

Estimators for Substitution Rates in Genomes from Read Data

Shiv Pratap Singh Rathore and Navin Kashyap
Department of Electrical Communication Engineering
Indian Institute of Science, Bengaluru
Email: {shivprataps,nkashyap}@iisc.ac.in

Intro: what is a genome? what is a nucleotide? what is a kmer? substrings of nucleotides of length k

Abstract—We study the problem of estimating the mutation rate between two sequences from noisy sequencing reads. Existing alignment-free methods typically assume direct access to the full sequences. We extend these methods to the sequencing framework, where only noisy reads from the sequences are observed. We use a simple model in which both mutations and sequencing errors are substitutions. We propose multiple estimators, provide theoretical guarantees for one of them, and evaluate the others through simulations.

underlying sequence, possibly corrupted by sequencing errors that occur during the physical process of acquiring the read; however, the location from which a given read originates is unknown.

The rest of the paper is organized as follows. Section II contains some definitions and notations used throughout the paper; it also introduces the problem statement. Section III presents, for benchmarking purposes, estimators for the scenario when there is direct access to k -mer data from the genomic sequences. In Section IV, we derive estimators for the situation where only noisy sequencing reads are available. Section V provides some performance analysis, while Section VI shows simulation results. The paper concludes in Section VII with a summary and discussion of possible future research directions.

arXiv:2601.07546v1 [cs.IT] 12 Jan 2026

genetic differences within and among populations

comparison of the genome sequence of different species (human, mouse,...)

I. INTRODUCTION
Estimating mutation rates in genomes is a fundamental task in genomics, with applications in comparative genomics, phylogenetics, population genetics, evolutionary biology, and disease genomics, including understanding the dynamics of cancer mutations and viral evolution. Traditional alignment-based methods for estimating mutation rates can be both computationally intensive and prone to inaccuracies, especially in highly repetitive or structurally complex regions of the genome. To address these limitations, alignment-free approaches based on k -mer statistics (a k -mer being a substring of a fixed length k) have gained attention in recent years. In this study, we focus on a simplified problem: estimating mutation rates between two genomes under the assumption that the only mutations are substitution errors following an independent and identically distributed (i.i.d.) model.

II. NOTATION AND PRELIMINARIES

Genomic sequences¹ are composed of nucleotides, represented by the four letters A, C, G, T. Let x be a circular² genomic sequence of length G . Fixing an arbitrary starting position, we use x_i to denote the nucleotide at the i -th position of x , i.e., $x = \{x_1, \dots, x_G\}$. We use $x[i : j]$ to denote the $(j - i + 1)$ -length substring $x[i : j] := \{x_i, x_{i+1}, \dots, x_j\}$. Let k be the parameter indicating k -mer size where a k -mer is a substring of length k , i.e., $x[i : i + k - 1], 1 \leq i \leq G$. Thus, sequence x contains exactly G (possibly repeating) k -mers since x is circular. (For a non-circular string, the number of k -mers would be $G - k + 1$.) Let \mathcal{K} denote the set of all distinct k -mers in x . For a k -mer $v \in \mathcal{K}$, we use f_v to denote the number of occurrences of k -mer v in x . For any two k -mers $v, w \in \mathcal{K}$, we use $d_H(v, w)$ to denote Hamming distance between the two k -mers. The abundance histogram of x is the sequence (a_1, \dots, a_G) where a_i is the number of k -mers in \mathcal{K} that occur i times in x , i.e., $a_i = |\{v \in \mathcal{K} : f_v = i\}|$.

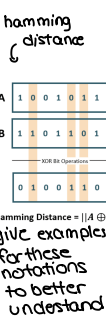
Let y be the string (also circular) obtained by passing the circular string x through a random substitution process with a rate parameter p : independently, the character at each position in x is unchanged with probability $1 - p$ and changed to one of three other nucleotides with probability $p/3$ per nucleotide.

Hypotheses:
¹In this paper, the terms *string* and *sequence* are used interchangeably. We distinguish between *substrings*, which are contiguous, and *subsequences*, which may be non-contiguous.

²The choice of a *circular* genome is purely for convenience, so that we avoid having to deal with edge effects. In any case, for long strings, the edge effects will be negligible.

This problem has previously been studied under the assumption that the genome contains no repeats [1], [2]. This assumption was later relaxed in [3], which allows repeated k -mers but assumes that whenever a k -mer mutates, it does not mutate to any other k -mer present in the string. The work in [1] estimates the substitution rate using the Jaccard similarity between k -mer sketches of two strings, while [2] and [3] adopt a method-of-moments approach based on the number of distinct common k -mers between the source and mutated strings. The former does not account for k -mer multiplicities and therefore performs poorly for highly repetitive sequences, whereas the latter remains empirically effective even in the presence of high repeat content. A common limitation of all these approaches is that they require access to the abundance histogram of the source string. In contrast, we extend these ideas and introduce new techniques for the more realistic case in which only unassembled sequencing reads are available, without access to the underlying sequence or its abundance histogram. Here, a sequencing read is a substring of the

Jaccard Similarity: $J(A, B) = |A \cap B| / |A \cup B|$



Hamming Distance = ||A ⊕ B|| give examples for these notations to better understand

We use f'_v to denote the number of occurrences of k -mer v in y , and \mathcal{K}' to denote set of all distinct k -mers in y .

A noiseless read of x is a random substring of x . We make the simplifying assumption that reads are of a fixed length L , and the starting position of a read is independently and uniformly distributed, i.e., $\Pr(\text{a read starts at } i\text{-th position}) = \frac{1}{G} \forall i \in 1, 2, \dots, G$. Reads are assumed to be acquired through a noisy physical sequencing process, with sequencing errors modeled as i.i.d. substitutions. Specifically, a noisy read of x is obtained by passing a random substring $x[i : i+L-1]$ through the random substitution process described above, but this time parameterized by the sequencing error rate s . Multiple independent noisy reads of x are obtained in this manner. If N is the number of reads so obtained, the sequencing coverage of the sequence x is defined to be $c := NL/G$; this is the expected number of times a given position in x is covered by the N acquired reads.

