcal{F}] \leq \frac{mn^{1-C}}{1-2/n}$ by the law of total expectation and non-negativity of N_S. Using this, we get that

$$\geq 1 - \frac{4}{\sqrt{m}} - \frac{\sqrt{m}n^{1-C}}{(1-2/n)}$$

and $\Pr(\mathcal{F}) \geq 1 - 2/n$, so

$$\mathbb{E}[R] \geq \mathbb{E}[R \mid \mathcal{F}] \Pr(\mathcal{F}) \geq (1 - \frac{4}{\sqrt{m}})(1 - 2/n) - n^{1-C}\sqrt{m} = 1 - O(\frac{1}{\sqrt{m}}).$$

Remark 2. In the proof of the intermediate lemmas associated with the proof of Corollary 3, we assumed that S' was not too close to the edges of S, i.e. quantities such as $p - \sqrt{m}$ were non-negative. It's not hard to see that if p were close to the edges of S, the breaks would actually be smaller because by definition breaks can not go past the ends of S. One could in fact show that our upper bound still work after accounting for these edge cases, but for simplicity, we omit these edge cases from our proof.

D.5 Proof of Lemma 8

We prove Lemma 8 by proving a series of lemmas.

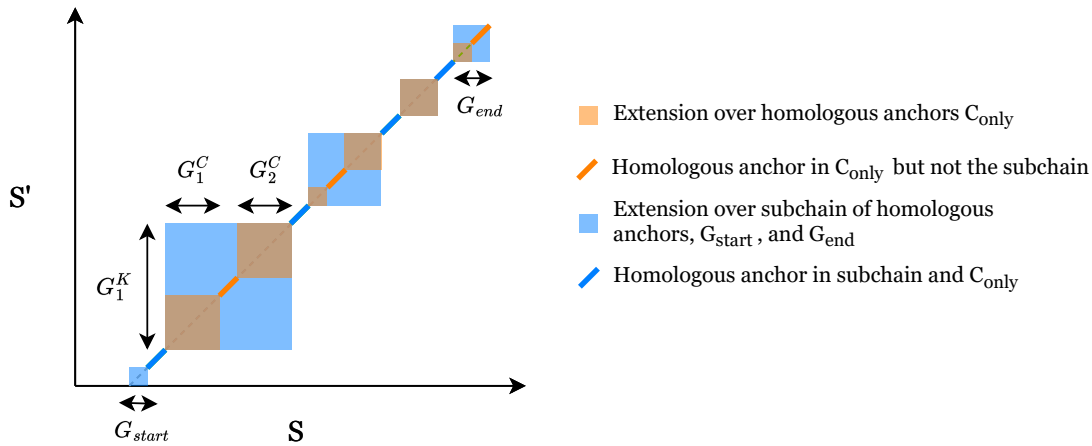
Supplementary Lemma 8. *For a given instance of S, S', let $\mathcal{C}$ be any optimal chain. Let $\mathcal{C}_{only}$ be the chain consisting of only homologous anchors, in S, S'. Defining T^H_{Ext} to be the runtime of extension of $\mathcal{C}$ over only the homologous gaps in $\mathcal{C}$ and $T_{Ext}(\mathcal{C}_{only})$ as the runtime of extension over the homologous gaps of $\mathcal{C}_{only}$, we have that*

$$T^H_{Ext} \leq T_{Ext}(\mathcal{C}_{only}).$$

Proof. Let (i_ℓ, i_ℓ) and $(i_{\ell+1}, i_{\ell+1})$ be two consecutive homologous anchors in an optimal chain $\mathcal{C}$. We can guarantee that no $A(i_\ell + 1, i_\ell + 1), \ldots, A(i_{\ell+1} - 1, i_{\ell+1} - 1)$ random variables are equal 1, otherwise adding such an anchor would improve an optimal chain (it does not incur a linear gap cost penalty due to being flanked by two anchors). Thus (i_ℓ, i_ℓ) and $(i_{\ell+1}, i_{\ell+1})$ are also consecutive anchors in $\mathcal{C}_{only}$, and the homologous gap corresponding to those anchors is also present in $\mathcal{C}_{only}$. Therefore, $T^H_{Ext} \leq T_{Ext}(\mathcal{C}_{only})$ since all homologous gaps in $\mathcal{C}$ are also in $\mathcal{C}_{only}$.

Definition 6. *Define $\mathcal{K} \subset \mathcal{C}_{only}$ to be the subchain of homologous anchors for which the starting positions of the k-mers on S are in $\{p+1, p+1+k, p+1+2k, \ldots, p+1+(\lfloor \frac{|S'|}{k} \rfloor - 1)k\}$.*

That is, the subchain $\mathcal{K}$ consists of only homologous anchors restricted to k-mers that are spaced k bases apart starting from the first index. Since $\mathcal{K}$ is sparser than $\mathcal{C}_{only}$, it should take longer to extend through $\mathcal{K}$ than $\mathcal{C}_{only}$. We formalize this below.



Supplementary Figure 6: A graphical proof of Supplementary Lemma 9. Given a subchain of homologous anchors, the extension time is longer because the square of the larger gaps in the subchain contains the square of the smaller gaps over all homologous anchors. The last gap on the right is not accounted for in the subchain, but adding G_{end}^2 and G_{start}^2 fixes this. Note: the second last orange square on the right is a gap for both chains, i.e. the square is both orange and blue.

