Estimation of substitution and indel rates via k-mer statistics

Mahmudur Rahman Hera
School of Electrical Engineering and Computer Science, Pennsylvania State University, USA

Paul Medvedev
Department of Computer Science and Engineering, The Pennsylvania State University, University Park, USA
Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, USA
Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, USA

David Koslicki
School of Electrical Engineering and Computer Science, Pennsylvania State University, USA
Department of Biology, Pennsylvania State University, USA
Huck Institutes of the Life Sciences, Pennsylvania State University, USA

Antonio Blanca
School of Electrical Engineering and Computer Science, Pennsylvania State University, USA

Abstract

Methods utilizing k-mers are widely used in bioinformatics, yet our understanding of their statistical properties under realistic mutation models remains incomplete. Previously, substitution-only mutation models have been considered to derive precise expectations and variances for mutated k-mers and intervals of mutated and non-mutated sequences. In this work, we consider a mutation model that incorporates insertions and deletions in addition to single-nucleotide substitutions. Within this framework, we derive closed-form k-mer-based estimators for the three fundamental mutation parameters: substitution, deletion rate, and insertion rates. We provide theoretical guarantees in the form of concentration inequalities, ensuring accuracy of our estimators under reasonable model assumptions. Empirical evaluations on simulated evolution of genomic sequences confirm our theoretical findings, demonstrating that accounting for insertions and deletions signals allows for accurate estimation of mutation rates and improves upon the results obtained by considering a substitution-only model. An implementation of estimating the mutation parameters from a pair of fasta files is available here: github.com/KoslickiLab/estimate_rates_using_mutation_model.git. The results presented in this manuscript can be reproduced using the code available here: github.com/KoslickiLab/est_rates_experiments.git.

2012 ACM Subject Classification Applied computing → Computational biology; Theory of computation → Theory and algorithms for application domains; Mathematics of computing → Probabilistic inference problems

Keywords and phrases k-mers, mutation rate, indel, alignment-free, estimation, substitution, insertion, deletion

Related Version A preliminary version of this paper appeared in the proceedings of WABI 2025 [10].

Funding This material is based upon work supported by the National Science Foundation under Grant No. DBI2138585. Research reported in this publication was supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM146462. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1 Introduction

Estimating the mutation rate between two evolutionarily related sequences is a classical question in molecular evolution, with roots that pre-date the genomics era [25]. Early quantitative efforts focused on amino-acid substitution: the seminal PAM matrices of Dayhoff et al. converted curated alignments of close homologues into an evolutionary time-scale [5], while the BLOSUM series by Henikoff and Henikoff mined ungapped blocks of conserved proteins to improve sensitivity for more diverged sequences [8]. These approaches and later profile-based HMM models [6] were derived from pairwise or multiple alignments and remain the gold standard when accurate alignments are available.

Over the last decade, however, high-throughput sequencing has shifted the scale of comparative genomics from dozens to millions of genomes, rendering high computational cost pipelines (e.g. quadratic-time) increasingly impractical. Consequently, alignment-free techniques that summarize sequences by inexpensive statistics have become indispensable [21, 26]. These approaches most commonly utilize k-mer sets and sketches thereof. Popular tools such as Mash [14], Skmer [17] and more recent sketch-corrected frameworks like Sylph [20] and FracMinHash-based methods [9, 11, 15, 19] can build whole-genome phylogenies, screen metagenomic samples, and compute millions of pairwise point-mutation rate estimates in minutes rather than days.

Despite their empirical success, theoretical understanding of alignment-free estimators has lagged behind practice. Nearly all existing models treat evolution as a pure-substitution process, ignoring insertions and deletions (indels), or else their performance in the presence of indels is often not thoroughly evaluated [16]. When indels are frequent, substitution-only estimators systematically inflate divergence and can misplace taxa—even in otherwise well-resolved trees of primates constructed from k-mer Jaccard similarities [14]. Recent work has quantified how k-mer-based statistics are affected by substitutions and are also used to estimate substitution-only mutation rates [11], yet a principled treatment that jointly infers substitution and indel parameters from k-mer statistics is still absent. This omission is particularly significant because indels represent a substantial fraction of genomic variation and play crucial roles in evolution [23]. Such indels cause substitution-only k-mer-based methods to underperform, as just like with substitutions, disruption of k-mer content by indel events affects at least k k-mers, often leading to overestimates of mutation rates [3, 11].

In this paper we introduce the first closed-form, alignment-free estimators for the three fundamental mutation parameters: substitution rate p_s, deletion rate p_d, and mean insertion length g under a model that explicitly incorporates single-nucleotide substitutions, deletions, and geometrically-distributed insertions. Starting from elementary counts of non-mutated and single-deletion k-mers, we derive algebraic expressions for p_s, p_d, and g and prove their accuracy via sub-exponential concentration bounds (Theorem 2). Simulations on synthetic and real bacterial genomes demonstrate that modeling indels yields markedly more accurate distance estimates than substitution-only approaches. The remainder of this paper is organized as follows: Section 2 gives preliminary definitions and defines our mutation model, Section 3 presents our main results, including the estimators and their theoretical guarantees, Section 4 provides a proof outline of the main theorem, Section 5 details our implementation, Section 6 presents experimental results, and Section 7 concludes with a discussion of implications and future directions. The Appendix contains most of the proofs.

2 Preliminaries

We will use S to denote a string over the alphabet $\{A, C, G, T\}$. We define L to denote the number of characters in S. In this paper, we assume that there is a fixed integer $1 \leq k \leq L$, which forms the basis of our analysis. We will use L_0 as a shorthand for $L - k + 1$. We will also assume that $L \geq k$. We use S_i to denote the i-th character in S, where $1 \leq i \leq L$. Let c_A be the number of 'A's in S.

We use the mutation model from [18]. This model is a type of indel channel and various variations of it have been used to model sequence evolution (e.g. [7]). The mutation process takes as input a string S and three parameters p_s, p_d, and g, where $0 \leq p_s, p_d < 1$, $p_s + p_d < 1$, and $g \geq 0$. It performs the following steps on S:

1. For each i, independently choose an operation from the set {Sub, Del, Stay} with respective probabilities p_s, p_d, and $1 - p_s - p_d$.
2. Let *track* be a function mapping from a position in the original string S to its position in the modified S. Initially, $track(i) = i$ for all i. We assume that *track* is updated accordingly whenever S will be modified by either deleting or inserting characters.
3. For each i with a Sub operation, select a character $c \in \{A, C, G, T\} \setminus \{S_i\}$ uniformly at random and replace S_i by c.
4. For each i, let $I_i \geq 0$ be a sample from a geometric distribution with mean g. Then generate a random string of length exactly I_i by drawing each character of the string from $\{A, C, G, T\}$ independently and uniformly at random. We call this the *insert string associated with position i*. If $I_i = 0$, then this is the empty string.
5. For every $i > 1$, insert the string associated with position i between $S_{track(i)-1}$ and $S_{track(i)}$, updating *track* accordingly after each insertion.
6. Prepend the insert string associated with position 1 to the start of S, updating *track* accordingly.
7. For every i with a Del operation, remove the character $S_{track(i)}$ from S, updating *track* accordingly after each deletion.

A k-span is an interval of length k on the interval $[1, L]$. We define K_i as the k-span starting at position $1 \leq i \leq L_0$, i.e. the interval $[i, i + k - 1]$. We say that K_i *contains a mutation* if there exists a $i \leq j < i + k$ with an operation other than Stay or there exists a $i < j < i + k$ such that $l_i \neq 0$.

We use S' to denote the random string output after the mutation process is applied on S, and we define L', S'_i, and $c_A{}'$ for S' analogously to how they are defined for S. We will also make use of three random variables based on the mutation process.

- $N \triangleq$ the number of k-spans that have no mutation.
- $D \triangleq$ the number of k-spans that have no mutation except for exactly one deletion.
- $P \triangleq L' - 4c_A{}'$.

3 Main result

For the remainder of the manuscript, we assume that the mutation process described in Section 2 is applied to string S of length L, with the unknown parameters p_s, p_d, and g, and a mutated string S' is returned. Rather than observing the full string S', we limit ourselves to deriving estimators from L', $c_A{}'$, D, and N. We assume that L' and $c_A{}'$ can be directly observed from the data, but we postpone the discussion on how we can observe D and N until Section 5.

To derive our estimators, we take the following approach. We derive the expected values of the observed variables as a function of our model parameters, and then plug in the observed variables in place of their expectations and solve for the model parameters. The expectations of the observed variables are straightforward to derive and given by the following lemma.

▶ **Lemma 1.**

$$\mathbf{E}[L'] = L\,(1 + g - p_d), \tag{1}$$

$$\mathbf{E}[N] = L_0\,(1 - p_s - p_d)^{\,k}\,\frac{1}{(g+1)^{\,k-1}}. \tag{2}$$

$$\mathbf{E}[D] = L_0\,k\,(1 - p_s - p_d)^{k-1}\,p_d\,\frac{1}{(g+1)^{\,k-1}}. \tag{3}$$

$$\mathbf{E}[c_A{}'] = c_A(1 - p_s - p_d) + \frac{p_s(L - c_A)}{3} + \frac{gL}{4}. \tag{4}$$

$$\mathbf{E}[P] = (L - 4c_A)(1 - \frac{4}{3}p_s - p_d) \tag{5}$$

Next, we algebraically manipulate the equations in Lemma 1 in order to get formulas for p_s, p_d, and g. From Lemma 1 we get

$$\frac{\mathbf{E}[D]}{\mathbf{E}[N]} = \frac{k\,p_d}{1 - p_s - p_d}. \tag{6}$$

Note that we are not dividing by zero because our mutation model assumes $p_s + p_d < 1$. From (1), (4) and (6), we obtain the following system of linear equations with variables p_s, p_d, and g.

$$\begin{aligned} -p_d + g &= \frac{\mathbf{E}[L']}{L} - 1 \\ \frac{L - 4c_A}{3}\,p_s - c_A\,p_d + \frac{L}{4}g &= \mathbf{E}[c_A{}'] - c_A \\ \mathbf{E}[D]p_s + (k\mathbf{E}[N] + \mathbf{E}[D])\,p_d &= \mathbf{E}[D]. \end{aligned}$$

Solving this system of equations, we obtain that:

$$p_s = 3\frac{k\,\mathbf{E}[N]\Big(\mathbf{E}[L'] - L + 4c_A - 4\mathbf{E}[c_A{}']\Big) + \mathbf{E}[D]\Big(\mathbf{E}[L'] - 4\mathbf{E}[c_A{}']\Big)}{\Big(4c_A - L\Big)\Big(\mathbf{E}[D] + 4k\mathbf{E}[N]\Big)}, \tag{7}$$

$$p_d = \frac{\mathbf{E}[D](4c_A + 12\mathbf{E}[c_A{}'] - L - 3\mathbf{E}[L'])}{(4c_A - L)(\mathbf{E}[D] + 4k\mathbf{E}[N])}, \tag{8}$$

$$g = \frac{\mathbf{E}[L']}{L} - 1 + \frac{\mathbf{E}[D](-4c_A - 12\mathbf{E}[c_A{}'] + L + 3\mathbf{E}[L'])}{(-4c_A + L)(\mathbf{E}[D] + 4k\mathbf{E}[N])}. \tag{9}$$

By replacing the expectations above with these observations, we obtain our estimators. That is,

$$\hat{p}_s \triangleq 3\frac{k\,N\Big(L' - L + 4c_A - 4c_A{}'\Big) + D\Big(L' - 4c_A{}'\Big)}{\Big(4c_A - L\Big)\Big(D + 4kN\Big)}, \tag{10}$$

$$\hat{p}_d \triangleq \frac{D\,[3(L' - 4c_A{}') + (L - 4c_A)]}{(L - 4c_A)\,(D + 4kN)}, \tag{11}$$

$$\hat{g} \triangleq \frac{L'}{L} - 1 + \frac{D\,[3(L' - 4c_A{}') + (L - 4c_A)]}{(L - 4c_A)\,(D + 4kN)}. \tag{12}$$

We briefly comment on our choice of estimators, as various statistical approaches based on a different set of observables could yield a different set of estimators. For example, we considered a variant of the estimators based on the counts of k-spans with a single insertion, substitution, or deletion (i.e. D). These quantities contain enough information to estimate the mutation parameters, specifically, by solving the non-linear system of equations given by (3) and similarly derived formulas for the expected number of k-spans with a single insertion and a single substitution; the resulting

estimators performed quite well in real data. However, establishing theoretical guarantees for these estimators proved challenging, as they were defined as roots of degree-k polynomials. Our current estimators address this theoretical limitation as they involve solely linear equations. As we shall see in Section 6, the performance of our estimators in real data is strong, and they strike a more favorable balance by offering both reasonable accuracy and rigorous theoretical guarantees.