A. Problem Statement

Throughout, the source string x is assumed to be arbitrary but fixed. The mutated string y obtained through the substitution process is random. The problem we consider is the following: *Given N independent noisy reads acquired from each of the sequences x and y (so $2N$ reads in all), estimate the rate p of the substitution process that generates y from x .* The sequences x and y are not known to the estimator. The sequencing error rate s may be known or unknown, depending on the specific application.

B. Concentration Inequalities

In our analysis of estimator performance, we make use of the following two well-known concentration inequalities.

1) *Hoeffding's Inequality* [4]: Let X_1, \dots, X_n be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely. If $X = \sum_{i=1}^n X_i$, then

$$\Pr(|X - \mathbb{E}[X]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

2) *McDiarmid's Inequality* [5]: Let X_1, \dots, X_n be independent random variables taking values in sets $\mathcal{X}_1, \dots, \mathcal{X}_n$. Suppose $f : \mathcal{X}_1 \times \dots \times \mathcal{X}_n \rightarrow \mathbb{R}$ satisfies the bounded differences condition: for every $i = 1, \dots, n$, $|f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x'_i, \dots, x_n)| \leq c_i$. Then, for all $t > 0$,

$$\Pr(|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

III. ESTIMATORS WITH ACCESS TO k -MER DATA FROM THE TWO STRINGS

In this section and the next, we introduce a number of k -mer-based estimators for p : some use $k = 1$ but others use larger k . The performance of these estimators depends on the underlying source strings x as well as on the mutation rate

p . To benchmark their behavior, we first work directly with k -mer data from the strings x and y . We give the estimators in this section access to the k -mer frequency tables, i.e., the number of occurrences of each k -mer in x and y . In the next section, we extend these estimators to the more realistic case of sequencing data, where only unassembled reads from x and y are observed.

A. $k = 1$

The idea of using $k = 1$ is inspired by the work of Hera et al. [6] who use the number of occurrences of a given nucleotide, say A , in the source string x and the mutated string y as one of the parameters to estimate substitution and indel rates. In our work, we restrict attention to the substitution-only case and use this statistic accordingly, before extending the same idea to any value of k .

For 1-mer A , f_A and f'_A represents the number of times A occurs in x and y respectively. Now, $f'_A = \sum_{i=1}^G Y_i$ where Y_i 's are indicator random variables such that $Y_i = 1$ iff $y_i = A$. Now, if $x_i = A$, then $\Pr(y_i = A) = 1 - p$, otherwise $\Pr(y_i = A) = p/3$. Thus, $\mathbb{E}[f'_A] = \sum_{i:x_i=A} (1-p) + \sum_{i:x_i \neq A} \frac{p}{3} = f_A(1-p) + (G - f_A)\frac{p}{3}$. Replacing $\mathbb{E}[f'_A]$ with the observed value f'_A , we get

$$\hat{p} = \frac{3 \cdot (f'_A - f_A)}{G - 4f_A} \quad (1)$$

We provide a performance analysis of this estimator in Section V-A.

We can use a subset of 1-mers as well, let's say, we take G and C 1-mers. Then, GC fraction can be used for estimation. Note that GC fraction is given as $\frac{f_G + f_C}{G}$. Let x_{GC} and y_{GC} represent GC fraction in x and y respectively.

Now, $\mathbb{E}[f'_G + f'_C] = (f_G + f_C)(1 - \frac{2p}{3}) + (G - f_G - f_C)\frac{2p}{3} \implies \mathbb{E}[y_{GC}] = x_{GC}(1 - \frac{2p}{3}) + (1 - x_{GC})\frac{2p}{3}$. Thus,³

$$\hat{p} = \frac{3 \cdot (y_{GC} - x_{GC})}{2 - 4x_{GC}} \quad (2)$$

This idea can be extended to general values of k as well.

B. $k \geq 1$

Observe that for a k -mer v , we have $\mathbb{E}[f'_v] = \sum_{w \in \mathcal{K}} f_w (1 - p)^{k - d_H(v,w)} (p/3)^{d_H(v,w)}$. The idea is to consider a subset $\mathcal{S} \subset \mathcal{K}$ such that $\sum_{v \in \mathcal{S}} f_v$ is large enough that $\sum_{v \in \mathcal{S}} f'_v$ can be confidently assumed to be close to its expected value, $\sum_{v \in \mathcal{S}} \mathbb{E}[f'_v] = \sum_{v \in \mathcal{S}} \sum_{w \in \mathcal{K}} f_w (1 - p)^{k - d_H(v,w)} (p/3)^{d_H(v,w)}$. Then, the estimate \hat{p} can be obtained numerically as the solution of

$$\sum_{v \in \mathcal{S}} f'_v = \sum_{v \in \mathcal{S}} \sum_{w \in \mathcal{K}} f_w (1 - \hat{p})^{k - d_H(v,w)} \left(\frac{\hat{p}}{3}\right)^{d_H(v,w)} \quad (3)$$

³For now, to keep notation simple, we use \hat{p} to denote all our estimators; the specific estimator in use is determined by the context.

C. Large k : $k \geq 20$

For large values of k , $|\mathcal{K}|$ is quite large and $|\mathcal{K}| \cdot |\mathcal{S}|$ pairwise Hamming distances need to be computed, which is computationally expensive, and moreover, it makes the estimator quite complex. To address this issue, we adopt the simplifying assumption used in [3], namely that if a k -mer v mutates, it doesn't mutate to any other k -mer $w \in \mathcal{K}$ with $w \neq v$.