Supplementary Lemma 9. *Let G_{start} be the distance from the first anchor of $\mathcal{K}$ to the start of S' and similarly for G_{end} and the last anchor of $\mathcal{K}$ to the end of S'. Then*

$$T_{Ext}(\mathcal{C}_{only}) = O(T_{Ext}(\mathcal{K}) + G_{start}^2 + G_{end}^2)$$

Proof. See Supplementary Figure 6 for a visualization of the proof. Let $\{G_1^K, G_2^K, \ldots, G_r^K\}$ be the homologous gap sizes in $\mathcal{K}$, where we think of G_1^K as the gap size (possibly 0) between the first two anchors, and so forth. Similarly, let $\{G_1^C, G_2^C, \ldots, G_q^C,\}$ be the homologous gap sizes in $\mathcal{C}_{only}$. The extension runtime is

$$T_{Ext}(\mathcal{K}) = \sum_{i=1}^{r} O((G_i^K)^2)$$

and similarly for $T_{Ext}(\mathcal{C}_{only})$. Any two consecutive anchors in $\mathcal{K}$ give rise to a gap G_i^K. These two anchors also exist on $\mathcal{C}_{only}$, but there may be intermediate anchors, so this gives rise to multiple gaps $G_j^C, \ldots, G_{j+\ell}^C$ between these two anchors on $\mathcal{C}_{only}$. Now $G_j^C + \ldots + G_{j+\ell}^C \leq G_i^K$ because the sum of all intermediate gaps is at most the size of the larger gap, so

$$(G_j^C)^2 + \ldots + (G_{j+\ell}^C)^2 \leq (G_i^K)^2.$$

Let G_a^C be the leftmost gap on $\mathcal{C}_{only}$ to the right of the first anchor of $\mathcal{K}$, and G_b^C be the gap to the left of the last anchor of $\mathcal{K}$. Considering every pair of consecutive anchors on $\mathcal{K}$, we get

$$\sum_{i=1}^{r} (G_i^K)^2 \geq \sum_{i=a}^{b} (G_i^C)^2$$

by the previous inequality. However, we're not done yet because the leftmost gaps in C_{only} may not be contained by any gap in $\mathcal{K}$. Using the definition of G_{start} as the "gap" on $\mathcal{K}$ containing all of the leftmost gaps on $\mathcal{C}_{only}$ and similarly for G_{end}, we get

$$G_{start}^2 \geq \sum_{i=1}^{a-1}(G_i^C)^2, G_{end}^2 \geq \sum_{i=b+1}^{q}(G_i^C)^2$$

by the same arguments as above. Combining both inequalities gives the result.

By the above results, we can bound T_{Ext}^H by either $T_{Ext}(\mathcal{C}_{only})$ or $T_{Ext}(\mathcal{K})$. We will work with $\mathcal{K}$ in this section, but we will actually use $\mathcal{C}_{only}$ for the sketched version of the main theorem. Below we give a random variable to calculate these extension times.

Definition 7. *Let Y_i be the random variable representing the number of uncovered bases between a homologous anchor starting at position i, if it exists, and the next homologous anchor. So $Y_i = \ell \geq 1$ if the bases $[i..i+k-1]$ are unmutated, the bases at $[i+\ell+k..i+\ell+2k-1]$ are unmutated, and there are no homologous anchors covering any bases in $[i+k..i+\ell+k-1]$. Otherwise, $Y_i = 0$. Let Y_i^K be the same random variable except only considering homologous anchors with starting positions in $\{p+1, p+1+k, p+1+2k, \ldots, p+1+(\lfloor\frac{|S'|}{k}\rfloor - 1)k\}$.*

Under our original definition, the extension time over $\mathcal{C}_{only}$ would be $\sum_{\ell=1}^{u-1} O(G_\ell^2)$ where G_ℓ is the size of the gap, but u is a random variable. It's clear that $\sum_{i=p+1}^{p+m} O(Y_i^2)$ is the runtime of extension over $\mathcal{C}_{only}$, but it will be easier to handle for our proof now that the upper index is not a random variable. Similarly, $\sum_{i=p+1}^{p+m} O((Y_i^K)^2)$ is the runtime of extension over $\mathcal{K}$.

We will now work with the chain $\mathcal{K}$ restricted to equally spaced k-mers and their unmutated homologous anchors. We upper bound this extension time in expectation, which will upper bound $\mathbb{E}[T_{Ext}^H]$ as well.

Lemma 8 *Let T_{Ext}^H be the time of extension over only the homologous gaps of an optimal chain. $\mathbb{E}[T_{Ext}^H] = O(mn^{C\alpha}\log n)$.*

Proof. By Supplementary Lemmas 9 and 8, to bound $\mathbb{E}[T_{Ext}^H]$ we can bound $\mathbb{E}[G_{start}^2] + \mathbb{E}[G_{end}^2] + \sum_{i=p+1}^{p+m}\mathbb{E}[(Y_i^K)^2]$. The random variable $G_{start} = \ell \cdot k$ if for the first ℓ k-mers, $A(1,1) = 0, A(1+k, 1+k) = 0, \ldots, A(1+(\ell-1)k, 1+(\ell-1)k) = 0$ but $A(1+\ell k, 1+\ell k) = 1$. In other words, the first ℓ k-mers that are spaced k bases apart are mutated, but the $\ell+1$th such k-mer is not mutated. Thus

$$\Pr(G_{start} = \ell \cdot k) \leq (1-(1-\theta)^k)^\ell (1-\theta)^k.$$

Importantly, we used the fact that k-mers spaced k distance apart are independent of each other. The $\leq$ comes from the fact that G_{start} can not be larger than $|S'|$. It follows that

$$\mathbb{E}[G_{start}^2] \leq (1-\theta)^k \sum_{\ell=1}^{\infty}(\ell k)^2(1-(1-\theta)^k)^\ell = O\left(\frac{k^2(1-\theta)^k}{(1-\theta)^{3k}}\right) = O(n^{2C\alpha}\log^2 n)$$

by the formula

$$\sum_{i=1}^{\infty} i^2 x^i = \frac{x(x+1)}{(1-x)^3}$$

which follows from geometric series manipulations. Notice that $O(n^{2C\alpha}\log^2 n) = O(m)$ because $m = \Omega(n^{2C\alpha+\epsilon})$. It's clear that the same argument holds for G_{end}^2, so both terms are $O(m)$.