We now provide our central theoretical result. Our main theorem, below, shows that our estimates of p_s, p_d, and g are concentrated in a symmetric interval around their true values.

▶ **Theorem 2.** *Let* $\theta_1 = \min\left(\frac{g^2}{2(g+1)^2}, \frac{p_d}{3}\right)$ *and* $\theta_2 = \max(g+1, 8)$. *Suppose* $4c_A < L$ *and* $\frac{4}{3}p_s + p_d < 1$. *Then, for* $\delta \in (0, 1/5)$*:*

1.
$$\Pr[|\hat{p}_d - p_d| \geq 4\delta] \leq 6k\exp\left(-\frac{\delta^2\mathbf{E}[D]}{3k}\right) + 3k\exp\left(-\frac{\delta^2\mathbf{E}[N]}{3k}\right) + 6\exp\left(-\frac{\delta^2\mathbf{E}[P]^2}{72Le^{2(4\theta_2+1)}}\right).$$

2.
$$\Pr[|\hat{p}_s - p_s| \geq 12\delta] \leq 9k\exp\left(-\frac{\delta^2\mathbf{E}[N]}{3k}\right) + 6k\exp\left(-\frac{\delta^2\mathbf{E}[D]}{3k}\right) + 6\exp\left(-\frac{\delta^2\mathbf{E}[P]^2}{72Le^{2(4\theta_2+1)}}\right).$$

3.
$$\Pr[|\hat{g}-g| \geq (5+g)\delta] \leq 6k\exp\left(-\frac{\delta^2\mathbf{E}[D]}{3k}\right)+3k\exp\left(-\frac{\delta^2\mathbf{E}[N]}{3k}\right)+6\exp\left(-\frac{\delta^2\mathbf{E}[P]^2}{72Le^{2(4\theta_2+1)}}\right)+4\exp\left(-\theta_1 L\delta^2\right).$$

The requirement that $4c_A < L$ in this theorem does not restrict generality: aside from equal nucleotide frequency (where the estimators are not defined), at least one character $c \in \{A, C, T, G\}$ must satisfy that 4 times its frequency is at most L. In addition, the assumptions that $\frac{4}{3}p_s + p_d < 1$ and $p_d \leq 1/2$ hold when p_s and p_d are small which is the case most frequently encountered in practice. They are also non-essential and are taken to simplify our theoretical proofs.

The error probability in Theorem 2 is small when each of the terms in the sum are small. Since each of these terms decays (at least) exponentially with δ^2 times an expectation that grows linearly with the length of the string (when the mutation parameters are fixed), they will all generally be small. The following corollary of Theorem 2 formalizes this idea.

▶ **Corollary 3.** *Let* $\theta_1 = \min\left(\frac{g^2}{2(g+1)^2}, \frac{p_d}{3}\right)$ *and* $\theta_2 = \max(g+1, 8)$. *Suppose there exists* $\varepsilon \in (0, 1)$ *independent of L such that* $(4+\varepsilon)c_A < L$, *and* $\frac{4}{3}p_s + p_d < 1$. *Suppose further that* p_s, p_d, g, *and* k *are constants independent of L. Then, there exist constants* $\gamma = \gamma(p_s, p_d, g, \varepsilon, k) > 0$ *and* $C = C(k) > 0$ *such that for any* $\delta \in (0, 1/5)$*:*

1. $\Pr[|\hat{p}_d - p_d| \geq \delta] \leq Ce^{-\gamma\delta^2 L}$;
2. $\Pr[|\hat{p}_s - p_s| \geq \delta] \leq Ce^{-\gamma\delta^2 L}$;
3. $\Pr[|\hat{g} - g| \geq \delta] \leq Ce^{-\gamma\delta^2 L}$.

4 Proof outline of main theorem and corollary

In this section, we will outline the proof of the main theorem, with the missing proofs presented in Appendix B. We start by rewriting (7)-(12) in a more convenient form. Recall that $P = L' - 4c'_A$ and define the following additional three variables:

$$Q \triangleq kN + D \qquad R \triangleq k(L - 4c_A)N \qquad T \triangleq 4kN + D.$$

Applying these definitions to (7)-(12), we get the following alternate formulations for the true and estimated parameters.

$$p_s = -3 \cdot \frac{\mathbf{E}[P]\mathbf{E}[Q] - \mathbf{E}[R]}{(L-4c_A)\mathbf{E}[T]} \qquad \hat{p}_s = -3 \cdot \frac{PQ - R}{(L-4c_A)T}, \tag{13}$$

$$p_d = \frac{\mathbf{E}[D]\,[3\mathbf{E}[P] + (L-4c_A)]}{(L-4c_A)]\mathbf{E}[T]} \qquad \hat{p}_d = \frac{D\,[3P + (L-4c_A)]}{(L-4c_A)T}, \tag{14}$$

$$g = \frac{\mathbf{E}[L']}{L} - 1 + p_d \qquad \hat{g} = \frac{L'}{L} - 1 + \hat{p}_d. \tag{15}$$

In order to derive the concentration bounds of Theorem 2, we will first derive concentration bounds for P, Q, R, T, and L'. Q, R, and T are all linear combinations of N and D and their concentration bounds follow easily once the concentration bounds for N and D are derived. The concentration bounds for N and D stem from the fact that, from the perspective of the mutation process k-spans located more than k positions apart are independent of each other. We can therefore partition the set of L_0 k-spans into k subsets, such that the k-spans in each subset are independent of each other. We can then apply Chernoff bounds to each part and combine these bounds together via a union bound into a bound for N. In this way, we can obtain the following lemma.

▶ **Lemma 4.** *Suppose that $4c_A < L$. For any $\delta \in (0,1)$, all the following hold:*

$$\Pr\Big[\big|N - \mathbf{E}[N]\big| \geq \delta\mathbf{E}[N]\Big] \leq 3k\exp\Big(-\frac{\delta^2\mathbf{E}[N]}{3k}\Big), \tag{16}$$

$$\Pr\Big[\big|D - \mathbf{E}[D]\big| \geq \delta\mathbf{E}[D]\Big] \leq 3k\exp\Big(-\frac{\delta^2\mathbf{E}[D]}{3k}\Big), \tag{17}$$

$$\Pr\Big[\big|R - \mathbf{E}[R]\big| \geq \delta\mathbf{E}[R]\Big] \leq 3k\exp\Big(-\frac{\delta^2\mathbf{E}[N]}{3k}\Big), \tag{18}$$

$$\Pr\Big[\big|Q - \mathbf{E}[Q]\big| \geq \delta\mathbf{E}[Q]\Big] \leq 3k\exp\Big(-\frac{\delta^2\mathbf{E}[N]}{3k}\Big) + 3k\exp\Big(-\frac{\delta^2\mathbf{E}[D]}{3k}\Big), \tag{19}$$

$$\Pr\Big[\big|T - \mathbf{E}[T]\big| \geq \delta\mathbf{E}[T]\Big] \leq 3k\exp\Big(-\frac{\delta^2\mathbf{E}[N]}{3k}\Big) + 3k\exp\Big(-\frac{\delta^2\mathbf{E}[D]}{3k}\Big). \tag{20}$$

To prove concentration for the length of the mutated string L', we separately bound the number of total inserted and deleted characters, which are negative binomially (i.e. sum of geometrically distributed variables) and binomially distributed, respectively.

▶ **Lemma 5.** *Let $\theta_1 = \min\Big(\frac{g^2}{2(g+1)^2}, \frac{p_d}{3}\Big)$. For any $\delta \in (0,1)$:*

$$\Pr\Big[|L' - \mathbf{E}[L']| \geq \delta(Lg + Lp_d)\Big] \leq 4e^{-\theta_1 L\delta^2}.$$

The proof of the concentration bound for P is more challenging, as P is a sum of independent but unbounded random variables and therefore the Chernoff and Hoeffding bounds cannot be applied directly. However, we are able to show that P can be expressed as the sum of independent random variables that are sub-exponential. The concentration of such sums can be bounded via Bernstein inequalities. We first derive a clean form of Bernstein's inequality well-suited to our setting.

▶ **Lemma 6.** *Let $\theta_2 = \max(g+1, 8)$. Then, the following holds for any $\delta > 0$:*

$$\Pr\Big[|P - \mathbf{E}[P]| \geq 2\delta\Big] \leq 4\exp\Big(-\frac{\delta^2}{8Le^{2(4\theta_2+1)}}\Big) + 2\exp\Big(-\frac{\delta}{2e^{4\theta_2+1}}\Big).$$

If we also assume that $4c_A < L$, $\frac{4}{3}p_s + p_d < 1$, and for any $\delta_0 \in (0,1)$ then

$$\Pr\Big[|P - \mathbf{E}[P]| \geq \delta_0\mathbf{E}[P]\Big] \leq 6\exp\Big(-\frac{\delta_0^2\mathbf{E}[P]^2}{72Le^{2(4\theta_2+1)}}\Big).$$

In order to combine these individual concentration inequalities into concentration inequalities for our estimators, which are ratios, we make use of the following technical lemma.

▶ **Lemma 7.** *Let X, Y, and Z be random variables with strictly positive means. Let $\delta \in (0, 1/5)$. Suppose that $(1-\delta)\,\mathbf{E}[X] \le X \le (1+\delta)\,\mathbf{E}[X]$ and $(1-\delta)\,\mathbf{E}[Y] \le Y \le (1+\delta)\,\mathbf{E}[Y]$. Then*

$$(1-4\delta)\frac{\mathbf{E}[X]}{\mathbf{E}[Y]} \le \frac{X}{Y} \le (1+4\delta)\frac{\mathbf{E}[X]}{\mathbf{E}[Y]}.$$

If it also holds that $(1-\delta)\,\mathbf{E}[Z] \le Z \le (1+\delta)\,\mathbf{E}[Z]$, then

$$(1-4\delta)\frac{\mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{E}[Z]} \le \frac{XY}{Z} \le (1+4\delta)\frac{\mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{E}[Z]},$$

We can now prove Theorem 2. Here, we give the proof of the concentration of $\hat{p}_d$, and leave the rest for Appendix B.