Under this assumption, $\mathbb{E}[f'_v] = f_v(1-p)^k \implies \sum_{v \in \mathcal{S}} \mathbb{E}[f'_v] = \sum_{v \in \mathcal{S}} f_v(1-p)^k = (1-p)^k \sum_{v \in \mathcal{S}} f_v$. If we take $\mathcal{S} = \mathcal{K}$, then $\sum_{v \in \mathcal{K}} \mathbb{E}[f'_v] = G(1-p)^k$. We thus obtain the estimator

$$\hat{p} = 1 - \left(\frac{\sum_{v \in \mathcal{K}} f'_v}{G} \right)^{1/k}. \quad (4)$$

Note that our estimator is similar to the estimators designed for this problem in [2], [3], the difference in our approach being that we account for multiplicities of k -mers in both x and y .

Next, we allow our estimators to access only the noisy reads observed from sequencing data.

IV. ESTIMATORS WITH ACCESS ONLY TO NOISY SEQUENCING READS

Now, the goal is to estimate the mutation rate p given N noisy reads each for x and y of fixed length L .

For a k -mer v , we use h_v and h'_v to denote number of occurrences of k -mer v across all the N reads of x and y respectively.

Now, consider the reads of string x . Then, $\Pr(\text{positions } x[i:i+k-1] \text{ are covered by a particular read}) = \frac{L-k+1}{G} \forall i \in \{1, \dots, G\}$. Let r_i and r'_i denote number of times positions $x[i:i+k-1]$ and $y[i:i+k-1]$ are covered across all the reads of x and y respectively. Then, $\mathbb{E}[r_i] = \mathbb{E}[r'_i] = \frac{N(L-k+1)}{G}$.

Thus, for any k -mer v ,

$$\begin{aligned} \mathbb{E}[h_v] &= \sum_{i=1}^G \mathbb{E}[r_i] (1-s)^{k-d_H(v, x[i:i+k-1])} \left(\frac{s}{3} \right)^{d_H(v, x[i:i+k-1])} \\ &= \sum_{w \in \mathcal{K}} f_w \left(\frac{N(L-k+1)}{G} \right) (1-s)^{k-d_H(v, w)} \left(\frac{s}{3} \right)^{d_H(v, w)} \end{aligned} \quad (5)$$

A. $k = 1$

For $k = 1$, we use fraction of A's in reads of x and y , i.e., $\frac{h_A}{NL}$ and $\frac{h'_A}{NL}$ to estimate p . From (5), we get

$$\begin{aligned} \mathbb{E}[h_A] &= \frac{NL}{G} \left(f_A(1-s) + (G-f_A)\frac{s}{3} \right) \\ &= \frac{NL}{G} \left(f_A \left(1 - \frac{4s}{3} \right) + G\frac{s}{3} \right) \end{aligned}$$

If f'_A is known, then analogously we get

$$\mathbb{E}[h'_A | f'_A] = \frac{NL}{G} \left(f'_A \left(1 - \frac{4s}{3} \right) + G\frac{s}{3} \right)$$

Then,

$$\begin{aligned} \mathbb{E}[h'_A] &= \mathbb{E}[\mathbb{E}[h'_A | f'_A]] \\ &= \frac{NL}{G} \left(\mathbb{E}[f'_A] \left(1 - \frac{4s}{3} \right) + G\frac{s}{3} \right) \\ &= \frac{NL}{G} \left(\left(f_A \left(1 - \frac{4p}{3} \right) + G\frac{p}{3} \right) \left(1 - \frac{4s}{3} \right) + G\frac{s}{3} \right) \\ &= \frac{NL}{G} \left(\left(f_A \left(1 - \frac{4s}{3} \right) + G\frac{s}{3} \right) \left(1 - \frac{4p}{3} \right) + G\frac{p}{3} \right) \\ &= \mathbb{E}[h_A] \left(1 - \frac{4p}{3} \right) + NL\frac{p}{3} \end{aligned}$$

Thus,

$$\hat{p} = \frac{3(h'_A - h_A)}{NL - 4h_A} \quad (6)$$

Note that our estimator is independent of s . An analysis of the performance of this estimator is provided in Section V-B.

B. Large k : $k \geq 20$

In this case, we assume the following:

- (1) If a k -mer v mutates, it doesn't mutate to any other k -mer $w \in \mathcal{K}$, with $w \neq v$.
- (2) Due to sequencing errors, a k -mer v in the string x (or y) doesn't change to any other k -mer $w \in \mathcal{K}$ (or \mathcal{K}'), with $w \neq v$, in the reads of x (or y).

Then, from (5), we obtain $\mathbb{E}[h_v] = f_v \cdot \left(\frac{N(L-k+1)}{G} \right) (1-s)^k$. Analogously, we get

$\mathbb{E}[h'_v | f'_v] = f'_v \cdot \left(\frac{N(L-k+1)}{G} \right) (1-s)^k$. Using similar arguments and calculations as before, we get $\mathbb{E}[h'_v] = \mathbb{E}[h_v](1-p)^k$. For a subset $\mathcal{S} \subseteq \mathcal{K}$, $\sum_{v \in \mathcal{S}} \mathbb{E}[h'_v] = (1-p)^k \sum_{v \in \mathcal{S}} \mathbb{E}[h_v]$. Thus,

$$\hat{p} = 1 - \left(\frac{\sum_{v \in \mathcal{S}} h'_v}{\sum_{v \in \mathcal{S}} h_v} \right)^{1/k} \quad (7)$$

Ideally, we would like to take $\mathcal{S} = \mathcal{K}$. However, in practice, some k -mers present in the string may not appear in the reads at all, while additional k -mers may be introduced due to sequencing errors. Let \mathcal{R} and \mathcal{R}' denote the sets of all k -mers observed in the reads of x and y , respectively. We define $\mathcal{F} = \mathcal{R} \setminus \mathcal{K}$ and $\mathcal{F}' = \mathcal{R}' \setminus \mathcal{K}'$ as the sets of spurious k -mers introduced by sequencing errors in the reads of x and y .

