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 nonmutated sequences. In this work, we consider a mutation model that uses 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 rate, deletion rate, and average insertion length. We provide statistics of k-mers under this model and theoretical guarantees via concentration inequalities, ensuring correctness under reasonable conditions. Empirical evaluations on simulated evolution of genomic sequences confirm our theoretical findings, demonstrating that accounting for indel 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/mahmudhera/estimate_rates_using_mutation_model.git. The results presented in this manuscript can be reproduced using the code available here: github.com/mahmudhera/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-mer, mutation rate, indel, alignment-free, estimation, substitution, insertion, deletion

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 [21]. 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 & Henikoff mined ungapped blocks of conserved proteins to improve sensitivity for more diverged sequences [7]. 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 complexity pipelines (e.g. quadratic-time) increasingly impractical. Consequently, alignment-free techniques that summarize sequences by inexpensive statistics have become indispensable [17, 22]. These approaches most commonly utilize k-mer sets and sketches thereof. Popular tools such as Mash [11], Skmer [13] and more recent sketch-corrected frameworks like Sylph [16] and FracMinHash-based methods [8, 15] 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 [12]. 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 [11]. Recent work has quantified how k-mer based statistics are affected by substitutions and are also used to estimate substitution-only mutation rates [8], 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 [19]. Such indels cause substitution-only k-mer-based methods to underperform, as disruption of k-mer content by indel events affects at least k k-mers, often leading to overestimates of mutation rates [3, 8].

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 d under a model that explicitly incorporates single-nucleotide substitutions, deletions, and geometrically-distributed insertions. This mutation model is closely related to the "indel channel" model of [14]. Starting from elementary counts of unmutated and single-deletion k-mers, we derive algebraic expressions for p_s, p_d, and d and prove sub-exponential concentration bounds that guarantee a strong form of consistency as the sequence length grows as detailed in our main contribution: Theorem 10. Simulations on synthetic and real bacterial genomes demonstrates that modeling indels yields markedly more accurate distance estimates than substitution-only approaches. The remainder of this paper is organized as follows: Section 2 defines our mutation model, Section 3 derives basic statistical properties, Section 4 presents our estimators, Section 5 provides theoretical guarantees about these estimators, Section 6 details our implementation, Section 7 presents experimental results, and Section 8 concludes with a discussion of implications and future directions.

2 The mutation model

We first describe the mutation model under consideration. Let S be a string over the alphabet $\{A, C, G, T\}$, and let L be the number of characters in S. Let S_i denote the i-th character in S where $1 \leq i \leq L$. We define three operations on a string S as follows:

- Substitute(i): select a character $c \in \{A, C, G, T\} \setminus \{S_i\}$ uniformly at random and replace S_i by c.
- Delete(i): remove the character S_i.
- Stay(i): do nothing.
- Insert(i, s): insert string s between S_i and S_{i+1} if $i < L$. If $i = L$, append s at the end of S instead.

The mutation process takes as input a string S and three parameters p_s, p_d, and d, where $0 \leq p_s, p_d < 1$, $p_s + p_d < 1$, and $d \geq 0$. For every position in S, it does the following:

1. For each i, let a_i be the operation to be performed at position i. Then, a_i = Substitute(i) with probability p_s, a_i = Delete(i) with probability p_d, or a_i = Stay(i) with the remaining probability $1 - p_s - p_d$. (These operations are not performed at this point, but only recorded.)
2. For each i, let $l_i \geq 0$ be a sample from a geometric distribution with mean d. Then generate a random string $Q^{(i)}$ of length exactly l_i by drawing each character in $Q^{(i)}$ from $\{A, C, G, T\}$ independently and uniformly at random. This is equivalent to sampling a uniformly random string among all strings of length l_i with characters in $\{A, C, G, T\}$. If $l_i = 0$, $Q^{(i)}$ is the empty string.
3. Go through each position in S. For position i, execute the corresponding operation a_i, and then execute the operation Insert(i, $Q^{(i)}$).
4. Return the resulting S'.