Proof of Theorem 2, for $\hat{p}_d$. Consider the case that D, P, and T are indeed close to their means, i.e. for all $X \in \{D, P, T\}$ and all $\delta \in (0, 1/5)$, we have $(1-\delta)\,\mathbf{E}[X] \le X \le (1+\delta)\,\mathbf{E}[X]$. Because we assume that $\frac{4}{3}p_s + p_d < 1$ and $4c_A < L$, we use (5) to get that $\mathbf{E}[P] > 0$; Lemma 1 gives $\mathbf{E}[D], \mathbf{E}[T] > 0$. Therefore, we can apply Lemma 7 to the definition of $\hat{p}_d$ (14) to get

$$\hat{p}_d = \frac{3DP}{(L-4c_A)T} + \frac{D}{T} \le (1+4\delta)\left(\frac{3\mathbf{E}[D]\mathbf{E}[P]}{(L-4c_A)]\mathbf{E}[T]} + \frac{\mathbf{E}[D]}{\mathbf{E}[T]}\right) \le (1+4\delta)p_d \le p_d + 4\delta.$$

In a symmetric manner, we obtain that $\hat{p}_d \ge p_d - 4\delta$. Hence, under the assumptions that P, D, and T, are close to their means, $|\hat{p}_d - p_d| \le 4\delta$. Conversely, when $\hat{p}_d$ is outside this range, at least one of these three variables must be far from the mean. That is,

$$\Pr\left[|\hat{p}_d - p_d| \ge 4\delta\right] \le \Pr\left[\bigcup_{X\in\{D,P,T\}} |X - \mathbf{E}[X]| \ge \delta\mathbf{E}[X]\right] \le \sum_{X\in\{D,P,T\}} \Pr\left[|X - \mathbf{E}[X]| \ge \delta\mathbf{E}[X]\right]$$

by a union bound. The result for $\hat{p}_d$ follows by plugging in the individual concentration bounds for P (Lemma 6), Q, and T (Lemma 4). ◀

Proof of Corollary 3. We prove the result for $\hat{p}_d$; the results for $\hat{p}_s$ and $\hat{g}$ follow analogously. Under the assumption that the model parameters are fixed and independent of L, it follows from Lemma 1 that there exist positive constants $\gamma_D = \gamma_D(p_s, p_d, g, k)$, $\gamma_N = \gamma_N(p_s, p_d, g, k)$, and $\gamma_P = \gamma_P(p_s, p_d, \varepsilon)$ independent of L such that $\mathbf{E}[D] \ge \gamma_D L$, $\mathbf{E}[N] \ge \gamma_N L$, and $\mathbf{E}[P] \ge \gamma_P L$. Plugging these bounds into part 1 of Theorem 2, we obtain:

$$\begin{aligned}\Pr[|\hat{p}_d - p_d| \ge 4\delta] &\le 6k\exp\left(-\frac{\delta^2\gamma_D L}{3k}\right) + 3k\exp\left(-\frac{\delta^2\gamma_N L}{3k}\right) + 6\exp\left(-\frac{\delta^2\gamma_P^2 L}{72e^{2(4\theta_2+1)}}\right)\\ &\le (9k+6)\exp\left(-\gamma\delta^2 L\right)\end{aligned}$$

where $\gamma = \min\left(\frac{\gamma_D}{3k}, \frac{\gamma_N}{3K}, \frac{\gamma_P^2}{72e^{2(4\theta_2+1)}}\right)$. ◀

5 Implementation details

Our estimators for the three rates require counting the number of k-spans with single deletion, D, and the number of k-spans with no mutation, N. Counting D and N can be challenging, particularly because k-mers do not contain the contextual information, and so we do not have access to their corresponding k-spans. An additional layer of complexity comes into play from the fact that identifying a k-mer with no mutation (or a single kind of mutation) is more difficult, considering many

edge cases that may arise from inserting the same character that has been deleted. These challenges are cimcumvented when we use k-spans, and therefore, counting D and N solely from k-mers is not trivial and can be considered an interesting problem in and of itself. We therefore implemented an ad hoc solution to estimate D and N given the two strings S and S'. The steps for estimating D and N are as follows.

We start by extracting all k-mers in S, and building a de Bruijn graph using these k-mers using the cuttlefish tool [13]. We then extract the unitigs from this graph. Let the set of unitigs computed from S be U. We also compute the unitigs in S' in a similar manner and call this U'. We next take an arbitrary unitig u from U, and align every unitig in U' with u. To allow for partial overlap, we use semi-global alignment by using the infix option in edlib [22], which makes sure that gaps at the beginning and at the end of the alignment are not penalized. For a particular $u \in U$, we align every $u' \in U'$ to make sure all relevant alignments are considered. We use these alignments to look at all windows of length k, and count D and N accordingly for u. We repeat this for all $u \in U$, and accumulate the measurements from individual u's into a single global count.

The motivation behind using unitigs is that if there is an isolated mutation, and if the mutation is in the first or the last position of a k-mer, then there is no way to understand if the mutation is a substitution, an insertion, or a deletion only from the k-mers. The only way to resolve this ambiguity (and other similar ambiguities) is to scan beyond the context of k characters – and unitigs are a natural way to do this. The core goal of our implementation of these steps described above was not to make it efficient, but rather to obtain a working solution. We found that executing these steps estimates D and N reasonably well, and the estimated rates are also acceptable. As such, we leave finding an efficient way to compute D and N as an open research question.

6 Experiments and results

In this section, we present a series of experiments to evaluate the performance of the estimators detailed in Section 3. As discussed earlier, these estimators are sensitive to several input parameters, including k-mer size, sequence length, and the fraction of 'A' characters in the sequence. Sections 6.1 through 6.3 explore the sensitivity of the estimators with respect to these parameters. In Sections 6.4 and 6.5, we estimate mutation rates across a wide range of known rate combinations. And finally, in Section 6.6, we demonstrate that our estimated substitution rate outperforms estimates obtained under a substitution-only mutation model. For the experiments in Sections 6.1 through 6.4, the original sequence is a randomly generated synthetic sequence. In these cases, we compute the number of k-spans containing a single deletion and the number of k-spans with no mutation directly from the known mutation process. For the experiments in Sections 6.5 and 6.6, we use real reference genomes as the original sequences. In these cases, the two types of k-span counts are estimated using the steps described in Section 5.

6.1 Sensitivity of the estimators to k-mer lengths

We begin our analysis by examining how the choice of k-mer length affects the estimation of the mutation rates. To investigate this effect, we first generated a synthetic reference sequence of 1 million nucleotides, randomly sampling bases with fixed frequencies: 30% 'A', and equal proportions of 'C', 'G', and 'T' – making sure total frequency is 100%. From this reference, we simulated 20 mutated sequences, independently from each other, using the mutation model described in Section 2. For each of these mutated sequences, we estimated mutation rates using the estimators defined in Section 3 using a range of values for k. The results of this analysis are summarized in Figure 1.

As illustrated, the choice of k has a substantial impact on the stability of the estimators. In particular, longer k-mers tend to produce estimates with higher variability. This behavior is consistent

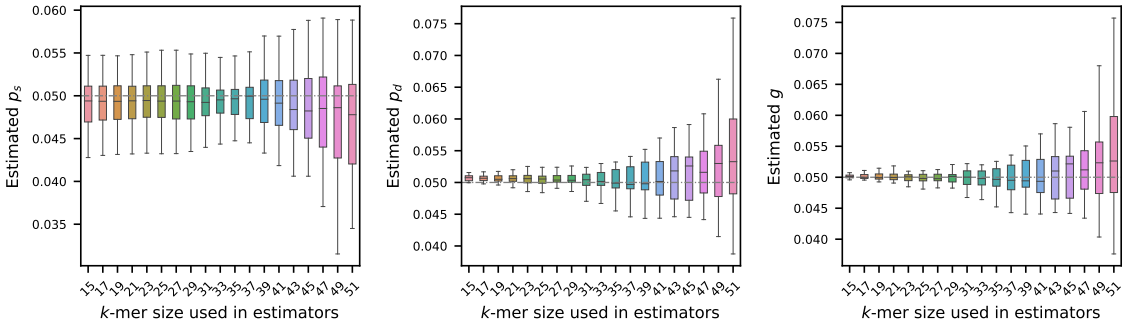

Figure 1 *Effect of k-mer length on mutation rate estimation (true rates set to 0.05).* A synthetic genome of 1 million nucleotides, mutated genomes were generated by setting p_s, p_d, and $g = 0.05$ – shown by the gray dashed horizontal line. Estimated rates were then computed using a range of k-mer sizes. Each boxplot shows the variability in estimation across 20 simulations, with error bars showing one standard deviation. The plots show that the estimators become more accurate and more precise for shorter k-mers.

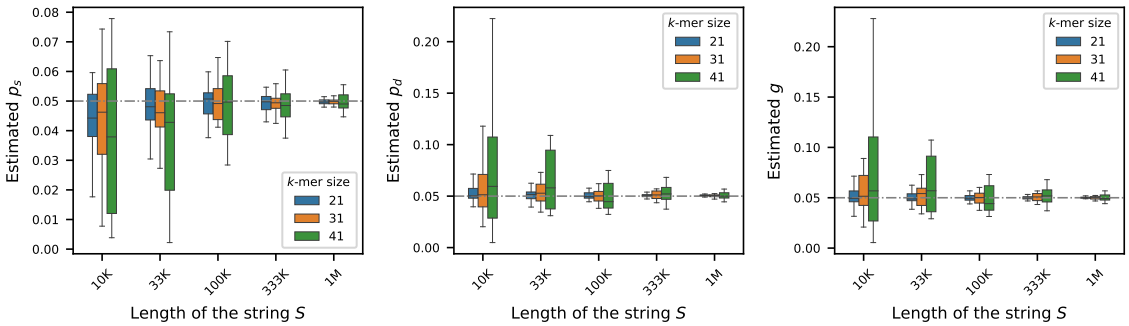

Figure 2 *Effect of sequence length on mutation rate estimation (true rates set to 0.05).* For synthetic genomes of varying lengths, mutated genomes were generated by setting p_s, p_d, and $g = 0.05$ – shown by the gray dashed horizontal line. Estimated rates were computed using three k-mer sizes: 21, 31, and 41. Each boxplot shows the variability in estimation across 20 simulations. The plots show that the estimators become more accurate and more precise for longer sequences.

with the known sensitivity of k-mers to mutations: since a single mutation can disrupt up to k consecutive k-mers, the longer the k-mer, the more susceptible it becomes to such perturbations. Our theoretical result in Theorem 2 also captures this: with a larger k, the error probabilities become larger, and the probabilistic guarantee for the estimators' performances decreases accordingly.

Interestingly, for the estimator of substitution probability p_s, we observe that the variability in the estimated values does not change significantly from 15 to 39. The reason behind this behavior is not immediately clear and warrants further investigation. It is possible that incorporating the number of k-spans with a single substitution into the estimators may correct this behavior, but additional analyses are required to substantiate this hypothesis.

6.2 Sensitivity of the estimators to sequence length

To investigate how the length of the original sequence S influences estimation of the mutation rates, we simulated synthetic genomes ranging from 10K to 1M nucleotides in length. For each genome length, we generated 10 independent synthetic sequences to capture variability due to random

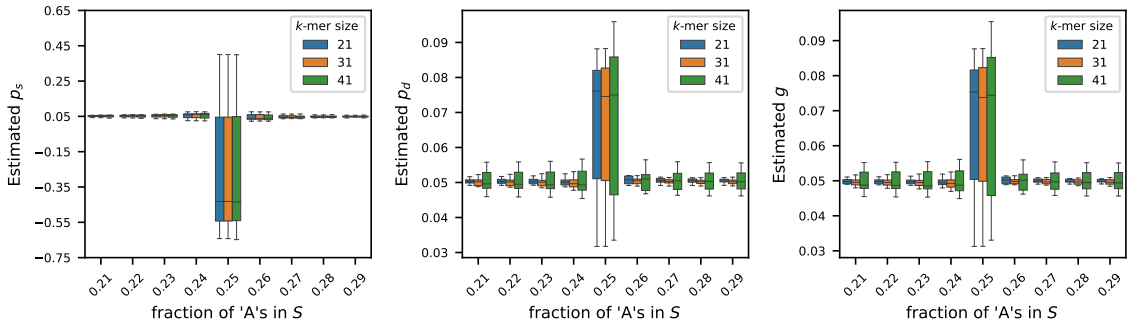

Figure 3 *Effect of nucleotide composition on mutation rate estimation (true rates set to 0.05).* For synthetic genomes of length $L = 1$M, the fraction of 'A's is varied from 0.21 to 0.29, and the frequencies of 'G's, 'C's, and 'T's are set equally. For each setting, the mutated string S' was generated by setting $p_s, p_d, g = 0.05$. Estimated rates were computed for three k-mer sizes: 21, 31, and 41. Each boxplot shows the variability in estimation across 20 simulations. The results show that the estimators generally work well for all three k-mer sizes, except when $c_A \approx L/4$, in which case the estimators become unstable – as predicted by Theorem 2.

sampling. The nucleotide composition of each sequence was fixed, with the frequency of 'A' set to 30% and frequencies of 'C', 'G', and 'T' set equally – making sure total frequency is 100%. For each synthetic sequence, we generated its mutated version by running the mutation process described in Section 2, setting each of p_s, p_d, and g to 0.05. We then estimated the mutation rates using the estimators outlined in Section 3 for three k-mer sizes: $k = 21, 31, 41$.