Now for Y_i^K, we have that $Y_i^K = \ell \cdot k$ where $\ell > 0$ only if $i \in \{p+1, p+1+k, p+2+k, \ldots\}$, the k-mer at i is unmutated, the k-mers that are $k, 2k, 3k, \ldots, \ell k$ bases ahead from i are mutated, and the k-mer $(\ell+1)\cdot k$ bases ahead of i is unmutated. Thus

$$\Pr\big(Y_i^K = \ell \cdot k\big) \leq (1-\theta)^{2k}(1-(1-\theta)^k)^\ell \text{ if } i \in \{1, 1+k, \ldots\} \text{ and 0 otherwise.}$$

We can compute the expectation over all Y_i^K and only pick out the non-zero random variables where $i \in \{p+1, p+1+k, \ldots\}$. There are at most m/k such random variables, so we get

$$\sum_{i=p+1}^{p+m} \mathbb{E}[(Y_i^K)^2] \leq \frac{m}{k}(1-\theta)^{2k} \sum_{\ell=1}^{\infty} (\ell k)^2 (1-(1-\theta)^k)^\ell$$

$$= O\left(\frac{mk^2(1-\theta)^{2k}}{k(1-\theta)^{3k}}\right) = O(mn^{C\alpha}\log n).$$

The expectations in all the terms are $O(mn^{C\alpha}\log n)$, so this finishes the proof.

D.6 Proof of Lemma 9

Supplementary Lemma 10 (F1 + F2). *Under the same assumptions of Lemma 7 all non-homologous gaps in an optimal chain have size* $< \sqrt{m} + \frac{50C}{8}\log(n)\ln(n)n^{C\alpha}$ *on both* S *and* S' *with probability* $\geq 1 - 2/n$ *and large enough* n.

Proof. Let us be in the space $\mathcal{F} = (\text{F1}) \cap (\text{F2})$, which holds with probability $\geq 1-2/n$. There are two types of non-homologous gaps; non-homologous gaps that are flanked by two spurious anchors and gaps that are flanked by only one.

If a non-homologous gap is flanked by two spurious anchors, then it is part of a break. The gap size must be smaller than the break, which has length less than $\sqrt{m}$.

Suppose a non-homologous gap is flanked by one spurious anchor $A(i,j)$ and one homologous anchor $A(h,h)$. Suppose WLOG that $|i-h| > |j-h|$. We know with high probability that $|j-h| < g(n) = 50/8C\log(n)\ln(n)n^{C\alpha}$ as otherwise there will be a homologous anchor between $A(i,j)$ and $A(h,h)$ (by property F2/Lemma 6) and the chain is not optimal as we could just insert the additional homologous anchor between j and h. Furthermore, $|i-j| < \sqrt{m}$ as $|i-j|$ is less than the break size. We can see then that $|i-h| \leq |i-j| + |j-h| < \sqrt{m} + 50/8C\log(n)\ln(n)n^{C\alpha}$ as desired.

Lemma 9 *Let* T_{Ext}^S *be the runtime of extension through only the non-homologous gaps of an optimal chain. Under the same assumptions as in Lemma 7,* $\mathbb{E}[T_{Ext}^S] = O(m)$.

Proof. Defining γ as the maximum non-homologous gap size as in Supplementary Lemma 10 conditional on $\mathcal{F}$ =F1 $\cap$ F2, $\gamma = O(m^{1/2} + n^{C\alpha}\log^2(n)) = O(m^{1/2})$ when $m = \Omega(n^{2C\alpha+\epsilon})$. We bound the expected value of the non-homologous gaps as follows.

$$\mathbb{E}[\sum_{G_j \text{ not homologous}} G_j G_j' \mid \mathcal{F}] \leq \mathbb{E}[2\mathcal{C}_S \cdot \gamma^2 \mid \mathcal{F}]$$

The inequality uses $\gamma > G_j$ and that the number of non-homologous gaps is at most 2 times $\mathcal{C}_S$, which we define to be the number of spurious anchors in the chain (each anchor gives rise to at most two unique gaps). $\mathcal{C}_S \leq N_S$ follows trivially, so

$$\leq \mathbb{E}[2\gamma^2 \cdot N_S \mid \mathcal{F}] \leq \frac{2n^{2-C}\gamma^2}{1-2/n} = O(mn^{2-C}).$$

The first inequality follows from $\mathbb{E}[N_S \mid \mathcal{F}]\Pr(\mathcal{F}) \leq \mathbb{E}[N_S]$ and non-negativity of N_S, as well as $\Pr(\mathcal{F}) \geq 1-2/n$. Finally, under $\overline{\mathcal{F}}$, the worst case extension through non-homologous gaps is just $O(nm)$ as in section "Extension and chaining runtimes". Since $\Pr(\overline{\mathcal{F}}) \leq 2/n$, the expected runtime is $O(nm/n + mn^{2-C}) = O(m)$ as desired.

