Table of Contents

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 $\mathit{Align}(\mathcal{C})$

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

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

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. $\mathit{Ext}(\ell)$ corresponds to the green boxes in Figure 5b. The set $\mathit{Ext}(\ell)$ is empty if $i_\ell + k > i_{\ell+1} - 1$ and similarly for j, i.e. there are no gaps between possibly overlapping anchors. $\mathit{Align}(\mathcal{C})$ also takes into account the k-mer matches given by $\mathcal{C}$. Thus we define $\mathit{Align}(\mathcal{C})$ formally as

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

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 $\mathit{Align}(\mathcal{C})$ with the homologous diagonal, we can manipulate the expression for $\mathit{Align}(\mathcal{C})$ to get something more tractable. Let $\mathrm{Diag}[a..b] = \{(x, x) : x \in [a..b]\}$, $\mathcal{C}_H$ 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

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

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

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

We define

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

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 $(x, x) \in Ext(\ell) \iff i_\ell + k \le x \le i_{\ell+1} - 1$ and $j_\ell + k \le x \le j_{\ell+1} - 1$. Finally,

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

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

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

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.

Supplemental Lemma S1.

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

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

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

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 $D_H = [p + 1..p + m] \times [p + 1..p + m]$ (we let $|S'| \sim m$ here), 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 $(j_u - j_1 - L(\mathcal{C}))/|S'| \le R(\mathcal{C})$.*

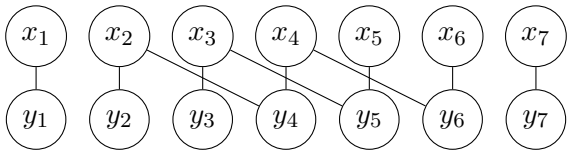
Proof. Using the supplemental lemma above, we note that $j_u \le p + m$ and $j_1 \ge p + 1$ since the anchors must lie between $[p + 1..p + m]$, so after ignoring the k term we get

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

We now prove that the breaks "cover" NR, so $|NR| \le L$ 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), $\min(i, j) \le x \le \max(i, j) + k - 1$. 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 $\bigcup_{\text{breaks } B \text{ in } \mathcal{C}} B \supset \pi(NR)$. $L(\mathcal{C}) \ge |\pi(NR)| = |NR|$, 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



Supplemental Figure S1: 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).

Supplemental Lemma S2. *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 Supplemental Lemma S2, 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

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

Finally, all $C_\sigma^n(A), n = 1, \dots, \sigma - 1$ 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.

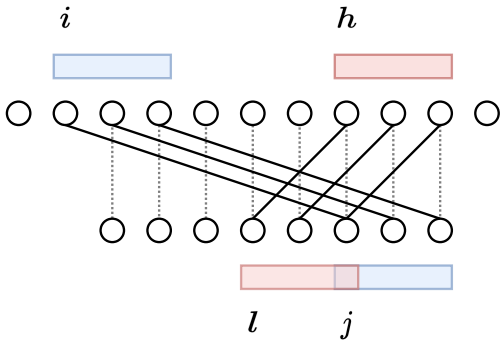
C.2 Proof of Corollary 1

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 $M(i,j), M(i+1,j+1), \dots, M(i+k-1, j+k-1)$ 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 $(x_i, y_j), \dots, (x_{i+k-1}, y_{j+k-1})$ but we give a rigorous argument below. See Supplemental Figure S1 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 $M(i,i), M(i+1,i+1), \dots$ are functions of sets of mutually independent random variables, and $\Pr(M(i,i) = 1) = (1-\theta)$.



Supplemental Figure S2: 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 $k = 3$.

C.3 Proof of Lemma 2

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 Supplemental Figure S2 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 $\mathbb{E}[N_S^2] \leq 8k^2mn\frac{1}{\sigma^k} + \mathbb{E}[N_S]^2$. *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* $\mathbb{E}[N_H N_S] \leq 4k\frac{mn}{\sigma^k} + m^2n(1-\theta)^k\frac{1}{\sigma^k}$.