Figure 2 displays the estimated rates across the varying sequence lengths. As shown, the estimators are less stable for shorter sequences. However, with longer sequences, the estimators yield more accurate results – a trend expected from our core theoretical result in Theorem 2, which states that the associated error is asymptotically vanishing in L, the length of the string S: as L increases, the number of k-spans increases, and therefore the probability of error decreases, leading to a more precise estimation.

6.3 Sensitivity of the estimators to base composition

We generated synthetic genomes of 1M nucleotides to investigate how the fraction of 'A' characters affects the estimation of the mutation rates. We varied the fraction of 'A's from 21% to 29% in increments of 1%. For each fraction of 'A's, we set the frequency of 'C's, 'G's, and 'T's equally. For each preset fraction of 'A's, we generated 10 random genomes to capture stochastic variation. For each of these genomes, we generated its mutated version using the mutation process described in Section 2, setting each of p_s, p_d, and g to 0.05. We then estimated the mutation rates using the estimators described in Section 3 using three k-mer sizes: $k = 21, 31$, and 41. Figure 3 shows the sensitivity of the estimators to the fraction of 'A's in the original string S.

We observe that the estimators $\hat{p}_s$, $\hat{p}_d$, and $\hat{g}$ work reasonably well to estimate the true rates when the frequency of 'A' characters, c_A is not $L/4$. On the contrary, when the fraction of 'A's is exactly 25% in the original string S, the estimator gives inaccurate values, some of which are even negative (see estimated values of p_s). This behavior is captured in Theorem 2: when $L - 4c_A = 0$, $\mu_P = \mathbf{E}[P] = 0$, and the probabilistic guarantees become unbounded. We only get meaningful guarantees of concentration when $4c_A$ is strictly smaller than L. While Theorem 2 does not guarantee concentration when $4c_A > L$, this does not restrict generality, as explained in Section 3, and as demonstrated by the estimators' performances when the fraction of 'A's is larger than 1/4.

6.4 Estimating rates from a randomly generated synthetic sequence

After testing our estimators for varying k-mer lengths, sequence lengths, and base compositions, we next turn to estimating mutation rates by varying the true rates across a range of values. To do this, we generated a synthetic reference genome of 1 million base pairs, fixing the base composition at 30% 'A', and equal proportions of 'C', 'G', and 'T' – making sure total frequency is 100%. Using the mutation model described in Section 2, we then simulated mutated genomes from the synthetic reference by varying the mutation rates p_s, p_d, and g across the values $\{0.01, 0.02, 0.03, 0.04, 0.05\}$. For every parameter combination, we generated 10 independent mutated genomes to capture stochastic variability. We then estimated the mutation rates using the estimators detailed in Section 3 for each of these mutated genomes.

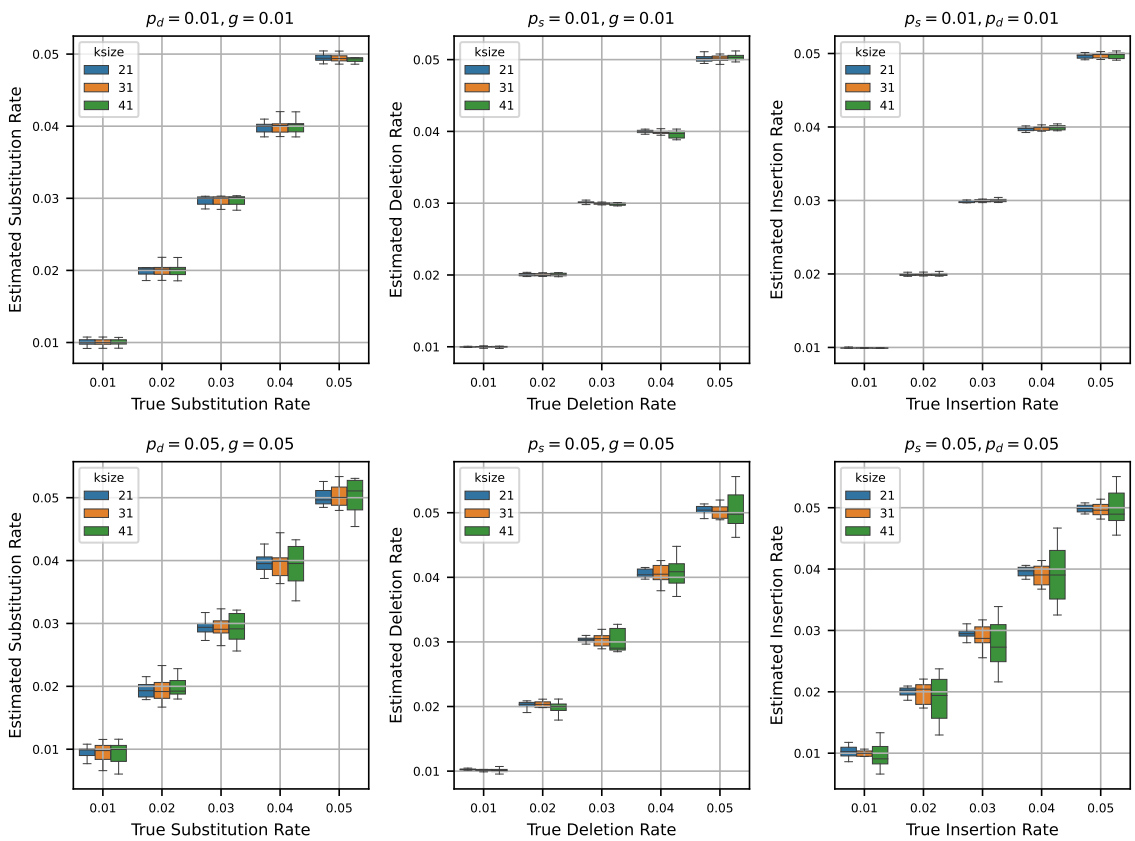

Figure 4 *Estimated mutation rates versus true values, where the original string is a synthetic sequence*. Each subplot corresponds to a case where two mutation rates are fixed (either at 0.01 or 0.05) and the third is varied from 0.01 to 0.05. Each boxplot shows the variability in estimation across 10 simulations. The results show that the rate estimation is very accurate when the other two rates are small, and is reasonably accurate when the other two rates are larger.

In Figure 4, we show two sets of results:

- **Fixed low rates (0.01):** Fixing two of the rates at 0.01, we show the estimates of the third rate as the true rate varies from 0.01 to 0.05. We repeat this process independently for p_s, p_d, and g.
- **Fixed high rates (0.05):** Fixing two of the rates at a higher value of 0.05, we also show the estimates of the third rate as the true rate varies from 0.01 to 0.05. Again, we repeat this for all

three rates.

When we set the other two rates to 0.01 and estimate the third rate, we observe that the estimated rates are highly accurate across all trials. In many cases, the boxplots of the estimates nearly vanish due to minimal variance, indicating tight clustering around the true values. This trend remains consistent across multiple k-mer sizes, suggesting that the estimators are robust at low rates of mutation.

In contrast, when we fix the other two rates at 0.05 and estimate the third rate, the accuracy of the estimation decreases slightly. While the estimates still track the true values reasonably well, the variance increases, and the boxplots become more prominent. Notably, the median estimate remains close to the true rate in most settings, which indicates that the estimators retain their central tendency even under higher mutation rates. However, for larger k-mer sizes, we observe increased variability in the estimates – an effect that mirrors our earlier observations in Section 6.1, where longer k-mers resulted in decreased precision of the estimators.

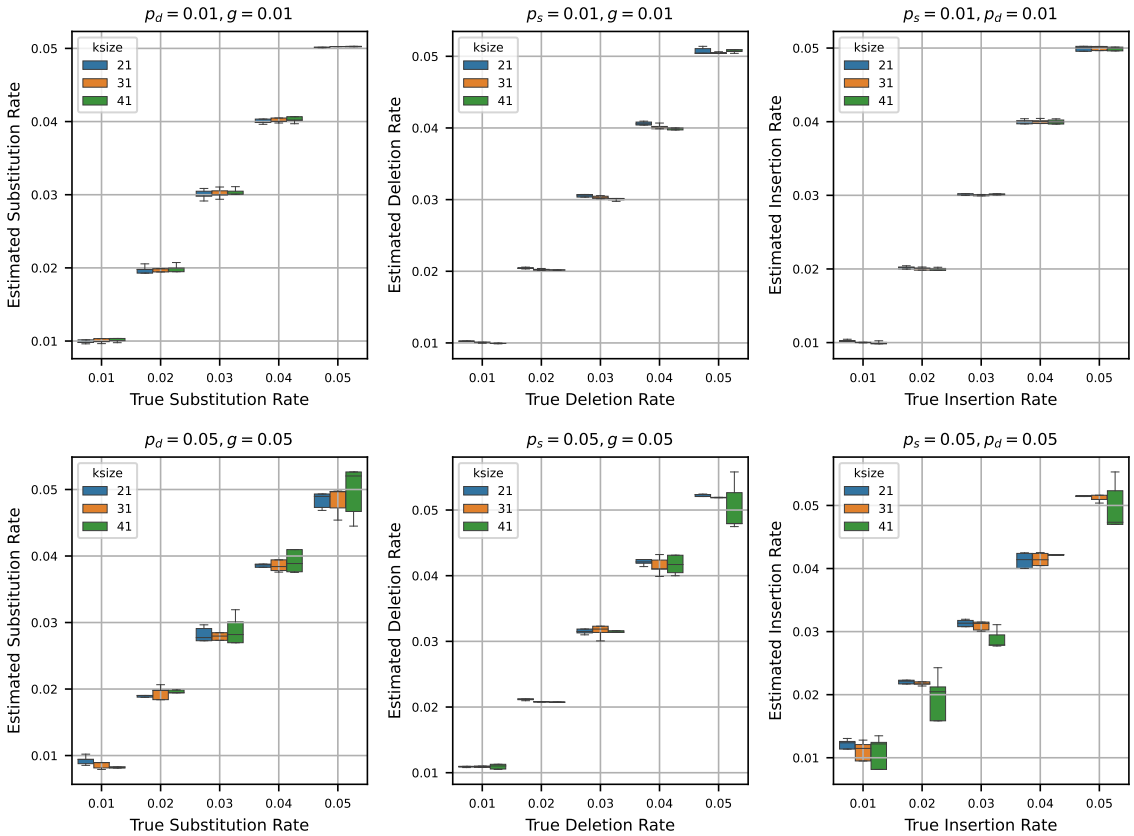

Figure 5 *Estimated mutation rates versus true values, where the original string is the reference genome of Staphylococcus (length 2.8 million).* Each subplot corresponds to a case where two mutation rates are fixed (either at 0.01 or 0.05) and the third varies from 0.01 to 0.05. Each boxplot shows the variability in estimation across 10 simulations. Estimated mutation rates closely match true values when other rates are low, but estimation becomes less precise under high total mutation rates (>10%) due to increased difficulty in real genomic contexts.

6.5 Estimating rates from real sequences

Having the estimators tested for a synthetic reference, we next estimate rates from a real genome sequence. For this set of experiments, we used the reference assembly of *Staphylococcus aureus* (subspecies: aureus USA300_TCH1516), which has 2.8 million nucleotides. We simulated mutated sequences from this reference by running the mutation process described in Section 2 by varying the mutation rates p_s, p_d, and g from 0.01 to 0.05. Similar to Section 6.4, we generated 10 independent mutated sequences for each combination of p_s, p_d, and g to capture stochastic variability. We then estimated the mutation rates using the estimators outlined in Section 3 for each of these simulated sequences.

Figure 5 shows the estimated mutation rates plotted against the true rates that were used to run the mutation process. We observe that the results shown in Figure 5 are consistent with previously discussed results. Specifically, when estimating a given mutation rate while keeping the other two rates low (0.01), the corresponding estimator performs with high precision, closely tracking the true value. On the other hand, when the estimation is carried out with the other two mutation rates set to higher values (0.05), the estimates appear more confounded. This is likely due to the increased difficulty of accurately estimating the number of k-spans with a single deletion (D) or no mutation (N) in a real genomic context. Notably, these experiments involve challenging conditions, with total mutation rates exceeding 10%. Despite this, the estimators yield reasonably accurate results, indicating potential for practical effectiveness.