Given a k -mer $v \in \mathcal{R}$, it is not possible to determine whether $v \in \mathcal{K}$ or $v \in \mathcal{F}$. Therefore, we use the following estimator:

$$\hat{p} = 1 - \left(\frac{\sum_{v \in \mathcal{R}: h_v \geq \lambda} h'_v}{\sum_{v \in \mathcal{R}: h_v \geq \lambda} h_v} \right)^{1/k} \quad (8)$$

ici j'ai pas très bien compris

introduction de nouvelles notations

loi unif. ?

Observe that

$$\sum_{v \in \mathcal{R}: h_v \geq \lambda} h_v = \sum_{v \in \mathcal{K}: h_v \geq \lambda} h_v + \sum_{v \in \mathcal{F}: h_v \geq \lambda} h_v.$$

We choose the threshold λ such that the total mass $\sum_{v \in \mathcal{R}: h_v \geq \lambda} h_v$ is sufficiently large, while the contribution from spurious k -mers, $\sum_{v \in \mathcal{F}: h_v \geq \lambda} h_v$, is minimized.

Since a k -mer in the reads remains unaffected by sequencing errors with probability $(1-s)^k$, we have $\mathbb{E}[\text{fraction of } k\text{-mers in reads unaffected by sequencing errors}] = (1-s)^k$. Under Assumption (2), this implies

$$\mathbb{E} \left[\frac{\sum_{v \in \mathcal{K}} h_v}{\sum_{v \in \mathcal{R}} h_v} \right] = (1-s)^k.$$

Motivated by this observation, we select

$$\lambda = \max_{\lambda' \geq 2} \left\{ \lambda' : \sum_{v \in \mathcal{R}: h_v \geq \lambda'} h_v \geq (1-s)^k \sum_{v \in \mathcal{R}} h_v \right\}.$$

That is, we retain the most frequently occurring k -mers in the reads whose cumulative frequency accounts for approximately $(1-s)^k$ fraction of all observed k -mers. When the sequencing error rate s is small, true k -mers $v \in \mathcal{K}$ typically appear more frequently in the reads than spurious k -mers $w \in \mathcal{F}$, resulting in only a small number of outliers and a negligible contribution from $\sum_{v \in \mathcal{F}: h_v \geq \lambda} h_v$.

A limitation of this estimator is that selecting the parameter λ requires prior knowledge of the sequencing error rate s . Although this requirement can be relaxed by using a known upper bound on s , doing so leads to degraded performance. In contrast, the estimator for the $k=1$ case remains applicable even when s is unknown. On the other hand, the sequence length G is not required to be known for both the estimators.

V. PERFORMANCE ANALYSIS

We here analyze the performance of the 1-mer-based estimators \hat{p} in (1) and (6) using the concentration inequalities of Hoeffding and McDiarmid to bound the probability (with respect to the random substitution process) that $|\hat{p} - p| \leq \epsilon p$. We have not been successful in extending this style of analysis to the estimators that use large values of k , so we evaluate their performance using simulation results in the next section.

A. A concentration bound for the estimator in (1)

For the estimator \hat{p} given by (1), we have

$$\begin{aligned} \Pr(|\hat{p} - p| \leq \epsilon p) &= \Pr \left(\left| \frac{3 \cdot (f'_A - f_A)}{G - 4f_A} - p \right| \leq \epsilon p \right) \\ &= \Pr \left(\left| f'_A - \mathbb{E}[f'_A] \right| \leq \frac{|G - 4f_A| \cdot p\epsilon}{3} \right) \\ &\geq 1 - 2 \exp \left(- \frac{2|G - 4f_A|^2 \cdot p^2 \epsilon^2}{9G} \right) \\ &= 1 - 2 \exp \left(- \frac{32}{9} G \left| \frac{f_A}{G} - 0.25 \right|^2 p^2 \epsilon^2 \right) \end{aligned}$$

where the inequality above is obtained using Hoeffding's inequality.

Now, for fixed values of ϵ , G and p , we compute the fraction of A's required in the source string x such that $\Pr(|\hat{p} - p| \geq \epsilon p) \leq 2 \exp(-5.3) \leq 10^{-2}$. This requirement translates to $\left| \frac{f_A}{G} - 0.25 \right| \geq \sqrt{\frac{1.5}{p^2 \epsilon^2 G}}$. For this estimator to be effective, $\left| \frac{f_A}{G} - 0.25 \right| > 0$ and the minimum deviation needed to guarantee the desired performance decreases as p and G increase, as illustrated in Table I.

TABLE I: Minimum value of $\left| \frac{f_A}{G} - 0.25 \right|$ required to ensure performance guarantees for $\epsilon = 0.1$

$p \setminus G$	10^4	10^5	10^6	10^7
0.01	12.24	3.87	1.22	0.387
0.03	4.08	1.29	0.408	0.129
0.05	2.44	0.77	0.245	0.077
0.10	1.22	0.387	0.122	0.038
0.20	0.612	0.194	0.061	0.019
0.50	0.245	0.077	0.024	0.008

B. A concentration bound for the estimator in (6)

Theorem 1. Assume $p, s < \frac{3}{4}$ and

$$\left| \frac{f_A}{G} - 0.25 \right| \geq \frac{3}{4p\epsilon} \left(\sqrt{\frac{C_2}{2N}} + \sqrt{\frac{C_1}{2G}} \right) + \sqrt{\frac{C_3}{2N}} + \left| \frac{4s}{3} \frac{f_A}{G} - \frac{s}{3} \right|$$

for some $C_1, C_2, C_3 > 0$. Then, for the estimator \hat{p} given by (6), we have

$$\Pr(|\hat{p} - p| \leq \epsilon p) \geq 1 - 2e^{-C_1} - 2e^{-C_2} - 2e^{-C_3}.$$