E Missing proofs from "Sketching and local k-mer selection"

E.1 Proof of Lemma 11

Lemma 11 *The variance $Var(N_S^*)$ can be upper bounded by $\frac{1}{c}8k^2mn^{1-C}$. Furthermore, $\mathbb{E}[N_H^{*2}] \leq \frac{1}{c}2mk(1-\theta)^k + \frac{1}{c^2}m^2(1-\theta)^{2k}$ and $\mathbb{E}[N_H^* N_S^*] \leq \frac{1}{c}4k\frac{mn}{\sigma^k} + \frac{1}{c^2}m^2n(1-\theta)^k\frac{1}{\sigma^k}$.*

Proof. The proof follows almost exactly the same as Lemma 3. We only do the variance bound as an example, and the other moment bounds follow exactly the same way.

We upper bound the three sums S_1, S_2, S_3 in the proof of Lemma 3 but now with the sketched versions S_1^*, S_2^*, S_3^*. The bounds $\mathbb{E}[S_2^*]$ and $\mathbb{E}[S_3^*] \leq \frac{1}{c}4k^2mn^{1-C}$ hold by a restatement of the argument using $\mathbb{E}[A(i,j)^*A(h,l)^*] \leq E[A(i,j)^*] = \frac{1}{c}n^C$ and the set $|B_k(i,j)| \leq 4k^2$.

The bound for $\mathbb{E}[S_1^*] \leq \frac{1}{c^2}m^2(n-1)^2n^{-2C} = \mathbb{E}[N_S^*]^2$ also holds as well; we just have to calculate $\mathbb{E}[A(i,j)^*A(h,l)^*]$ when $A(i,j)$ and $A(h,l)$ are independent. This is

$$\mathbb{E}[J(i)J'(j)J(h)J'(l) \mid A(i,j)=1, A(h,l)=1]\Pr(A(i,j)=1, A(h,l)=1).$$

By independence of $A(i,j), A(h,l)$ in S_1^*, the second term is n^{-2C}. Now notice that under the conditions of S_1, either $|i \geq h| > k$ or $|j-l| \geq k$ meaning that two of the k-mers along either S or S' are independent. WLOG we can assume it is i and h; thus $J(i)$ and $J(h)$ are independent. The $A(i,j)=1$ condition implies $J(i)=J'(j)$ and similarly for h,l, so

$$\mathbb{E}[J(i)J'(j)J(h)J'(l) \mid A(i,j)=1, A(h,l)=1] = \mathbb{E}[J(i)J(h)] = \frac{1}{c^2}.$$

E.2 Proof of Theorem 8

Theorem 8 *Under the same assumptions as in Theorem 6, letting N^* be the total number of sketched k-mer anchors, the expected chaining time is $O(\mathbb{E}[N^* \log N^*]) = O(\frac{1}{c}mn^{-C\alpha}\log m)$.*

Proof. The argument follows in the same way as in the proof of Theorem 6, only using the sketched moment bounds in Lemma 11. We still get that $\mathbb{E}[(N^*)^2]/\mathbb{E}[N^*] = O(mn^{-C\alpha}/c)$ and the result follows.

E.3 Sketched concentration bounds and gap sizes

Supplementary Lemma 11. *Let $\tau = (c+1)/2$ and $1 > \beta \geq 0$. Given $\ell + \tau - 1$ k-mers indexed by the interval $[a..b]$ on a random string S,*

$$\Pr\left(\sum_{i=a}^{b} J(i) \leq \frac{(1-\beta)\ell}{c}\right) \leq \exp\left(-\frac{8\ell\beta^2}{(k+\tau)50}\right).$$

Remark 3. The variables $J(i)$ are k-dependent so we could use Theorem 5 for this sum. This however leads to a bound like $\exp(-O(m/(kc)))$ instead of our bound which is $\exp(-O(m/(k+c)))$ in the above lemma. While the former bound still leads to the same asymptotic behavior for Supplementary Lemma 13, which we are ultimately trying to prove, we believe it is enlightening to show how the dependence structure of the $J(i)$s can be used to obtain a better bound.

Proof. Define $J(i, i+\tau) = J(i) + \ldots + J(i+\tau-1)$. Then

$$\sum_{i=a}^{b-\tau+1} J(i, i+\tau) \leq \tau \sum_{i=a}^{b} J(i).$$

Let $\min(J(i, i+\tau), 1) = I(i, i+\tau)$ and clearly $\sum_{i=a}^{b-\tau+1} I(i, i+\tau) \leq \sum_{i=a}^{b-\tau+1} J(i, i+\tau)$.

We will show that $\sum_{i=a}^{b-\tau+1} I(i, i+\tau)$ is large, implying that $\tau \sum_{i=a}^{b} J(i)$ is also large. Notice that $I(i, i+\tau)$ is the random variable which is 1 when some k-mer in $[i..i+\tau)$ is an open syncmer. By Theorem 7, only one of the J variables in $J(i), \ldots, J(i+\tau-1)$ can be equal to one. Using $\Pr(J(i) = 1) = \frac{1}{c}$, we get that

$$\Pr(I(i, i+\tau) = 1) = \Pr\left(\bigcup_{j=i,\ldots,i+\tau-1} J(j) = 1\right) = \frac{\tau}{c}$$

by disjointness. $I(i, i+\tau)$ is $k+\tau-1$-dependent because it examines the $k+\tau-1$ bases starting from position i, so using Theorem 5 we get that

$$\Pr\left(\sum_{i=a}^{b-\tau+1} I(i, i+\tau) \leq \frac{(1-\beta)\tau\ell}{c}\right) \leq \exp\left(-\frac{\beta^2 8\tau\ell}{c(k+\tau)25}\right).$$