6.6 Comparison with substitution rates estimated using simple mutation model

We conclude the experiments section by contrasting the substitution rates estimated using (10) with substitution rates estimated considering a simple mutation model. We use the statistics of k-mers developed in a recent work [2] to estimate substitution rates under a simple mutation model. The simple mutation model captures only substitutions, and no insertions or deletions. Consequently, we can only compute substitution rates considering this simple model. Henceforth, we refer to the simple mutation model as SMM.

We estimated the substitution rates using the SMM for the same simulated mutated sequences described in Section 6.5. In Figure 6, we show the estimated substitution rates using our estimators in (10), and using SMM. The results highlight that the substitution rates estimated using the estimator we developed track the true substitution rates accurately. On the other hand, the substitution rates estimated using SMM make a gross overestimation. This is because the SMM does not consider indels, and therefore, the effects of all three mutation rates are subsumed in the single substitution rate we get using the SMM. As such, a simple mutation model cannot disentangle the distinct contributions of substitution, insertion, and deletion rates. In contrast, the mutation model we introduce effectively decomposes these components, enabling more accurate and meaningful estimation of individual mutation rates.

7 Conclusions

We have presented a mutation model that accommodates single-nucleotide substitutions, as well as insertions and deletions while retaining enough mathematical structure to admit closed-form rate estimators derived solely from k-mer statistics. From this model, we obtained algebraic estimators for the three elementary mutation rates: p_s, p_d, and g and proved relatively tight sub-exponential concentration bounds on these estimators. We also identified regimes in which the estimation becomes ill-conditioned (i.e. large k, $p_d = 0$, or sequence composition with 25% 'A'). These results establish a

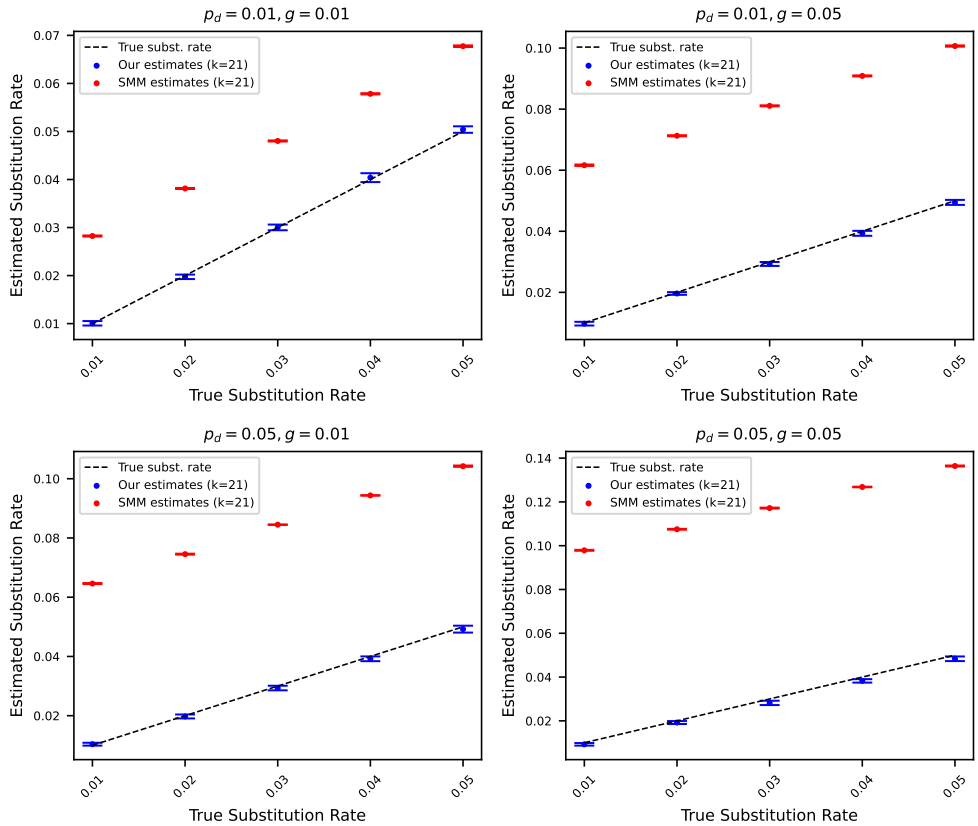

Figure 6 *Estimated substitution rates versus true substitution rates, where the original string is the reference genome of Staphylococcus (length 2.8 million).* Rates were estimated using (10) and using a simple mutation model (SMM) that only considers substitution. Each subplot corresponds to a case where p_d and g are fixed at either 0.01 or 0.05. The points show the average of 10 estimates, and the error bars show one standard deviation. The dashed black line corresponds to the true substitution rates. Estimated substitution rates using our method closely match true rates, whereas SMM overestimates due attributing to substitutions mismatching k-mers originating from insertions and deletions.

bridge between sequence evolution and combinatorial word statistics, thus providing additional tools for theoretical algorithmic computational biology.

In our prototype implementation, we demonstrated that our estimates remain accurate on simulated evolution of real genomes, and outperforms a substitution-only simple mutation model by avoiding spurious attribution of indel signals. While naive counting of non-mutated and single-deletion k-mers sufficed to show practical accuracy of our estimators, this raises an interesting open problem: estimating the number of these non-mutated and single-deletion k-mers efficiently for large scale data sets.

Several directions invite further investigation. First, incorporating the count of single-substitution k-spans may illuminate why p_s remains relatively stable even for moderately large k. Second, our framework can extend to heterogeneous or context-dependent rates by replacing global expectations with position-specific covariates. Third, coupling our estimators with sketch-based distance measures (such as in [11]) may provide a theory-backed avenue for larger scale applications such as phylogenetic placement in the presence of high indel activity. Finally, a more thorough investigation on real genomic data (where the unitig-based approach we used in the practical implementation starts to become

infeasible) will be necessary to understand the utility of the mutation estimates in practice.

In summary, by utilizing probabilistic modeling and concentration inequalities, we provide a theoretical foundation and initial practical implementation for quantifying the parameters of a relatively complex mutation process directly from k-mers. We anticipate that these ideas will continue to inform new alignment-free computational biology tasks, particularly relevant as sequencing data continues to outpace traditional alignment-based paradigms.

References

1. Dana Angluin and Leslie G Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. In *Proceedings of the ninth annual ACM symposium on Theory of computing*, pages 30–41, 1977.
2. Antonio Blanca, Robert S Harris, David Koslicki, and Paul Medvedev. The statistics of k-mers from a sequence undergoing a simple mutation process without spurious matches. *Journal of Computational Biology*, 29(2):155–168, 2022.
3. Ke Chen, Vinamratha Pattar, and Mingfu Shao. Sequence similarity estimation by random subsequence sketching. *bioRxiv*, pages 2025–02, 2025.
4. Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.
5. Margaret Oakley Dayhoff. Computer analysis of protein evolution. *Scientific American*, 221(1):86–95, 1969.
6. Richard Durbin, editor. *Biological sequence analysis*. Cambridge Univ. Press, Cambridge [u.a.], 17. print. edition, 2013.
7. Arun Ganesh and Aaron Sy. Near-Linear Time Edit Distance for Indel Channels. In Carl Kingsford and Nadia Pisanti, editors, *20th International Workshop on Algorithms in Bioinformatics (WABI 2020)*, volume 172 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 17:1–17:18. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.WABI.2020.17.
8. Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.
9. Mahmudur Rahman Hera, Shaopeng Liu, Wei Wei, Judith S Rodriguez, Chunyu Ma, and David Koslicki. Metagenomic functional profiling: to sketch or not to sketch? *Bioinformatics*, 40(Supplement_2):ii165–ii173, 2024.
10. Mahmudur Rahman Hera, Paul Medvedev, David Koslicki, and Antonio Blanca. Estimation of substitution and indel rates via k-mer statistics. In *25th International Workshop on Algorithms in Bioinformatics (WABI 2025)*, Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025.
11. Mahmudur Rahman Hera, N Tessa Pierce-Ward, and David Koslicki. Deriving confidence intervals for mutation rates across a wide range of evolutionary distances using FracMinHash. *Genome Research*, 33(7):1061–1068, 2023.
12. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
13. Jamshed Khan and Rob Patro. Cuttlefish: fast, parallel and low-memory compaction of de bruijn graphs from large-scale genome collections. *Bioinformatics*, 37(Supplement_1):i177–i186, 2021.
14. Brian D Ondov, Gabriel J Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B Buck, and Adam M Phillippy. Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biology*, 20:1–13, 2019.
15. Mahmudur Rahman Hera and David Koslicki. Estimating similarity and distance using fracminhash. *Algorithms for Molecular Biology*, 20(1):1–13, 2025.
16. Sophie Röhling, Alexander Linne, Jendrik Schellhorn, Morteza Hosseini, Thomas Dencker, and Burkhard Morgenstern. The number of k-mer matches between two dna sequences as a function of k and applications to estimate phylogenetic distances. *Plos one*, 15(2):e0228070, 2020.

17 Shahab Sarmashghi, Kristine Bohmann, M Thomas P Gilbert, Vineet Bafna, and Siavash Mirarab. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, 20(1):1–20, 2019.

18 Tizian Schulz and Paul Medvedev. ESKEMAP: exact sketch-based read mapping. *Algorithms for Molecular Biology*, 19(1):19, 2024.

19 Jim Shaw and Yun William Yu. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods*, 20(11):1661–1665, 2023.

20 Jim Shaw and Yun William Yu. Rapid species-level metagenome profiling and containment estimation with sylph. *Nature Biotechnology*, pages 1–12, 2024.

21 Kai Song, Jie Ren, Gesine Reinert, Minghua Deng, Michael S Waterman, and Fengzhu Sun. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in bioinformatics*, 15(3):343–353, 2014.

22 Martin Šošić and Mile Šikić. Edlib: a c/c++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 2017.

23 Shiheng Tao, Yanhui Fan, Wenjuan Wang, Guoji Ma, Lijing Liang, and Qi Shi. Patterns of insertion and deletion in mammalian genomes. *Current Genomics*, 8(6):370–378, 2007.

24 Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

25 Kun Wu, Danqi Qin, Yang Qian, and Haoxuan Liu. A new era of mutation rate analyses: Concepts and methods. *Zoological Research*, 45(4):767, 2024.

26 Andrzej Zielezinski, Susana Vinga, Jonas Almeida, and Wojciech M Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology*, 18:1–17, 2017.

A Standard concentration inequalities

This section gathers several definitions and standard concentration inequalities that we use in our proofs. Certain sums of independent random variables exhibit sharp concentration, and this is captured by the following standard inequalities. The first one is the Chernoff bound for sums of independent and identically distributed Bernoulli random variables [1, 4].

▶ **Lemma 8.** *Let $X_1, \ldots, X_n$ be a sequence of independently and identically distributed Bernoulli random variables. Consider $X = \sum_{i=1}^{n} X_i$ and let $\mathbf{E}[X] = \mu$. Then, the following three inequalities hold:*

$$\Pr[X \geq (1+\delta)\mu] \leq e^{-\delta^2\mu/(2+\delta)}, \qquad 0 \leq \delta. \tag{21}$$

$$\Pr[X \leq (1-\delta)\mu] \leq e^{-\delta^2\mu/2}, \qquad 0 < \delta < 1. \tag{22}$$

$$\Pr[|X-\mu| \geq \delta\mu] \leq 2e^{-\delta^2\mu/3}, \qquad 0 < \delta < 1. \tag{23}$$

A more general inequality that applies to sums of independent and bounded random variables is Hoeffding's inequality [12].

▶ **Lemma 9.** *Let $X_1, \ldots, X_n$ be a sequence of independent random variables taking values in $[a, b]$. Let $X = \sum_{i=1}^{n} X_i$, and $\mathbf{E}[X] = \mu$. Then, for any $\delta > 0$:*

$$\Pr\left[|X-\mu| \geq \delta\right] \leq 2\exp\left(-\frac{2\delta^2}{n(b-a)^2}\right).$$

We will also analyze sums of certain *unbounded* random variables. For this, we recall the definition of sub-exponential random variables. For a comprehensive survey about this type of random variables, the reader is referred to [24, Chapter 2.7].