Proof.

$$\begin{aligned} \Pr(|\hat{p} - p| \leq \epsilon p) &= \Pr \left(\left| \frac{3 \cdot (h'_A - h_A)}{NL - 4h_A} - p \right| \leq \epsilon p \right) \\ &= \Pr \left(\left| h'_A - \mathbb{E}[h'_A] \right| \leq \frac{|NL - 4h_A| \cdot p\epsilon}{3} \right) \end{aligned}$$

Define the event $B_1 := \{|f'_A - \mathbb{E}[f'_A]| \leq \delta \cdot \mathbb{E}[f'_A]\}$ for some suitable choice of δ to be specified later. Let $t = |NL - 4h_A| \cdot p\epsilon/3$. We bound $\Pr(|h'_A - \mathbb{E}[h'_A]| > t) = 1 - \Pr(|h'_A - \mathbb{E}[h'_A]| \leq t)$ from above as follows:

$$\begin{aligned} \Pr(|h'_A - \mathbb{E}[h'_A]| > t) &= \Pr(|h'_A - \mathbb{E}[h'_A|f'_A] + \mathbb{E}[h'_A|f'_A] - \mathbb{E}[h'_A]| > t) \\ &\leq \Pr(|h'_A - \mathbb{E}[h'_A|f'_A]| > t_1 | B_1) \\ &\quad + \Pr(|\mathbb{E}[h'_A|f'_A] - \mathbb{E}[h'_A]| > t_2 | B_1) + \Pr(B_1^c) \end{aligned}$$

where $t_1 + t_2 \leq t$.

Now, we will bound these three terms individually. First, using Hoeffding's inequality,

$$\Pr(B_1^c) \leq 2 \exp \left(-2\delta^2 G \left(\frac{f_A}{G} \left(1 - \frac{4p}{3} \right) + \frac{p}{3} \right)^2 \right)$$

To have $\Pr(B_1^c) \leq 2\exp(-C_1)$, it suffices to have $2\delta^2 G \left(\frac{f_A}{G} \left(1 - \frac{4p}{3} \right) + \frac{p}{3} \right)^2 \geq C_1$. Thus, we take

$$\delta := \frac{\sqrt{C_1}}{\sqrt{2G} \left(\frac{f_A}{G} \left(1 - \frac{4p}{3} \right) + \frac{p}{3} \right)}.$$

Note that $\mathbb{E}[h'_A | f'_A] = \frac{NL}{G} \left(f'_A \left(1 - \frac{4s}{3} \right) + G \frac{s}{3} \right)$. Also, $\mathbb{E}[h'_A] = \frac{NL}{G} \left(\mathbb{E}[f'_A] \left(1 - \frac{4s}{3} \right) + G \frac{s}{3} \right)$. Thus, conditioned on the event B_1 , with probability 1,

$$\begin{aligned} |\mathbb{E}[h'_A | f'_A] - \mathbb{E}[h'_A]| &\leq \delta NL \frac{\mathbb{E}[f'_A]}{G} \left(1 - \frac{4s}{3} \right) \\ &= \delta NL \left(\frac{f_A}{G} \left(1 - \frac{4p}{3} \right) + \frac{p}{3} \right) \left(1 - \frac{4s}{3} \right) \\ &= NL \left(1 - \frac{4s}{3} \right) \sqrt{\frac{C_1}{2G}} \\ &\leq NL \sqrt{\frac{C_1}{2G}} := t_2 \end{aligned}$$

Now, we will bound $\Pr(|h'_A - \mathbb{E}[h'_A | f'_A]| \geq t_1 | B_1)$. Let r.v. P_i determines the starting position of i -th read. Let W_i be random substitution process acting on i -th read. The pair $S_i := (P_i, W_i)$ completely specifies the i -th read. Now, S_1, \dots, S_N are independent, and this independence holds even when conditioned on f'_A or on the event B_1 . Let $Z = g(S_1, \dots, S_N)$ denote the total number of occurrences of the symbol A across all N reads. Since each read has fixed length L , changing a single read S_i can change the value of g by atmost L , i.e., for each $i = 1, \dots, N$, $|g(s_1, \dots, s_{i-1}, s_i, s_{i+1}, \dots, s_N)| - |g(s_1, \dots, s_{i-1}, s'_i, s_{i+1}, \dots, s_N)| \leq L$. Thus, using McDiarmid's inequality, we get

$$\begin{aligned} \Pr(|g(S_1, \dots, S_N) - \mathbb{E}[g(S_1, \dots, S_N)]| \geq t) \\ \leq 2 \exp \left(- \frac{2t^2}{NL^2} \right). \end{aligned}$$

Conditioning on the event B_1 and $f'_A = q$, we obtain

$$\Pr(|h'_A - \mathbb{E}[h'_A | f'_A = q]| \geq t_1 | B_1, f'_A = q) \leq 2 \exp \left(- \frac{2t_1^2}{NL^2} \right).$$

Since this bound holds uniformly for all values of q , it follows that

$$\Pr(|h'_A - \mathbb{E}[h'_A | f'_A]| \geq t_1 | B_1) \leq 2 \exp \left(- \frac{2t_1^2}{NL^2} \right).$$

To make $\Pr(|h'_A - \mathbb{E}[h'_A | f'_A]| \geq t_1 | B_1) \leq 2\exp(-C_2)$, we set $t_1 := L\sqrt{NC_2/2}$.