From our inequality $\tau \sum_{i=a}^{b} J(i) \geq \sum_{i=a}^{b-\tau+1} 1(i, i+\tau)$ and using $\tau = (c+1)/2$ we obtain

$$\Pr\left(\sum_{i=a}^{b} J(i) \leq \frac{(1-\beta)\ell}{c}\right) \leq \exp\left(-\frac{8\ell\beta^2}{[k+(c+1)/2]50}\right)$$

Supplementary Lemma 12. *For any interval $[a..b]$ containing $\ell + \frac{c+1}{2} - 1$ k-mers and $1 > \beta \geq 0$, the probability that all $A^*(a,a), \ldots, A^*(b,b)$ are 0 is upper bounded by*

$$\Pr\left(\sum_{i=a}^{b} A^*(i,i) = 0\right) \leq \exp\left(-\frac{8\ell\beta^2}{[k+(c+1)/2]50}\right) + \exp\left(-\frac{8\ell(1-\beta)(1-\theta)^k}{50k+25c}\right).$$

Proof. Defining $\mathcal{J} = \{i \in [a..b] : J(i) = 1)\}$, we have $\sum_{i=a}^{b} A^*(i,i) = \sum_{i \in \mathcal{J}} A(i,i)$ where we just sum only over syncmer anchors instead of all k-mer anchors in the range.

$$\Pr\left(\sum_{i=a}^{b} A^*(i,i) \leq 0\right) = \Pr\left(\sum_{i \in \mathcal{J}} A(i,i) \leq 0\right)$$

$$\leq \Pr\left(\sum_{i \in \mathcal{J}} A(i,i) \leq 0 \mid |\mathcal{J}| > \frac{(1-\beta)\ell}{c}\right) + \Pr\left(|\mathcal{J}| \leq \frac{(1-\beta)\ell}{c}\right).$$

Now by Theorem 7, the distance between consecutive positions in $\mathcal{J}$ is at least $(c+1)/2$ (remembering that we assume c is odd for simplicity), which means that given the ith open syncmer, the $(i + \lceil \frac{2k}{c+1} \rceil)$th open syncmer is more than k bases apart from the ith open syncmer and hence independent from each other. Thus the $A(i,i)$s in $\mathcal{J}$ are now $\lceil \frac{2k}{c+1} \rceil < \frac{2k}{c+1} + 1$ dependent. Conditioning on $|\mathcal{J}| > \frac{(1-\beta)\ell}{c}$, we can bound the first term using Theorem 5 to get

$$\Pr\left(\sum_{i\in\mathcal{J}} A(i,i) \le 0 \mid |\mathcal{J}| \ge \frac{(1-\beta)\ell}{c}\right) \le \exp\left(-\frac{8(1-\theta)^{2k}|\mathcal{J}|^2}{25(\frac{2k}{c+1}+1)|\mathcal{J}|(1-\theta)^k}\right) \le \exp\left(-\frac{8\ell(1-\beta)(1-\theta)^k}{50k+25c}\right).$$

Where we substituted in $|\mathcal{J}| = \frac{(1-\beta)\ell}{c}$ in the above equation and used $\frac{c}{c+1} \le 1$ to remove the term. To bound the second term, we just use Supplementary Lemma 11. This furnishes the final result

$$\Pr\left(\sum_{i=a}^{b} A^*(i,i) = 0\right) \le \exp\left(-\frac{8\ell\beta^2}{[k+(c+1)/2]50}\right) + \exp\left(-\frac{8\ell(1-\beta)(1-\theta)^k}{50k+25c}\right).$$

Supplementary Lemma 13. *No homologous gap has size greater than*

$$g'(n) = \left(\frac{3}{2(1-\theta)^k}+2\right)\frac{200k}{8}\ln(n) = \frac{200C}{8}\log(n)\ln(n)\cdot\left(\frac{3}{2}n^{C\alpha}+2\right)$$

with probability $\ge 1-2/n$ *after ignoring an additive* $(c+1)/2-1+C\log n$ *term.*

Proof. We use Supplementary Lemma 12 after plugging in the value of $\ell = g'(n)$ and letting $\beta = 1/2$. Algebraic manipulations show that the probability is less than $\le 2/n^2$ using $k = C\log n$ and the inequality $c < k$. Then as in the proof of Theorem 6 we can use indicator random variables and Markov's inequality in the same way to get the result.

E.4 Proving Lemma 13

We use the same definitions and supplementary lemmas as in Section D.5 for Y_i, $\mathcal{C}_{only}$ and $\mathcal{K}$, now considering the sketched versions Y_i^*, $\mathcal{C}_{only}^*$, and $\mathcal{K}^*$. We first note that the sketched version for Supplementary Lemmas 8 and 9 hold by the same arguments. The analogous version of Lemma 8 holds as well.

Supplementary Lemma 14. *Let* T_{Ext}^{H*} *be the time of sketched extension over only the homologous gaps of an optimal chain.* $\mathbb{E}[T_{Ext}^{H*}] = O(c\cdot mn^{C\alpha}\log n)$.

Proof. Essentially the same argument as in the proof of Lemma 8 outlined in Section D.5. We end up bounding $\mathbb{E}[T_{Ext}(\mathcal{K}^*)] + \mathbb{E}[(G_{start}^*)^2] + \mathbb{E}[(G_{end}^*)^2]$; the only difference is that $(1-\theta)^k \mapsto \frac{(1-\theta)^k}{c}$ due to sketching. Ultimately, the main term becomes

$$\sum_{i=p+1}^{p+m} \mathbb{E}[(Y_i^{K*})^2] \le \frac{m}{k}\frac{(1-\theta)^{2k}}{c^2}\sum_{\ell=1}^{\infty}(\ell k)^2\left(1-\frac{(1-\theta)^k}{c}\right)^{\ell}.$$

The same geometric series manipulation as before gives us that this is

$$= O\left(\frac{mkc^3}{(1-\theta)^k c^2}\right) = O(c\cdot mn^{C\alpha}\log n).$$

This finishes the proof.