▶ **Definition 10.** *Let a, b be fixed positive real numbers. A random variable X such that $\mathbf{E}[X] = 0$ is called (a, b)-sub-exponential if for all $t \geq 0$*

$$\Pr\left[|X| \geq t\right] \leq ae^{-t/b}.$$

For sums of (a, b)-sub-exponential random variables, the following version of Bernstein's inequality can be derived following the argument in the proof of [24, Theorem 2.8.1].

▶ **Lemma 11.** *Let a, b be fixed positive real numbers and let $X_1, \ldots, X_n$ be a sequence of independent (a, b)-sub-exponential random variables. Let $X = \sum_{i=1}^{n} X_i$. Then, for every $t \geq 0$:*

$$\Pr\left[|X| \geq t\right] \leq 2\exp\left(-\min\left(\frac{t^2}{8ne^{2(ab+1)}}, \frac{t}{2e^{ab+1}}\right)\right).$$

For completeness, we provide a proof of this concentration inequality in Appendix B. We derive and use this variant of Bernstein's inequality instead of the one in [24, Theorem 2.8.1] since that one does not track the dependency on a and b in the tail bound.

We will also use the following lemma to combine concentration inequalities, with a proof given in Appendix C.

▶ **Lemma 12.** *For all random variables $X_1, \ldots, X_n$ and numbers $t_1, \ldots, t_n$,*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i\right| \geq \sum_{i=1}^{n} t_i\right] \leq \sum_{i=1}^{n} \Pr\left[|X_i| \geq t_i\right].$$

B Proof of main result

▶ **Lemma 1.**

$$\mathbf{E}[L'] = L\,(1 + g - p_d), \tag{1}$$

$$\mathbf{E}[N] = L_0\,(1 - p_s - p_d)^{\,k}\,\frac{1}{(g+1)^{\,k-1}}. \tag{2}$$

$$\mathbf{E}[D] = L_0\,k\,(1 - p_s - p_d)^{k-1}\,p_d\,\frac{1}{(g+1)^{\,k-1}}. \tag{3}$$

$$\mathbf{E}[c_A{}'] = c_A(1 - p_s - p_d) + \frac{p_s(L - c_A)}{3} + \frac{gL}{4}. \tag{4}$$

$$\mathbf{E}[P] = (L - 4c_A)(1 - \frac{4}{3}p_s - p_d) \tag{5}$$

Proof. *Proof of* (1)*:* Let W_i be an indicator random variable that is 1 in the event that character S_i was deleted. Then $\mathbf{E}[L'] = \mathbf{E}[L + \sum_{i=1}^{L}(W_i + I_i)] = L + \sum_{i=1}^{L}(\mathbf{E}[W_i] + \mathbf{E}[I_i]) = L + L(p_d + g)$.

Proof of (2)*:* Let N_i be the indicator that k-span K_i has no mutations. This happens when each of its k positions has a Stay operation and there are no string inserted inside the k-span, i.e. $l_j = 0$ for each $j \in (i+1, \ldots, i+k-1)$. This happens with probability $(1 - p_s - p_d)^k/(g+1)^{k-1}$. The result follows since $N = \sum_{i=1}^{L_0} N_i$.

Proof of (3)*:* Let D_i be an indicator that K_i has a single deletion and no other mutation. This happens when there are no strings inserted inside the k-span, one position has a deletion, and $k-1$ other positions have a Stay operation. The single deletion operation could be at any of the k positions. Then, $\mathbf{E}[D_i] = k(1 - p_s - p_d)^{k-1}p_d/(g+1)^{k-1}$. The result follows since $D = \sum_{i=1}^{L_0} D_i$.

Proof of (4)*:* Let X_i be an indicator S_i is not deleted or substituted, which happens with probability $1 - p_s - p_d$. Let Y_i bet the indicator S_i was substituted into 'A', which happens with probability $p_s/3$. Let Z_i be the number of 'A's in the inserted string associated with position i. We have that

$$\mathbf{E}[Z_i] = \sum_{\ell \geq 0} \mathbf{E}[Z_i \mid I_i = \ell]\,\Pr[I_i = \ell] = \sum_{\ell \geq 0} \frac{\ell}{4}\,\Pr[I_i = \ell] = \frac{g}{4}. \tag{24}$$

We can express $c_A{}'$ as: $c_A{}' = \sum_{i:S_i=\text{'A'}} X_i + \sum_{i:S_i\neq\text{'A'}} Y_i + \sum_{i=1}^{L} Z_i$, and thus get the claimed result that

$$\mathbf{E}[c_A{}'] = \sum_{i:S_i=\text{'A'}} \mathbf{E}[X_i] + \sum_{i:S_i\neq\text{'A'}} \mathbf{E}[Y_i] + \sum_{i=1}^{L} \mathbf{E}[Z_i] = c_A(1 - p_s - p_d) + (L - c_A)p_s/3 + gL/4.$$

Proof of (5)*:* It follows directly by linearity of expectation and (1) and (4). ◀

▶ **Lemma 4.** *Suppose that* $4c_A < L$*. For any* $\delta \in (0,1)$*, all the following hold:*

$$\Pr\Big[\big|N - \mathbf{E}[N]\big| \geq \delta\mathbf{E}[N]\Big] \;\leq\; 3k\exp\Big(-\frac{\delta^2\mathbf{E}[N]}{3k}\Big), \tag{16}$$

$$\Pr\Big[\big|D - \mathbf{E}[D]\big| \geq \delta\mathbf{E}[D]\Big] \;\leq\; 3k\exp\Big(-\frac{\delta^2\mathbf{E}[D]}{3k}\Big), \tag{17}$$

$$\Pr\Big[\big|R - \mathbf{E}[R]\big| \geq \delta\mathbf{E}[R]\Big] \;\leq\; 3k\exp\Big(-\frac{\delta^2\mathbf{E}[N]}{3k}\Big), \tag{18}$$

$$\Pr\Big[\big|Q - \mathbf{E}[Q]\big| \geq \delta\mathbf{E}[Q]\Big] \;\leq\; 3k\exp\Big(-\frac{\delta^2\mathbf{E}[N]}{3k}\Big) + 3k\exp\Big(-\frac{\delta^2\mathbf{E}[D]}{3k}\Big), \tag{19}$$

$$\Pr\Big[\big|T - \mathbf{E}[T]\big| \geq \delta\mathbf{E}[T]\Big] \;\leq\; 3k\exp\Big(-\frac{\delta^2\mathbf{E}[N]}{3k}\Big) + 3k\exp\Big(-\frac{\delta^2\mathbf{E}[D]}{3k}\Big). \tag{20}$$

Proof. *Proof of* (16)*:* Recall from the proof of Lemma 1 that we can express $N = \sum_{i=1}^{L_0} N_i$ where N_i is the indicator that K_i has no mutations. The N_i's are not independent, but N_i and N_j are independent if $|i-j| \geq k$. Then, for $\ell \in \{1, \ldots, k\}$, define

$$M_\ell \triangleq \sum_{j:\, 1 \leq \ell + kj \leq L-k+1} N_{\ell+kj}.$$

Observe that the M_ℓ's are defined such that each is a sum of independent identically distributed Bernoulli random variables with success probability $q = (1-p_s-p_d)^k/(g+1)^{k-1}$ and $N = \sum_{i=1}^{k} M_\ell$. The Chernoff bound in (23) implies that for any $\delta \in (0,1)$,

$$\Pr\left[\left|M_\ell - \mathbf{E}[M_\ell]\right| \geq \delta \mathbf{E}[M_\ell]\right] \;\leq\; 2\exp\Big(-\frac{\delta^2 \mathbf{E}[M_\ell]}{3}\Big),$$

Using the fact that $N = \sum_{i=1}^{k} M_\ell$ and Lemma 12, we get

$$\begin{aligned}
\Pr\left[\left|N - \mathbf{E}[N]\right| \geq \delta \mathbf{E}[N]\right] &= \Pr\left[\left|\sum M_\ell - \mathbf{E}[M_\ell]\right| \geq \sum \delta \mathbf{E}[M_\ell]\right] \\
&\leq \sum \Pr\left[|M_\ell - \mathbf{E}[M_\ell]| \geq \delta \mathbf{E}[M_\ell]\right] \\
&\leq \sum 2\exp\Big(-\frac{\delta^2 \mathbf{E}[M_\ell]}{3}\Big), \\
&\leq 2k\exp\Big(-\frac{\delta^2 \min_\ell \mathbf{E}[M_\ell]}{3}\Big),
\end{aligned}$$

Finally, observe that each M_ℓ is the sum of at least $\lfloor \frac{L_0}{k} \rfloor$ Bernoulli variables. Then $\mathbf{E}[M_\ell] \geq \lfloor \frac{L_0}{k} \rfloor q \geq \frac{\mathbf{E}[N]}{k} - q$. Using the fact $e^{q\delta^2/(3k)} \leq e^{1/3} \leq 1.5$, we get

$$\Pr\left[\left|N - \mathbf{E}[N]\right| \geq \delta \mathbf{E}[N]\right] \;\leq\; 2k\exp\Big(-\frac{\delta^2(\mathbf{E}[N]-q)}{3k}\Big) \;\leq\; 3k\exp\Big(-\frac{\delta^2 \mathbf{E}[N]}{3k}\Big).$$

Proof of (17)*:* The concentration of D is proven in an identical manner to the concentration of N. Recall from the proof of Lemma 1 that we can express $D = \sum_{i=1}^{L_0} D_i$, where D_i is the indicator that the K_i has no mutations except exactly one deletion. These D_is have the same dependence structure as the N_is, so the proof for the concentration of D is identical to the one for N.

Proof of (18)*:* Since we assumed $4c_A < L$, we can multiply both sides of the inequality inside the probability of (16) by $k(L-4c_A)$ to obtain (18).

Proof of (19)*:* The concentration from Q comes by using Lemma 12 to combine the the concentration inequalities of N and D. That is,

$$\begin{aligned}
\Pr\left[|Q - \mathbf{E}[Q]| \geq \delta \mathbf{E}[Q]\right] &= \Pr\left[|kN + D - k\mathbf{E}[N] - \mathbf{E}[D]| \geq \delta k \mathbf{E}[N] + \delta \mathbf{E}[D]\right] \\
&\leq \Pr\left[|kN - k\mathbf{E}[N]| \geq \delta k \mathbf{E}[N]\right] + \Pr\left[|D - \mathbf{E}[D]| \geq \delta \mathbf{E}[D]\right] \\
&= \Pr\left[|N - \mathbf{E}[N]| \geq \delta \mathbf{E}[N]\right] + \Pr\left[|D - \mathbf{E}[D]| \geq \delta \mathbf{E}[D]\right] \\
&\leq 3k\exp\Big(-\frac{\delta^2 \mathbf{E}[N]}{3}\Big) + 3k\exp\Big(-\frac{\delta^2 \mathbf{E}[D]}{3}\Big).
\end{aligned}$$

Proof of (20)*:* The concentration of T is derived in a way identical to the concentration of Q. ◀

▶ **Lemma 5.** *Let* $\theta_1 = \min\Big(\frac{g^2}{2(g+1)^2}, \frac{p_d}{3}\Big)$*. For any* $\delta \in (0,1)$*:*

$$\Pr\left[|L' - \mathbf{E}[L']| \geq \delta(Lg + Lp_d)\right] \leq 4e^{-\theta_1 L \delta^2}.$$

Proof. Let D_i be an indicator that is 1 in the event that the operation at i was Del. Let us define $\mathcal{L}_{\text{ins}} \triangleq \sum_{i=1}^{L} I_i$, and $\mathcal{L}_{\text{del}} \triangleq \sum_{i=1}^{L} D_i$, so that $L' = L + \mathcal{L}_{\text{ins}} - \mathcal{L}_{\text{del}}$. We will obtain the desired concentration inequality for L' from separate concentration inequalities for $\mathcal{L}_{\text{ins}}$ and $\mathcal{L}_{\text{del}}$.