Now, $t_1 + t_2 \leq t = |NL - 4h_A| \cdot p\epsilon/3$ yields the condition

$$L\sqrt{\frac{NC_2}{2}} + NL\sqrt{\frac{C_1}{2G}} \leq \frac{|NL - 4h_A| p\epsilon}{3},$$

which is equivalent to

$$\left| \frac{h_A}{NL} - 0.25 \right| \geq \frac{3}{4p\epsilon} \left(\sqrt{\frac{C_2}{2N}} + \sqrt{\frac{C_1}{2G}} \right) =: u$$

Define the event $B_2 := \{ \left| \frac{h_A}{NL} - \mathbb{E} \left[\frac{h_A}{NL} \right] \right| \leq u_2 \}$. Conditioned on B_2 , a sufficient condition for $\left| \frac{h_A}{NL} - 0.25 \right| \geq u$ is $|\mathbb{E} \left[\frac{h_A}{NL} \right] - 0.25| \geq u + u_2$. Since $\mathbb{E} \left[\frac{h_A}{NL} \right] = \frac{f_A}{G} \left(1 - \frac{4s}{3} \right) + \frac{s}{3}$, it suffices to have $\left| \frac{f_A}{G} - 0.25 \right| \geq u + u_2 + \left| \frac{4s}{3} \frac{f_A}{G} - \frac{s}{3} \right|$. But this, if we set $u_2 := \sqrt{C_3/(2N)}$, is precisely the inequality assumed in the statement of the theorem. With this choice of u_2 , we further have, by McDiarmid's inequality, $\Pr(B_2^c) = \Pr(|h_A - \mathbb{E}[h_A]| > u_2 NL) \leq 2\exp(-2u_2^2 N) = 2\exp(-C_3)$.

Putting it all together using the union bound, we obtain

$$\begin{aligned} \Pr(|\hat{p} - p| \leq \epsilon p) &\geq 1 - \Pr(|h'_A - \mathbb{E}[h'_A | f'_A]| > t_1 | B_1) \\ &\quad - \Pr(|\mathbb{E}[h'_A | f'_A] - \mathbb{E}[h'_A]| > t_2 | B_1) \\ &\quad - \Pr(B_1^c) - \Pr(B_2^c) \\ &\geq 1 - 2e^{-C_2} - 0 - 2e^{-C_1} - 2e^{-C_3} \\ &= 1 - 2e^{-C_1} - 2e^{-C_2} - 2e^{-C_3}. \end{aligned}$$

□

A similar table to Table I can be obtained for different values of G, N, p, s , and ϵ . As an illustration, consider $G = 10^7$, $L = 10^3$, $N = 10^6$, $p = 0.2$, $s = 0.03$, and $\epsilon = 0.1$. For these parameters, ensuring $\Pr(|\hat{p} - p| \leq \epsilon p) \geq 1 - 10^{-3}$ requires $\left| \frac{f_A}{G} - 0.25 \right| \geq 0.112$.

The main bottleneck of this analysis is that achieving $|\hat{p} - p| \leq \epsilon p$ with high probability requires a large number of reads and a small sequencing error rate.

VI. SIMULATIONS AND RESULTS

We evaluate the performance of all our proposed estimators using both **synthetic and real genomic sequences**. Synthetic sequences are generated under an i.i.d. model, where each symbol is drawn independently according to a fixed probability distribution. Real sequences are extracted from the human T2T-CHM13v2.0 reference genome [7].

For real data, we consider three sequences. Two of them, **D-easy** and **D-hardest**, were previously used in [3] and have length 10^5 . In addition, we include a longer sequence of length 10^6 , referred to as **Chr9-HSat**.

For the purposes of this discussion, we refer to the estimators from Section III as “non-sequencing estimators” and to those from Section IV as “sequencing-based estimators”. In the case of non-sequencing estimators, where the complete k -mer frequency tables of both sequences x and y are available, we compare our estimators for $k = 1$ and $k = 30$ with the estimator proposed in [3], which we denote by \hat{p}_{AH} .⁴ When considering sequencing-based estimators, where only reads from x and y are available, we evaluate only our estimators for $k = 1$ and $k = 30$. We use \hat{p}_{NS} to denote our non-sequencing estimators and \hat{p}_S for the sequencing-based estimators. Unless stated otherwise, the sequencing coverage $c = NL/G$ is fixed to 30, with read length $L = 1000$.

We evaluate the performance of all estimators using box plots of the relative error $e = \frac{\hat{p}}{p} - 1$, as a function of the parameters p, s, c , and G .

⁴The subscript AH is meant to indicate that the estimator uses abundance histogram information.

A. Real sequences

For each real sequence and each value of the mutation rate p , we generate 100 mutated sequences. The characteristics of the three sequences used in our experiments are as follows:

introduce them at the start. ↪

- 1) **D-easy:** A length- 10^5 arbitrarily chosen substring from chr6, with no unusual repeat annotations. For $k = 30$, the number of distinct k -mers is 99442 and the maximum multiplicity of any k -mer is 9. For $k = 1$, we use the 1-mer T, for which $f_T/G = 0.301$.
- 2) **D-hardest:** A length- 10^6 highly repetitive subsequence from a region annotated as “Active α Sat HOR” in the chr21 centromere. For $k = 30$, the number of distinct k -mers is 3987 and the maximum multiplicity is 127. For $k = 1$, we use the 1-mer G, with $f_G/G = 0.168$.
- 3) **Chr9-HSsat:** A length- 10^6 repetitive substring from a pericentromeric region of chr9 annotated as satellite DNA (HSAT2/HSAT3). For $k = 30$, the number of distinct k -mers is 58244 and the maximum multiplicity is 1250. For $k = 1$, we use the 1-mer C, with $f_C/G = 0.169$.