When c is fixed to be a constant independent of n, we get the same asymptotic bound as before. This bound suggests that sketching makes extension slower. However, we can actually do better than this when c grows with n. This time, we proceed in a different manner, instead directly using $\mathcal{C}_{only}^*$ to bound the extension time. We let Y_i^* be the sketched version of Y_i as in Definition 7, which is a random variable measuring gap sizes between anchors.

Supplementary Lemma 15. *For $\ell \geq 1$,*

$$\Pr(Y_i^* = \ell) \leq \frac{(1-\theta)^{2k}}{c^2} \cdot \Pr\left(\sum_{j \in \mathcal{H}_i(k,\ell)} A(j,j)^* = 0\right) \tag{1}$$

where $\mathcal{H}_i(k,\ell) = [i+k..i+\ell]$ if $\ell \geq k$, and is empty otherwise. Thus $|\mathcal{H}_i(k,\ell)| \geq \ell - k + 1$.

Proof. $Y_i^* = \ell$ is equivalent to the k-mer covering $[i..i+k-1]$ being unmutated (i.e. no bases on the k-mer are mutated) and selected as an open syncmer, all of the k-mers in between these two flanking k-mers not being present, and the k-mer covering $[i+\ell+k..i+\ell+2k-1]$ being unmutated and selected as an open syncmer. Calling these events H_1, H_2, and H_3 respectively, $\Pr(Y_i = \ell) = \Pr(H_1 \cap H_2 \cap H_3)$.

The k-mers considered in H_2 that lie in between the flanking k-mers overlap the flanking k-mers in H_1 and H_3, so these events are not independent. To upper bound H_2, let H_2' be the event that the k-mers lying *completely in the interval* $[i+k..i+\ell+k-1]$ on S, and not just overlapping the interval, are mutated or not selected as an open syncmer. $H_2' \supset H_2$ as events. This set of k-mers is exactly the k-mers starting at positions in $\mathcal{H}_i(k,\ell)$ after some examination. Now notice that H_1, H_2', H_3 are all independent as the k-mers in each event lie on non-overlapping bases, so

$$\Pr(Y_i^* = \ell) \leq \Pr\big(H_1 \cap H_2' \cap H_3\big) = \frac{(1-\theta)^{2k}}{c^2} \Pr\left(\sum_{j \in \mathcal{H}_i(k,\ell)} A(j,j)^* = 0\right).$$

In the case that $[i+\ell+k..i+\ell+2k-1]$ does not lie on S', i.e. $i+\ell+k > p+m$, then $\Pr(Y_i = \ell) = 0$ as a homologous gap of this length can not exist near the edges, so the upper bound still holds. The upper bound also holds if $i-k+1 < p+1$ by the same reason. This finishes the proof.

Lemma 13 *The expected runtime of sketched extension through the homologous gaps in an optimal chain is $\mathbb{E}[T_{Ext}^{H*}] = O\Big(\min\big(\frac{1}{c^2} mn^{C\alpha}\log^3(n), c \cdot mn^{C\alpha}\log n\big)\Big)$. If $c = \Theta(\log n)$ then $\mathbb{E}[T_{Ext}^{H*}] = O(mn^{C\alpha}\log n)$.*

Proof. By Supplementary Lemma 12 and our discussion that $|\mathcal{H}_i(k,\ell)| \geq \ell - k + 1$,

$$\Pr\left(\sum_{j \in \mathcal{H}_i(k,\ell)} A(j,j)^* = 0\right) \leq O\left(\exp\left(-\frac{8(\ell-k+1)(1-\beta)(1-\theta)^k}{75k}\right)\right)$$

holds by noticing that the first term in Supplementary Lemma 12 is dominated by the second term and using $k > c$, where we also ignored the $-\frac{c+1}{2} + 1$ in Supplementary Lemma 12 because it is asymptotically small. We can now bound the expected time of extension over homologous gaps by $\mathbb{E}[T_{Ext}(\mathcal{C}_{only}^*)] = \sum_{i=p+1}^{p+m} O(\mathbb{E}[(Y_i^*)^2])$. Since the probability densities of each Y_i^* is upper bounded by Equation 1,

$$\sum_{i=p+1}^{p+m} O(\mathbb{E}[(Y_i^*)^2]) = O\left(m\frac{(1-\theta)^{2k}}{c^2}\exp((k-1)D)\sum_{\ell=1}^{\infty}\ell^2\exp(-D\ell)\right)$$

where $D = \frac{8(1-\beta)(1-\theta)^k}{75k}$ and β is some constant. Now we use the fact that

$$\sum_{\ell=1}^{\infty}\ell^2\exp(-D\ell) \leq O\left(\frac{1}{(1-e^{-D})^3}\right) = O\big(\frac{1}{D^3}\big)$$

after geometric series manipulations as before, using Taylor's theorem for $1 - e^{-D}$, and noticing that $D = o(1)$. Thus our final bound is

$$\mathbb{E}[T_{Ext}^{H*}] = O\left(m\frac{(1-\theta)^{-k}}{c^2}k^3\exp((k-1)D)\right) = O(\frac{1}{c^2}mn^{C\alpha}\log^3 n).$$

The other bound $c \cdot mn^{C\alpha}\log n$ follows from Supplementary Lemma 14, so taking the minimum over both bounds yields the result.