First, notice that $\mathcal{L}_{\text{ins}} \geq (1+\delta)Lg$ is equivalent to observing at least $(1+\delta)Lg$ failures before observing L successes from a series of independent Bernoulli trials with success probability $1/(g+1)$. This is in turn equivalent to observing at most L successes in $(1+\delta)Lg + L$ independent Bernoulli trials with success probability $1/(g+1)$ each. Hence, if $\{Z_i\}_{i\geq 1}$ is a sequence of independent Bernoulli random variable with success probability $1/(g+1)$, we have

$$\Pr[\mathcal{L}_{\text{ins}} \geq (1+\delta)Lg] = \Pr\Big[\sum_{i=1}^{(1+\delta)Lg+L} Z_i \leq L\Big] = \Pr\Big[\sum_{i=1}^{(1+\delta)Lg+L} Z_i \leq (1-\varepsilon)\mu\Big]$$

where

$$\mu = \mathbf{E}\Big[\sum_{i=1}^{(1+\delta)Lg+L} Z_i\Big] = \frac{(1+\delta)Lg + L}{g+1}$$

and $\varepsilon = \frac{\delta g}{\delta g + g + 1}$. From the Chernoff bound in (22) we obtain

$$\Pr[\mathcal{L}_{\text{ins}} \geq (1+\delta)Lg] \leq \exp\Big(-\frac{L\delta^2 g^2}{2(g+1+\delta g)(g+1)}\Big).$$

Proceeding in an analogous manner and using the Chernoff bound of (21), we can prove the probability for the other tail

$$\Pr[\mathcal{L}_{\text{ins}} \leq (1-\delta)Lg] \leq \exp\Big(-\frac{L\delta^2 g^2}{(2+\varepsilon')(g+1-\delta g)(g+1)}\Big),$$

where $\varepsilon' = \frac{\delta g}{g+1-\delta g}$. From a union bound we deduce that:

$$\Pr[|\mathcal{L}_{\text{ins}} - Lg| \geq \delta Lg] \leq 2\exp\Big(-\frac{L\delta^2 g^2}{2(g+1-\delta g)(g+1)}\Big). \tag{25}$$

To establish the concentration of $\mathcal{L}_{\text{del}}$, note that this random variable is a sum of L independent Bernoulli random variable with parameter p_d. Then, $\mathbf{E}[\mathcal{L}_{\text{del}}] = Lp_d$ and the Chernoff bound in (23) yields that:

$$\Pr[|\mathcal{L}_{\text{del}} - Lp_d| \geq \delta Lp_d] \leq 2\exp\Big(-\frac{Lp_d\delta^2}{3}\Big). \tag{26}$$

Using Lemma 12, we now combine the concentration bounds of (25) and (26) into a concentration bound for L':

$$\begin{aligned}
\Pr\Big[\Big|L' - \mathbf{E}[L']\Big| \geq \delta(Lg + Lp_d)\Big] &= \Pr\Big[\Big|\mathcal{L}_{\text{ins}} - \mathcal{L}_{\text{del}} + Lp_d - Lg\Big| \geq \delta(Lg + Lp_d)\Big] \\
&\leq \Pr[|\mathcal{L}_{\text{ins}} - Lg| \geq \delta Lg] + \Pr[|\mathcal{L}_{\text{del}} - Lp_d| \geq \delta Lp_d] \\
&\leq 2\exp\Big(-\frac{L\delta^2 g^2}{2(g+1-\delta g)(g+1)}\Big) + 2\exp\Big(-\frac{Lp_d\delta^2}{3}\Big) \\
&\leq 2\exp\Big(-\frac{L\delta^2 g^2}{2(g+1)^2}\Big) + 2\exp\Big(-\frac{Lp_d\delta^2}{3}\Big).
\end{aligned}$$

Setting $\theta_1 = \min\Big(\frac{g^2}{2(g+1)^2}, \frac{p_d}{3}\Big)$, we obtain:

$$\Pr\Big[\Big|L' - \mathbf{E}[L']\Big| \geq \delta(Lg + Lp_d)\Big] \leq 4e^{-\theta_1\delta^2 L}.$$

◀

▶ **Lemma 6.** *Let $\theta_2 = \max(g+1, 8)$. Then, the following holds for any $\delta > 0$:*

$$\Pr\Big[|P - \mathbf{E}[P]| \geq 2\delta\Big] \leq 4\exp\Big(-\frac{\delta^2}{8Le^{2(4\theta_2+1)}}\Big) + 2\exp\Big(-\frac{\delta}{2e^{4\theta_2+1}}\Big).$$

If we also assume that $4c_A < L$, $\frac{4}{3}p_s + p_d < 1$, and for any $\delta_0 \in (0,1)$ then

$$\Pr\Big[|P - \mathbf{E}[P]| \geq \delta_0\mathbf{E}[P]\Big] \leq 6\exp\Big(-\frac{\delta_0^2\mathbf{E}[P]^2}{72Le^{2(4\theta_2+1)}}\Big).$$

Proof. We will consider the contributions to $P = L' - 4c_A{}'$ from the characters in S and from newly inserted characters separately. For this, we define the random variable X_i that is 0 if S_i is deleted, it is -3 if S_i is not deleted and is an 'A' after the mutation process, and it is 1 otherwise.

We also define the random variable $Z_i = I_i - 4C_i$ where recall that I_i is the length of the inserted string at position i and C_i is the random variable corresponding to the number of 'A's in the inserted string at position i. Note that each inserted character contributes 1 to P, if it is a non-'A', or contributes -3 to P, if it is a 'A'. We can now write $P = \sum_{i=1}^{L} X_i + \sum_{i=1}^{L} Z_i$. For ease of notation, let $X = \sum_{i=1}^{L} X_i$ and $Z = \sum_{i=1}^{L} Z_i$.

Observe that the values that X_i can take on are limited to the range of $[-3, 1]$. Using this and the fact that the X_is are independent of each other, we apply Hoeffding's inequality (Lemma 9) to get that for any $\delta > 0$,

$$\Pr\left[|X - \mathbf{E}[X]| \geq \delta\right] \leq 2\exp\Big(-\frac{\delta^2}{8L}\Big). \tag{27}$$

To show that Z is strongly concentrated around its mean, we first prove that the Z_i's are sub-exponential random variables and then use Bernstein's inequality (Lemma 11).

▶ **Lemma 13.** *Let $\theta_2 = \max(g+1, 8)$. For each Z_i, $\mathbf{E}[Z_i] = 0$ and, for all $t \geq 0$:*

$$\Pr[|Z_i| \geq t] \leq 4e^{-t/\theta_2}.$$

We prove this lemma shortly, but first we finish the proof of Lemma 6. As the Z_i's are mean-zero $(4, \theta_2)$-sub-exponential random variables, it follows from Lemma 11 that for any $\delta > 0$,

$$\Pr\Big[|Z| \geq \delta\Big] \leq 2\exp\Big(-\min\Big(\frac{\delta^2}{8Le^{2(4\theta_2+1)}}, \frac{\delta}{2e^{4\theta_2+1}}\Big)\Big). \tag{28}$$

We now use Lemma 12 to combine the concentrations of X and Z into a concentration for P. For any $\delta > 0$,

$$\begin{aligned}
\Pr\Big[|P - \mathbf{E}[P]| \geq 2\delta\Big] &= \Pr\Big[|X - \mathbf{E}[X] + Z - \mathbf{E}[Z]| \geq 2\delta\Big] \\
&\leq \Pr\Big[|X - \mathbf{E}[X]| \geq \delta\Big] + \Pr\Big[|Z - \mathbf{E}[Z]| \geq \delta\Big] \\
&\leq 2\exp\Big(-\frac{\delta^2}{8L}\Big) + 2\exp\Big(-\min\Big(\frac{\delta^2}{8Le^{2(4\theta_2+1)}}, \frac{\delta}{2e^{4\theta_2+1}}\Big)\Big) \\
&\leq 2\exp\Big(-\frac{\delta^2}{8L}\Big) + 2\exp\Big(-\frac{\delta^2}{8Le^{2(4\theta_2+1)}}\Big) + 2\exp\Big(-\frac{\delta}{2e^{4\theta_2+1}}\Big) \\
&\leq 4\exp\Big(-\frac{\delta^2}{8Le^{2(4\theta_2+1)}}\Big) + 2\exp\Big(-\frac{\delta}{2e^{4\theta_2+1}}\Big).
\end{aligned}$$

Further, applying the assumptions that $\frac{4}{3}p_s + p_d < 1$ and $4c_A < L$ to the expected value of P given by (5), we have that $0 < \mathbf{E}[P] < L$. Therefore, given $\delta_0 \in (0,1)$, we can plug $\delta = \delta_0 \mathbf{E}[P]/2$ into the above to obtain

$$\Pr\Big[|P - \mathbf{E}[P]| \geq \delta_0 \mathbf{E}[P]\Big] \leq 4\exp\left(-\frac{\delta_0^2 \mathbf{E}[P]^2}{32Le^{2(4\theta_2+1)}}\right) + 2\exp\left(-\frac{\delta_0 \mathbf{E}[P]}{4e^{4\theta_2+1}}\right) \leq 6\exp\left(-\frac{\delta_0^2 \mathbf{E}[P]^2}{32Le^{2(4\theta_2+1)}}\right).$$

◀

Proof of Lemma 13. Recall that $Z_i = I_i - 4C_i$, where I_i is a Geometric random variable with parameter $p \triangleq 1/(g+1)$ and mean g, and C_i is a Binomial random variable with parameters I_i and $1/4$. Observe that $\mathbf{E}[Z_i] = 0$, since $\mathbf{E}[I_i] = g$ and, by (24), $\mathbf{E}[C_i] = g/4$. Let B_y be a Binomial random variable with parameters y and $1/4$. For $t \geq 0$, conditioning on the value of I_i and using Hoeffding's inequality (Lemma 9), we obtain:

$$\begin{aligned}
\Pr[|Z_i| \geq t] &= \sum_{y \geq 0} \Pr[I_i = y] \Pr[|B_y - y/4| \geq t/4] \\
&\leq 2p \sum_{y \geq 0} (1-p)^y \exp\Big(\frac{-t^2}{8y}\Big) \\
&\leq 2p \left(\sum_{y=0}^{\lfloor t \rfloor} (1-p)^y \exp\Big(-\frac{t^2}{8y}\Big) + \sum_{y=\lfloor t \rfloor + 1}^{\infty} (1-p)^y \exp\Big(-\frac{t^2}{8y}\Big) \right).
\end{aligned}$$

For $y \leq t$, we have $\frac{t^2}{8y} \geq \frac{t}{8}$, so $\exp(-\frac{t^2}{8y}) \leq \exp(-t/8)$ and we have

$$\sum_{y=0}^{\lfloor t \rfloor} (1-p)^y \exp(-\frac{t^2}{8y}) \leq \exp(-t/8) \sum_{y=0}^{\lfloor t \rfloor} (1-p)^y \leq \frac{e^{-t/8}}{p}.$$

When $y > t$, we use the fact that $e^{-x} \leq 1$ for all positive x and apply the formula for a geometric series; we have

$$\sum_{y=\lfloor t \rfloor + 1}^{\infty} (1-p)^y \exp(-\frac{t^2}{8y}) \leq \sum_{y=\lfloor t \rfloor + 1}^{\infty} (1-p)^y \leq \frac{(1-p)^t}{p} \leq \frac{e^{-pt}}{p}.$$

Putting these bounds together:

$$\Pr[|Z_i| \geq t] \leq 2\left(e^{-t/8} + e^{-pt}\right) \leq 4e^{-t/\theta_2},$$

where $\theta_2 = \min\{p, 1/8\} = \max(g+1, 8)$. ◀

▶ **Lemma 7.** *Let X, Y, and Z be random variables with strictly positive means. Let $\delta \in (0, 1/5)$. Suppose that $(1-\delta)\,\mathbf{E}[X] \leq X \leq (1+\delta)\,\mathbf{E}[X]$ and $(1-\delta)\,\mathbf{E}[Y] \leq Y \leq (1+\delta)\,\mathbf{E}[Y]$. Then*

$$(1-4\delta)\frac{\mathbf{E}[X]}{\mathbf{E}[Y]} \leq \frac{X}{Y} \leq (1+4\delta)\frac{\mathbf{E}[X]}{\mathbf{E}[Y]}.$$