Figures 1(a), 1(b), and 1(e) compare the estimators in the non-sequencing case. For $k = 30$, our estimator \hat{p}_{NS} (given by (4)) significantly outperforms \hat{p}_{AH} for the D-hardest and Chr9-HSsat sequences. This improvement arises because \hat{p}_{NS} accounts for k -mer multiplicities in the mutated sequence as well. For D-easy sequence, where most k -mers occur only once, the performance of \hat{p}_{NS} and \hat{p}_{AH} is similar.

We also observe that the estimators using $k = 30$ become unstable at higher mutation rates ($p \geq 0.3$), whereas the estimator that uses $k = 1$, given by (1), performs poorly at low mutation rates but improves as p increases. The Chr9-HSsat sequence provides an example where both the $k = 30$ estimators — \hat{p}_{NS} given by (4) as well as \hat{p}_{AH} — perform poorly even at low mutation rates, while the $k = 1$ estimator \hat{p}_{NS} given by (1) achieves relatively better performance.

Figures 1(c), 1(d), and 1(e) present results for the sequencing-based estimators. For D-easy and D-hardest, trends similar to the non-sequencing case are observed: the $k = 1$ estimator given by (6) performs better at higher mutation rates, whereas the $k = 30$ estimator given by (8) performs better at lower mutation rates. Going from non-sequencing to sequencing-based estimators leads to a slight degradation in performance, but the overall performance remains within acceptable limits.

For the Chr9-HSsat sequence, we only report results for the $k = 1$ sequencing-based estimator \hat{p}_S given by (6); for $k = 30$ even the non-sequencing estimators perform poorly. In this case, \hat{p}_S performs well due to the large sequence length and the significant deviation of f_C/G from 0.25.

Figures 1(f) and 1(g) illustrate the effect of varying sequencing coverage c and sequencing error rate s while fixing $p = 0.05$, for the D-easy and D-hardest sequences. The estimator \hat{p}_S given by (6) is largely insensitive to changes in s , whereas the performance of \hat{p}_S given by (8) degrades as s increases. This degradation affects D-easy more than D-hardest since most of the k -mers in D-easy occur only once.

Therefore, as s increases, most of the k -mers don’t appear in the reads, allowing spurious k -mers generated by sequencing errors to dominate.

B. Synthetic sequences

For randomly generated synthetic sequences, we fix $p = s = 0.05$ and vary f_A/G , c , and G . We consider sequence lengths $G = 10^5$ and 10^6 . For each value of f_A/G , we generate 20 independent reference sequences and simulate 5 mutated sequences from each reference.

Figure 2(a) shows the performance of the $k = 1$ sequencing-based estimator \hat{p}_S given by (6) as a function of f_A/G , c , and G . The estimator accuracy improves as $|f_A/G - 0.25|$ increases. Increasing the sequence length G leads to a substantial improvement in performance, while increasing the coverage c only slightly improves the performance. The estimator becomes unstable when $|f_A/G - 0.25|$ is small.

Fig. 2(b) evaluates the performance of \hat{p}_S given by (8). In contrast to the $k = 1$ estimator, this estimator performs well even when $f_A/G = 0.25$. Here, performance improves slightly with increasing G and more significantly with increasing c .

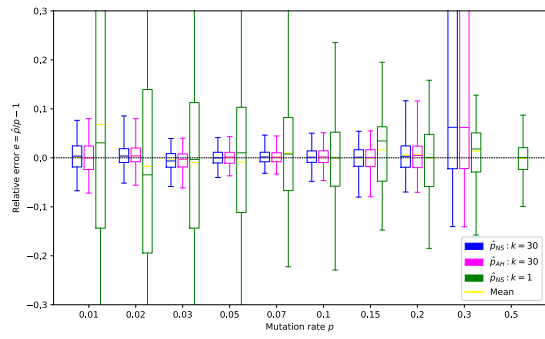
VII. CONCLUSION AND FUTURE WORK

In this work, we extend estimation of substitution rate between two sequences to the practically relevant setting in which only unassembled sequencing reads are available, and the underlying sequences are unknown. Our approach relies on k -mer statistics, and we propose two kinds of estimators: one based on $k = 1$ and another based on large k , for which we use $k = 30$ in practice. We also derived theoretical performance guarantees for the $k = 1$ estimator.

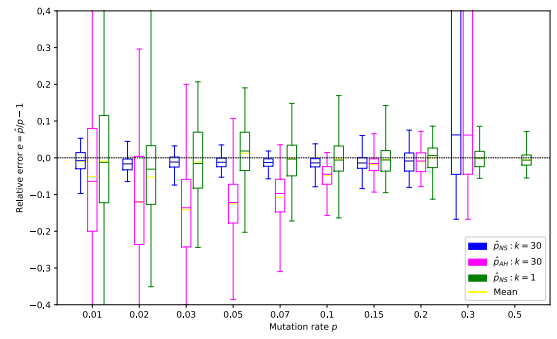
Both the estimators have their pros and cons. The $k = 1$ estimator is unbiased and performs well whenever $|f_v/G - 0.25| > 0$ for at least one $v \in \{A, G, C, T\}$. Its accuracy improves with increasing sequence length G and higher mutation rates p , and it is relatively robust to sequencing errors. Moreover, the $k = 1$ estimator has substantially lower time and space complexity and scales better with sequence length. In contrast, the estimator with $k = 30$ is slightly biased, with a bias that depends on the underlying sequence. Nevertheless, it performs better in practice for less repetitive sequences and in regimes where the probability that a k -mer mutates to another k -mer in the sequence is low, particularly at small mutation rates. However, this estimator is more sensitive to sequencing errors.

A direction for future work is to establish theoretical performance guarantees for the large- k estimator. In addition, there exist regimes and sequences for which both estimators become unstable or fail to provide accurate estimates. This motivates the open problem of designing an estimator that performs well across different regimes and sequence types, or at least for most real genomic sequences. Another possible future direction is to extend this framework to account for other types of mutations and sequencing errors such as insertions and deletions.