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

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

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

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

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

Supplemental Lemma S3. *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 supplemental 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 supplemental 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 $\sim m^{1/2}$.

Supplemental Lemma S4 (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, $N_S \leq n^{2-C} + \sqrt{8C}\log(n)m^{1/2}n^{1-C/2}$. Since $\zeta = \frac{1}{6g(n)}$,

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

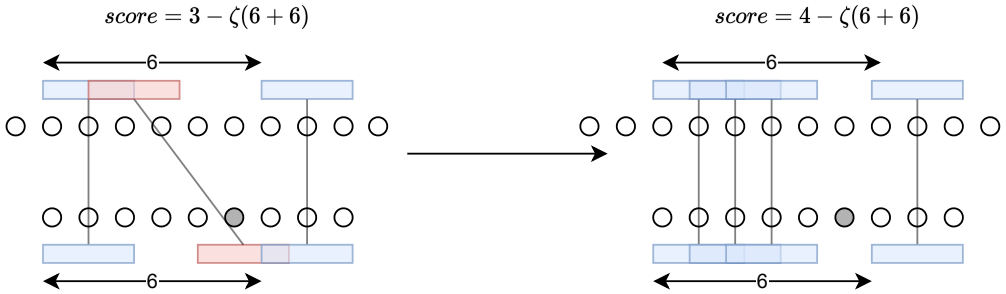
holds and also

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

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.



Supplemental Figure S3: 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 Supplemental Lemma S5 to show that for large enough breaks, such a procedure always improves the chaining score.

Supplemental Lemma S5 (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 $((i_r, j_r), \ldots, (i_t, j_t))$

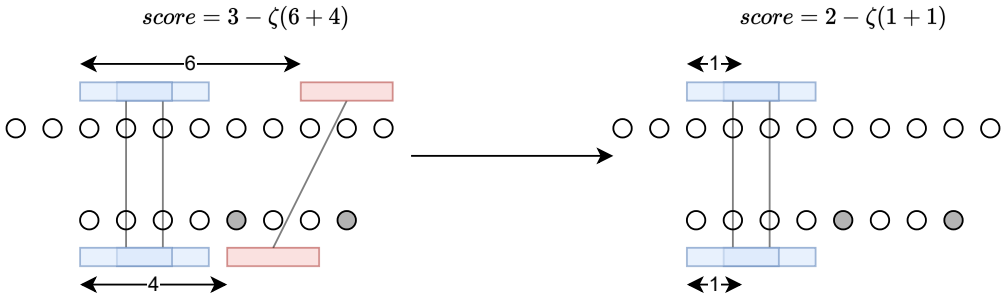
within our chain that lie in this break, and then adding all homologous anchors that are present within the break to the chain. See Supplemental Figure S3.

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.

Supplemental Lemma S6 (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 $< m^{1/2}$.*



Supplemental Figure S4: 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 Supplemental Lemma S6 to show that large breaks can not appear near the ends of chains.

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 Supplemental Figure S4.

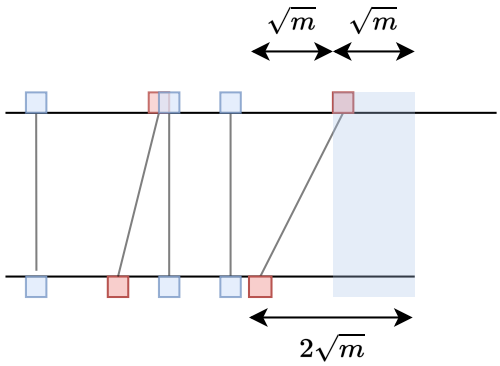
The score of the old chain is $A - \zeta[(i_u - i_{r-1}) + (j_u - j_{r-1})] + w$ where A is the score considering subchain of the anchors prior to r, and w is the number of anchors in the break. This is upper bounded by $A - \zeta(B-k) + N_S$ where $N_S = \Theta(\log(n) m^{1/2} n^{1-C/2})$ as before 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

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

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 Supplemental Lemma S4, 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 supplemental 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



Supplemental Figure S5: 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 Supplemental Lemma S7 and argue that adding in every possible homologous k-mer in the shaded region gives a more optimal score.