Remark 4. Intuitively, what is happening here is that open syncmers have the same key property as $\mathcal{K}$; they are spaced at least $\frac{c+1}{2} \sim \frac{k}{2}$ bases apart when $c \sim k$, and are "almost independent" like the anchors in $\mathcal{K}$. C^*_{only} has similar properties to the non-sketched $\mathcal{K}$ chain, showing why they give the same bounds. Interestingly, if were to bound $\mathcal{C}_{only}$ to get a result for the non-sketched version $\mathbb{E}[T_{Ext}^{H}]$ while using the same techniques as above (with the dependent Chernoff-Hoeffding bound), the result would be $O(mn^{C\alpha}\log^3 n)$, worse than bounding using $\mathcal{K}$ by $\log^2(n)$.

Remark 5. If we were to use the FracMinHash method or another k-mer selection/sketching method without the distance guarantee provided by Theorem 7, Supplementary Lemma 12 would not hold, and therefore the above argument would fail. Supplementary Lemma 14 would however still hold.

E.5 Re-proving sketched bounds

Supplementary Lemma 16 (F1* + F2*). *Using the same assumptions as in Lemma 7, there are no breaks of length $> m^{1/2}$ with probability greater than $1 - 3/n$ in an optimal chain for large enough n.*

Proof. In Section D.3, the structure of the problem does not change. The only difference is the bounds on gap size and spurious k-mers. Asymptotically the homologous gap size is still the same, and both the number of spurious anchors and homologous anchors are smaller by a factor of $\sqrt{c}$, which does not affect the inequalities. Thus all of the bounds in Section D.3 hold except with slightly different probabilities since F1* holds with probability $\geq 1 - 2/n$ instead of $1 - 1/n$.

Supplementary Lemma 17. *Using the same assumptions as in Lemma 7, the expected recoverability of any optimal chain is $\geq 1 - O(1/\sqrt{m})$ for large enough n.*

Proof. The proof is the same as the proof of Corollary 3 in Section D.4 except we use $\mathbb{E}[N_S^*] \leq \frac{1}{c}mn^{1-C}$ and the failure probability is $\geq 1 - 3/n$. This doesn't change the main $1 - 4/\sqrt{m}$ term, so the same big O bound holds.

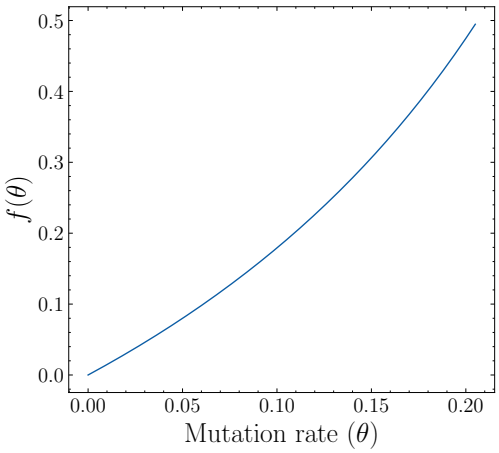
Supplementary Lemma 18. *Under the assumptions of Lemma 7, the expected value of runtime of extension through non-homologous gaps is $\mathbb{E}[T_{Ext}^{S*}] = O(m)$.*

Proof. The proof is the same as the proof of Lemma 9 in Section D.6 except the sketched maximum non-homologous gap length is $m^{1/2} + g'(n)$. This is still $O(m^{1/2})$, and $N_S^* < N_S$, so the expected value of the non-homologous runtime does not change.

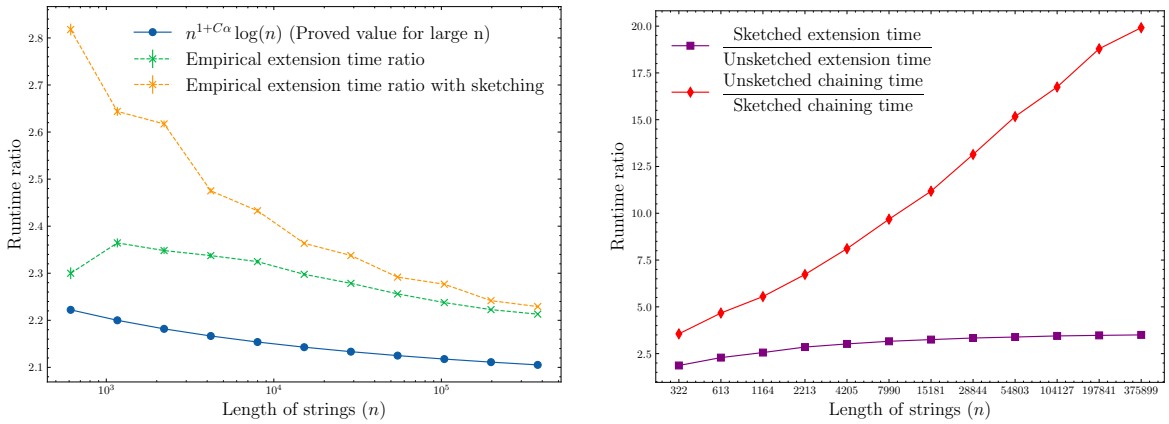
Theorem 2 *In addition to the hypotheses outlined in Theorem 1, let $c = O(\log n) < k$ and $\zeta = \frac{1}{6g'(n)}$ instead where $g'(n) = C\frac{200}{8}\log(n)\ln(n)(\frac{3}{2}n^{C\alpha} + 2)$. For open syncmer sketched seed-chain-extend, the expected running time is $O\big(\min(\frac{1}{c^2}mn^{C\alpha}\log^3(n), c \cdot mn^{C\alpha}\log n)\big)$ for extension and $O(\frac{1}{c}mn^{-C\alpha}\log m)$ for chaining. The expected recoverability of any optimal chain is $\geq 1 - O(\frac{1}{\sqrt{m}})$.*

Proof. The expected runtime of chaining and recoverability follows from Theorem 8 and Supplementary Lemma 17. Lemma 13 and Supplementary Lemma 18 give the runtimes under our conditional event space F1*, F2*, which occurs with probability $\geq 1 - 3/n$, so the same argument as in the proof of Theorem 1 gives the result.