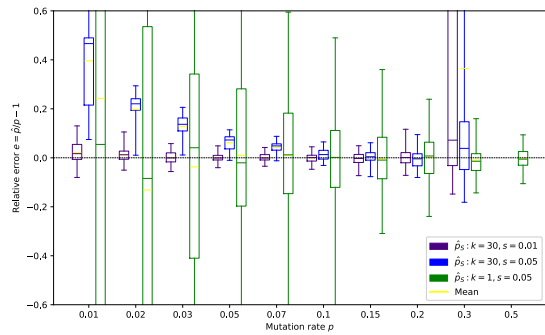
Acknowledgements. The authors would like to acknowledge useful discussions with Daanish Mahajan, Chirag Jain, and Paul Medvedev.



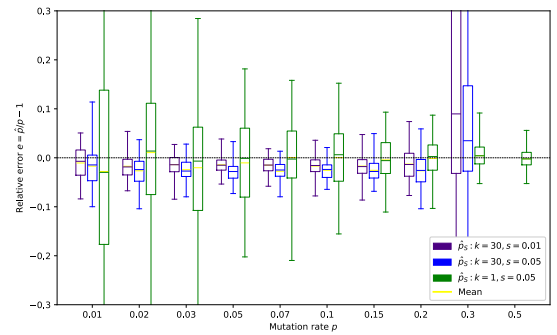
(a) D-easy non-sequencing



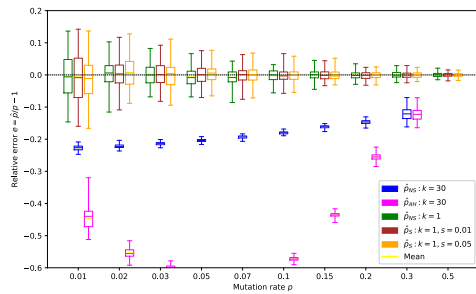
(b) D-hardest non-sequencing



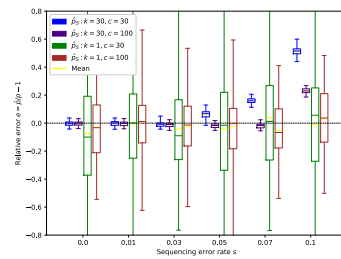
(c) D-easy sequencing



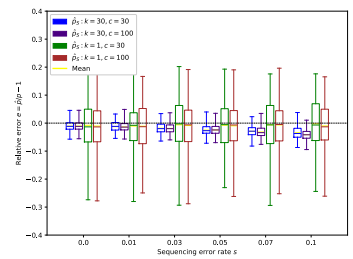
(d) D-hardest sequencing



(e) Chr9-HSAt

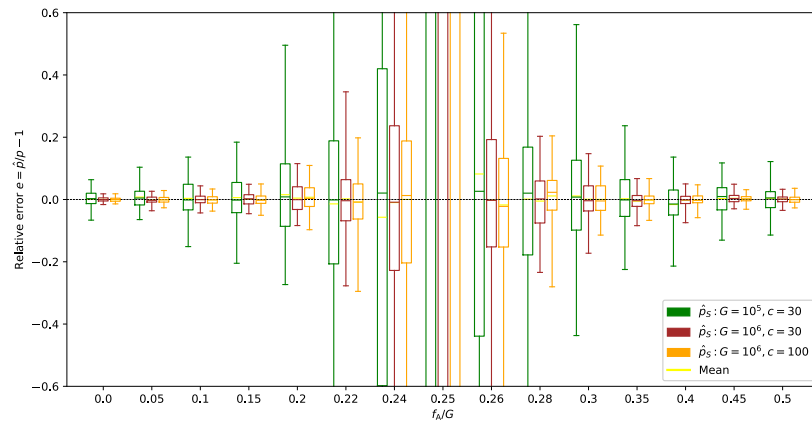


(f) D-easy: Fixed $p = 0.05$, varying s and c

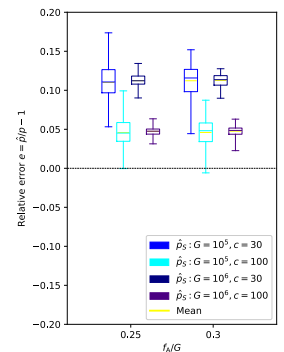


(g) D-hardest: Fixed $p = 0.05$, varying s and c

Fig. 1: Real genomic sequences



(a) $k = 1$



(b) $k = 30$

Fig. 2: Randomly generated synthetic sequences

REFERENCES

- [1] B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy, "Mash: Fast genome and metagenome distance estimation using MinHash," *Genome Biology*, vol. 17, no. 1, p. 132, 2016.
- [2] A. Blanca, R. S. Harris, D. Koslicki, and P. Medvedev, "The statistics of k-mers from a sequence undergoing a simple mutation process without spurious matches," *Journal of Computational Biology*, vol. 29, no. 2, pp. 155–168, 2022.
- [3] H. Wu, A. Blanca, and P. Medvedev, "A k-mer-based estimator of the substitution rate between repetitive sequences," 2025. bioRxiv e-print 2025.06.19.660607.
- [4] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.
- [5] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference* (J. Siemons, ed.), London Mathematical Society Lecture Note Series, pp. 148–188, Cambridge University Press, 1989.
- [6] M. R. Hera, P. Medvedev, D. Koslicki, and A. Blanca, "Estimation of substitution and indel rates via k-mer statistics," 2025. bioRxiv e-print 2025.05.14.653858.
- [7] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, et al., "The complete sequence of a human genome," *Science*, vol. 376, no. 6588, pp. 44–53, 2022.