Supplemental Lemma S7. *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 Supplemental 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 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 Supplemental Figure S5 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 $\mathcal{F} = \text{F1} \cap \text{F2}$, 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 Supplemental Lemma S7 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.

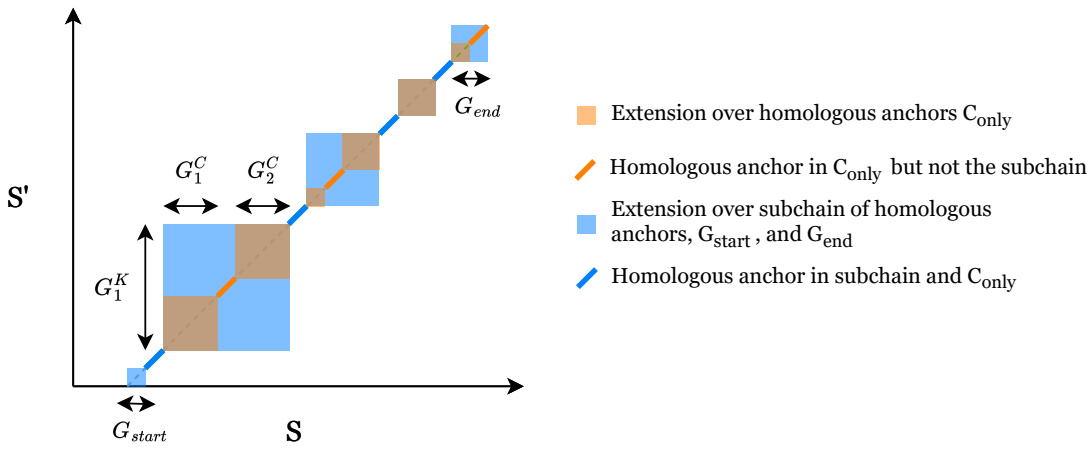
Supplemental Lemma S8. *For a given instance of S, S', let $\mathcal{C}$ be any optimal chain. Let $\mathcal{C}_{only}$ be the chain consisting of only homologous anchors, in S, S'. Defining T^H_{Ext} to be the runtime of extension of $\mathcal{C}$ over only the homologous gaps in $\mathcal{C}$ and $T_{Ext}(\mathcal{C}_{only})$ 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.



Supplemental Figure S6: A graphical proof of Supplemental Lemma S9. 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^2_{end} and G^2_{start} 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.

Supplemental Lemma S9. *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*

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

Proof. See Supplemental Figure S6 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 Supplemental Lemmas S9 and S8, to bound $\mathbb{E}[T^H_{Ext}]$ we can bound $\mathbb{E}[G^2_{start}] + \mathbb{E}[G^2_{end}] + \sum_{i=p+1}^{p+m} \mathbb{E}[(Y_i^K)^2]$. 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

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

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^2_{end}, 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

Supplemental Lemma S10 (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^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. Defining γ as the maximum non-homologous gap size as in Supplemental Lemma S10 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, $\mathbb{E}[N^{*2}_H] \leq \frac{1}{c}2mk(1-\theta)^k + \frac{1}{c^2}m^2(1-\theta)^{2k}$ 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

Supplemental Lemma S11. *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 Supplemental Lemma S13, 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

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

Supplemental Lemma S12. *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

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

Where we substituted in $|\mathcal{J}| = \frac{(1-\beta)\ell}{c}$ in the above equation and used $\frac{c}{c+1} \leq 1$ to remove the term. To bound the second term, we just use Supplemental Lemma S11. This furnishes the final result

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

Supplemental Lemma S13. *No homologous gap has size greater than*

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

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

Proof. We use Supplemental Lemma S12 after plugging in the value of $\ell = g'(n)$ and letting $\beta = 1/2$. Algebraic manipulations show that the probability is less than $\leq 2/n^2$ 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 supplemental 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 Supplemental Lemmas S8 and S9 hold by the same arguments. The analogous version of Lemma 8 holds as well.