F Additional figures and tables



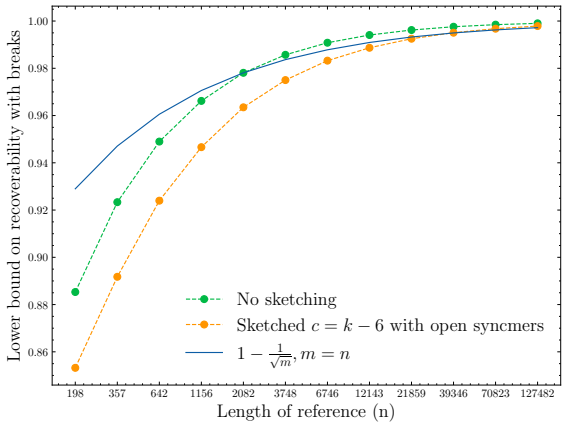
Supplementary Figure 7: The function $f(\theta)$ for the runtime $O(mn^{f(\theta)} \log n)$. Technically, $f(\theta) = \frac{2\alpha}{1-2\alpha} + \delta$ for any $\delta > 0$, so we plot $\frac{2\alpha}{1-2\alpha}$ where α is a function of θ defined as $\alpha = -\log_4(1-\theta)$. This function is convex on $[0, 0.206]$ so $f(\theta) < \frac{0.5}{0.206}\theta < 2.43 \cdot \theta$.



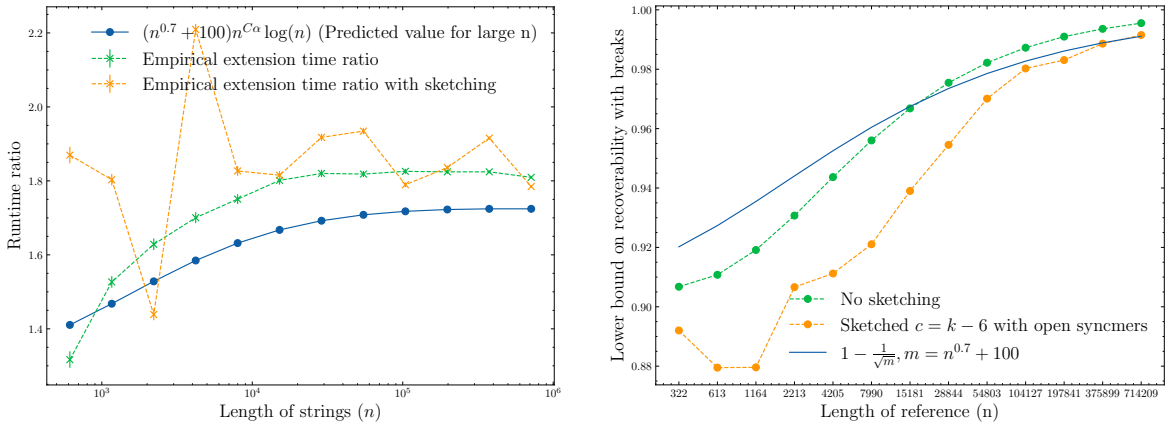
(a) Empirical extension runtime ratios.

(b) Multiplicative speed up of chaining vs slow-down of extension.

Supplementary Figure 8: The same experiment as in Figure 1 (a) but with $\theta = 0.05$.



Supplementary Figure 9: The recoverability of our alignments of two length n strings over 50,000 iterations as a function of sequence length n where $\theta = 0.10$ and k was chosen as described in section "Simulated genome alignment experiments". Breaks were uncommon and most of the recoverability loss is from the $(j_u - j_1)$ term in the recoverability definition, i.e. the chain length being smaller than the sequence length. The lower bound on recoverability is $1 - O(\frac{1}{\sqrt{m}})$ where in Theorem 1 ($m = n$ in this case), and we plotted $1 - \frac{1}{\sqrt{m}}$ as a non-asymptotic proxy for the asymptotic bound.



(a) Empirical extension runtime ratios when aligning a substring of length $m = n^{0.7} + 100$.

(b) Empirical recoverability when aligning a substring of length $m = n^{0.7} + 100$.

Supplementary Figure 10: Empirical runtimes and recoverability with $\theta = 0.05$ but aligning a mutated substring of length m where $m < n$. We let $m = n^{0.7} + 100$ vary as a function of n.

Genome Name	Genome ID	SRA Read Accession/URL link	Subsampling
SARS-CoV-2	NC_045512.2	SRR22765637	None
Escherichia coli	NC_007779.1	DRR198814	100k
Magnaporthe oryzae	CM001231.1	ERR9878936	100k
Drosophila melanogaster	NC_004354.4	SRR21160205	100k
Homo sapiens	GCA_009914755.3	`https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/nanopore/rel3/`	500k

Supplementary Table 1: Data sets used for the real nanopore data alignment experiments. Reads were randomly subsampled to the specified number when data sets were too large. We subsampled to 500k reads for the human data set because the reads were more erroneous and had a wider length distribution.