If it also holds that $(1-\delta)\,\mathbf{E}[Z] \leq Z \leq (1+\delta)\,\mathbf{E}[Z]$, then

$$(1-4\delta)\frac{\mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{E}[Z]} \leq \frac{XY}{Z} \leq (1+4\delta)\frac{\mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{E}[Z]},$$

Proof. First, using the fact that $\frac{(1+\delta)^2}{1-\delta} \leq 1 + 4\delta$, we get:

$$\frac{XY}{Z} \leq \frac{(1+\delta)^2}{1-\delta} \frac{\mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{E}[Z]} \leq (1+4\delta)\frac{\mathbf{E}[P]\mathbf{E}[Q]}{\mathbf{E}[T]}.$$

Using the fact that $\frac{(1-\delta)^2}{1+\delta} \geq 1 - 4\delta$, we get

$$\frac{XY}{Z} \geq \frac{(1-\delta)^2}{1+\delta}\frac{\mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{E}[Z]} \geq (1-4\delta)\frac{\mathbf{E}[X]\mathbf{E}[Y]}{\mathbf{E}[Z]}$$

Similarly, using the facts that $\frac{1-\delta}{1+\delta} \geq 1 - 3\delta$ and $\frac{1+\delta}{1-\delta} \leq 1 + 3\delta$, we obtain that $(1-3\delta)\frac{\mathbf{E}[X]}{\mathbf{E}[Y]} \leq \frac{X}{Y} \leq (1+3\delta)\frac{\mathbf{E}[X]}{\mathbf{E}[Y]}$. ◀

▶ **Theorem 2.** *Let* $\theta_1 = \min\left(\frac{g^2}{2(g+1)^2}, \frac{p_d}{3}\right)$ *and* $\theta_2 = \max(g+1, 8)$. *Suppose* $4c_A < L$ *and* $\frac{4}{3}p_s + p_d < 1$. *Then, for* $\delta \in (0, 1/5)$*:*

1.

$$\Pr[|\hat{p}_d - p_d| \geq 4\delta] \leq 6k\exp\left(-\frac{\delta^2\mathbf{E}[D]}{3k}\right) + 3k\exp\left(-\frac{\delta^2\mathbf{E}[N]}{3k}\right) + 6\exp\left(-\frac{\delta^2\mathbf{E}[P]^2}{72Le^{2(4\theta_2+1)}}\right).$$

2.

$$\Pr[|\hat{p}_s - p_s| \geq 12\delta] \leq 9k\exp\left(-\frac{\delta^2\mathbf{E}[N]}{3k}\right) + 6k\exp\left(-\frac{\delta^2\mathbf{E}[D]}{3k}\right) + 6\exp\left(-\frac{\delta^2\mathbf{E}[P]^2}{72Le^{2(4\theta_2+1)}}\right).$$

3.

$$\Pr[|\hat{g}-g| \geq (5+g)\delta] \leq 6k\exp\left(-\frac{\delta^2\mathbf{E}[D]}{3k}\right)+3k\exp\left(-\frac{\delta^2\mathbf{E}[N]}{3k}\right)+6\exp\left(-\frac{\delta^2\mathbf{E}[P]^2}{72Le^{2(4\theta_2+1)}}\right)+4\exp\left(-\theta_1 L\delta^2\right).$$

Proof. In Section 4, we proved the concentration bound for $\hat{p}_d$; here, we will prove the bound for $\hat{p}_s$ and g. Recall that $P = L' - 4c'_A$, $Q = kN + D$, $R = k(L - 4c_A)N$, and $T = 4kN + D$. Let us assume our variables are indeed close to their means, i.e. for all $X \in \{P, Q, R, T\}$, $(1-\delta)\,\mathbf{E}[X] \leq X \leq (1+\delta)\,\mathbf{E}[X]$ for $\delta \in (1, 1/5)$. Because we assume that $\frac{4}{3}p_s + p_d < 1$ and $4c_A < L$, we use (5) to obtain that $\mathbf{E}[P] > 0$; in addition, Lemma 1 gives $\mathbf{E}[Q], \mathbf{E}[R], \mathbf{E}[T] > 0$. Therefore, we can apply Lemma 7 to get

$$\frac{PQ-R}{T} \leq (1+4\delta)\frac{\mathbf{E}[P]\mathbf{E}[Q]}{\mathbf{E}[T]} - (1-4\delta)\frac{\mathbf{E}[R]}{\mathbf{E}[T]} \quad \leq \frac{\mathbf{E}[P]\mathbf{E}[Q] - \mathbf{E}[R]}{\mathbf{E}[T]} + 4\delta\frac{\mathbf{E}[P]\mathbf{E}[Q] + \mathbf{E}[R]}{\mathbf{E}[T]}$$

and, similarly, $\frac{PQ-R}{T} \geq \frac{\mathbf{E}[P]\mathbf{E}[Q]-\mathbf{E}[R]}{\mathbf{E}[T]} - 4\delta\frac{\mathbf{E}[P]\mathbf{E}[Q]+\mathbf{E}[R]}{\mathbf{E}[T]}$. Multiplying by $\frac{-3}{L-4c_A}$ and using the definition of (13), we obtain that $p_s - 4\delta W \leq \hat{p}_s \leq p_s + 4\delta W$ with $W = 3\frac{\mathbf{E}[P]\mathbf{E}[Q]+\mathbf{E}[R]}{(L-4c_A)\mathbf{E}[T]}$. Plugging in the values given by Lemma 1, we obtain that $W \leq 3$. Hence, $|\hat{p}_s - p_s| \leq 12\delta$. As in the proof for $\hat{p}_d$, the result follows by a union bound using the individual concentration bounds for P (Lemma 6), Q, R, and T (Lemma 4).

To obtain the bound for $\hat{g}$, recall that $\hat{g} = \frac{L'}{L} - 1 + \hat{p}_d$. From Lemma 5, for any $\delta \in (0, 1)$

$$\Pr[|L' - \mathbf{E}[L']| \geq \delta(g + p_d)L] \leq 4e^{-\theta_1 L\delta^2},$$

where $\theta_1 = \min\left(\frac{g^2}{2(g+1)^2}, \frac{p_d}{3}\right)$. Then,

$$\Pr\left[\left|\frac{L'}{L} - \frac{\mathbf{E}[L']}{L}\right| \geq \delta(g+1)\right] \leq \Pr\left[\left|\frac{L'}{L} - \frac{\mathbf{E}[L']}{L}\right| \geq \delta(g+p_d)\right] \leq 4e^{-\theta_1 L\delta^2}.$$

Combining this with Lemma 12, we get

$$\begin{aligned}\Pr[|\hat{g} - g| \geq (5+g)\delta] &= \Pr\left[\left|\frac{L'}{L} - \frac{\mathbf{E}[L']}{L} + \hat{p}_d - p_d\right| \geq (5+g)\delta\right] \\ &\leq \Pr\left[\left|\frac{L'}{L} - \frac{\mathbf{E}[L']}{L}\right| \geq (1+g)\delta\right] + \Pr\left[|\hat{p}_d - p_d| \geq 4\delta\right] \\ &\leq 4e^{-\theta_1 L\delta^2} + \Pr\left[|\hat{p}_d - p_d| \geq 4\delta\right].\end{aligned}$$

Plugging in the bounds on $\Pr\left[|\hat{p}_d - p_d| \geq 4\delta\right]$ from part 2 of this theorem gives the result. ◀

C Proofs of helper lemmas

In this section, we provide for completeness some of the proofs from Appendix A. To prove Lemma 11, we will first need the following fact.

▶ **Lemma 14.** *Let a, b be fixed positive real numbers. If X is (a, b)-sub-exponential, for $\lambda \in \mathbb{R}$ such that $|\lambda e^{ab+1}| < 1/2$ we have $\mathbf{E}[e^{\lambda X}] \leq e^{2(\lambda e^{a+1} b)^2}$.*

Proof. Using the Taylor expansion for the exponential function, the fact that $\mathbf{E}[X] = 0$, and standard inequality $p! \geq (\frac{p}{e})^p$, we deduce that

$$\mathbf{E}[e^{\lambda X}] = 1 + \lambda \mathbf{E}[X] + \sum_{p=2}^{\infty} \frac{\lambda^p \mathbf{E}[|X|^p]}{p!} = 1 + \sum_{p=2}^{\infty} \left(\frac{\lambda e}{p}\right)^p \mathbf{E}[|X|^p]. \tag{29}$$

Observe that for integer $p \geq 1$, we have the identity $\mathbf{E}[|X|^p] = \int_0^\infty \Pr[|X|^p \geq u] du$, and with the change of variable $u = bt^p$, we obtain:

$$\mathbf{E}[|X|^p] = \int_0^\infty \Pr[|X|^p \geq bt^p](bp)t^{p-1} dt \leq abp \int_0^\infty e^{-t} t^{p-1} du = abp\Gamma(p) \leq abpp^p \leq (e^{ab} p)^p;$$

Γ denotes the Gamma function and we use the fact that $\Gamma(x) \leq x^x$ for all $x \geq 1$. Plugging this bound into (29), we obtain:

$$\mathbf{E}[e^{\lambda X}] \leq 1 + \sum_{p=2}^{\infty} (\lambda e^{ab+1})^p = 1 + \frac{(\lambda e^{ab+1})^2}{1 - \lambda e^{ab+1}} \leq 1 + 2(\lambda e^{ab+1})^2 \leq e^{2(\lambda e^{ab+1})^2}$$

provided $|\lambda e^{ab+1}| < 1/2$. ◀

Proof of Lemma 11. Let $\lambda > 0$. Markov's inequality and the independence of the X_is imply that:

$$\Pr[X \geq t] = \Pr\left[e^{\lambda X} \geq e^{\lambda t}\right] \leq e^{-\lambda t} \mathbf{E}[e^{\lambda X}] = e^{-\lambda t} \prod_{i=1}^{n} \mathbf{E}[e^{\lambda X_i}].$$

If λ is such that $|\lambda e^{ab+1}| < 1/2$, Lemma 14 implies that $\mathbf{E}[e^{\lambda X_i}] \leq e^{2(\lambda e^{ab+1})^2}$ and thus

$$\Pr[X \geq t] \leq e^{-\lambda t + 2n(\lambda e^{ab+1})^2}.$$

For ease of notation set $y = e^{ab+1}$, and let us optimize for λ. The minimum of the parabola $f(\lambda) = -\lambda t + 2y^2 n \lambda^2$ in $[0, 1/(2y)]$ aligns with is global minimum at $\lambda = \frac{t}{4y^2 n}$ when $\frac{t}{4y^2 n} \leq \frac{1}{2y}$; the minimum value of f is $\frac{t^2}{8ny^2}$ in this case. Otherwise, if $\frac{t}{4y^2 n} > \frac{1}{2y}$, the the minimum occurs at the boundary point $\frac{1}{2y}$ and has value $\frac{-t}{2y} + \frac{n}{2} \geq \frac{-t}{2y}$. Combining these two bounds:

$$\Pr[X \geq t] \leq e^{-\min\left(\frac{t^2}{8ne^2(ab+1)}, \frac{t}{2e^{ab+1}}\right)}.$$

Repeating the same argument, we obtain the same bound for $\Pr[-X \geq t]$ and the result follows from a union bound. ◀

Proof of Lemma 12. First observe that $|\sum X_i| \leq \sum |X_i|$ and therefore $|\sum X_i| \geq \sum t_i$ implies that $\sum |X_i| \geq \sum t_i$. Hence $\Pr[|\sum X_i| \geq \sum t_i] \leq \Pr[\sum |X_i| \geq \sum t_i]$. Second, $\sum |X_i| \geq \sum t_i$ implies that there is at least one i such that $|X_i| \geq t_i$. Therefore, $\Pr[\sum |X_i| \geq \sum t_i] \leq \Pr[|X_1| \geq t_1 \cup \ldots \cup |X_n| \geq t_n]$. Third, by the union bound, $\Pr[|X_1| \geq t_1 \cup \ldots \cup |X_n| \geq t_n] \leq \sum \Pr[|X_i| \geq t_i]$. Combining the three observations in sequences, we get the result. ◀