Supplemental Lemma S14. *Let T^{H*}_{Ext} be the time of sketched extension over only the homologous gaps of an optimal chain. $\mathbb{E}[T^{H*}_{Ext}] = O(c \cdot mn^{C\alpha} \log n)$.*

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

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

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

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

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.

Supplemental Lemma S15. *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* $\mathbb{E}[T^{H*}_{Ext}] = O\Big(\min\big(\frac{1}{c^2}mn^{C\alpha}\log^3(n), c\cdot mn^{C\alpha}\log n\big)\Big)$. *If* $c = \Theta(\log n)$ *then* $\mathbb{E}[T^{H*}_{Ext}] = O(mn^{C\alpha}\log n)$.

Proof. By Supplemental Lemma S12 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 Supplemental Lemma S12 is dominated by the second term and using $k > c$, where we also ignored the $-\frac{c+1}{2}+1$ in Supplemental Lemma S12 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

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

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

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

The other bound $c\cdot mn^{C\alpha}\log n$ follows from Supplemental Lemma S14, 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 $\mathbb{E}[T^H_{Ext}]$ 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, Supplemental Lemma S12 would not hold, and therefore the above argument would fail. Supplemental Lemma S14 would however still hold.

E.5 Re-proving sketched bounds

Supplemental Lemma S16 (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 $1 - 1/n$.

Supplemental Lemma S17. *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.

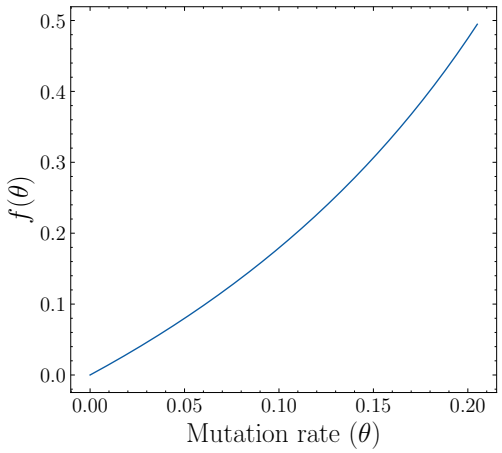
Supplemental Lemma S18. *Under the assumptions of Lemma 7, the expected value of runtime of extension through non-homologous gaps is* $\mathbb{E}[T_{Ext}^{S*}] = O(m)$.

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.

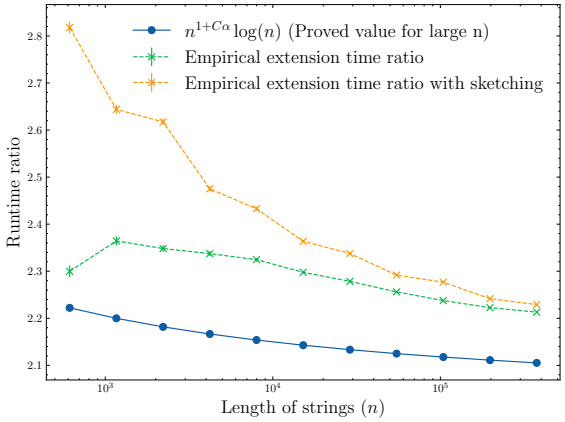
Theorem 2 *In addition to the hypotheses outlined in Theorem 1, let* $c = O(\log n) < k$ *and* $\zeta = \frac{1}{6g'(n)}$ *instead where* $g'(n) = C\frac{200}{8}\log(n)\ln(n)(\frac{3}{2}n^{C\alpha} + 2)$. *For open syncmer sketched seed-chain-extend, the expected running time is* $O\big(\min(\frac{1}{c^2}mn^{C\alpha}\log^3(n), c \cdot mn^{C\alpha}\log n)\big)$ *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 Supplemental Lemma S17. Lemma 13 and Supplemental Lemma S18 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.

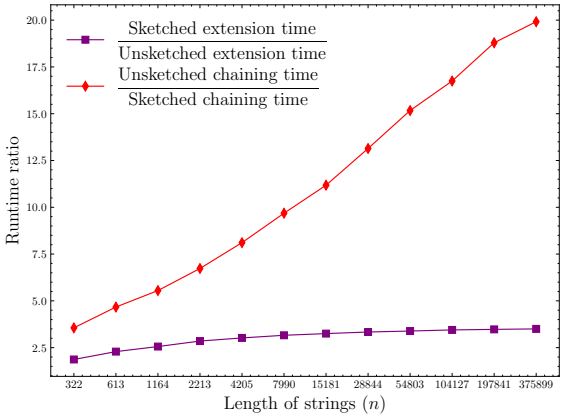
F Additional figures and tables



Supplemental Figure S7: 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 $f(\theta) < \frac{0.5}{0.206}\theta < 2.43 \cdot \theta$.

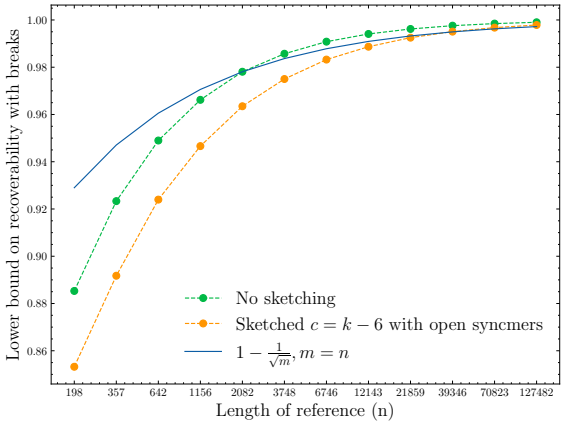


(a) Empirical extension runtime ratios.

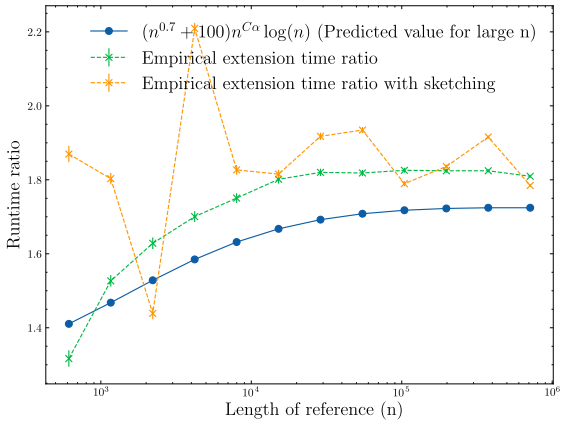


(b) Multiplicative speed up of chaining vs slow-down of extension.

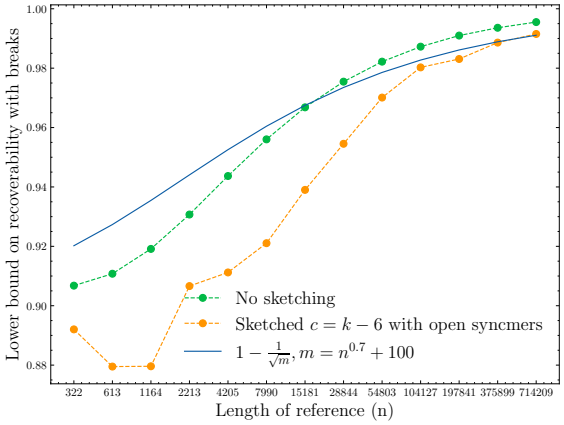
Supplemental Figure S8: The same experiment as in Figure 1 (a) but with $\theta = 0.05$.



Supplemental Figure S9: The recoverability lower bound (Lemma 1, calculated using breaks) 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 $m = n^{0.7} + 100$.

Supplemental Figure S10: Empirical runtimes and recoverability lower bound (using breaks) 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

Supplemental Table S1: 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.

Bibliography

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

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.

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