3 Preliminary statistics of the mutation process

In this section, we define several quantities whose statistics we use to design an estimator of the parameters of the mutation process. 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 d, and a mutated string S' is returned.

▶ **Lemma 1.** *Let L' be the length of S'. Then*

$$\mathbf{E}[L'] = L\,(1 + d - p_d), \tag{1}$$
$$\mathbf{Var}[L'] = L\Big(d(d+1) + p_d(1 - p_d)\Big).$$

▶ **Lemma 2.** *Let f_A and $f_A{}'$ be the number of 'A's in S and S', respectively. Then,*

$$\mathbf{E}[f_A{}'] = f_A(1 - p_s - p_d) + \frac{p_s(L - f_A)}{3} + \frac{dL}{4}. \tag{2}$$

We can analogously obtain expressions for the expectations of the corresponding counts of 'C's, 'G's, and 'T's in S', denoted by $f_C{}'$, $f_G{}'$, and $f_T{}'$ respectively.

Next, we say that a k-mer has no mutations if (a) for all k characters, the mutation process picks a Stay($\cdot$) action, and if (b) for all $k - 1$ intermediate positions of the k-mer, the mutation process does not insert anything. Let $\mathcal{N}$ be the number of such k-mers. Let $\mathcal{K}$ be the number of all k-mers in S.

▶ **Lemma 3.** $\mathbf{E}[\mathcal{N}] = \mathcal{K}\,(1 - p_s - p_d)^{\,k}\,\frac{1}{(d+1)^{\,k-1}}$.

The last quantities we will use correspond to the number of k-mers with a single kind of mutation. Let $\mathcal{S}$ be the number of k-mers with a single substitution, and no other mutations, let $\mathcal{D}$ be the number of k-mers with a single deletion and no other mutations, and let $\mathcal{I}$ be the number of k-mers with a single insertion and no other mutations.

▶ **Lemma 4.** *The expectations of $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{I}$ are given by:*

$$\mathbf{E}[\mathcal{S}] = \mathcal{K}\, k\, (1 - p_s - p_d)^{k-1}\, p_s\, \frac{1}{(d+1)^{\,k-1}}. \tag{3}$$

$$\mathbf{E}[\mathcal{D}] = \mathcal{K}\, k\, (1 - p_s - p_d)^{k-1}\, p_d\, \frac{1}{(d+1)^{\,k-1}}. \tag{4}$$

$$\mathbf{E}[\mathcal{I}] = \mathcal{K}\, (k-1)\, (1 - p_s - p_d)^{k}\, \frac{d}{(d+1)^{\,k}}. \tag{5}$$

4 Estimating p_s, p_d and d

We describe next our estimators for the parameters of the mutation model p_s, p_d, and d. We derive estimators for these rates based on the statistics defined in Section 3. Recall that S is the known input string to the mutation process, and S' is the resulting random string or observation.

Our mutation model assumes $p_s + p_d < 1$, and $d \geq 0$. In addition, we assume that $L \geq k$ which implies that $\mathbf{E}[\mathcal{N}] > 0$. In reality, strings are typically much longer than k-mer sizes and therefore this is a reasonable assumption. We will also assume that $p_d > 0$, which implies that $\mathbf{E}[\mathcal{D}] > 0$. We will handle the $p_d = 0$ case separately.

From Lemmas 3 and 4 we get

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

Reorganizing (1), (2) and (6), we have the following system of linear equations with variables p_s, p_d, and d.

$$-p_d + d = \frac{\mathbf{E}[L']}{L} - 1 \tag{7}$$

$$\frac{L - 4f_A}{3}\, p_s - f_A\, p_d + \frac{L}{4} d = \mathbf{E}[f_A{}'] - f_A \tag{8}$$

$$p_s + \left(\frac{k\mathbf{E}[\mathcal{N}]}{\mathbf{E}[\mathcal{D}]} + 1\right) p_d = 1. \tag{9}$$

Solving this system of equations, we obtain that:

$$p_s = 3\,\frac{k\,\mathbf{E}[\mathcal{N}]\big(\mathbf{E}[L'] - L + 4f_A - 4\mathbf{E}[f_A{}']\big) + \mathbf{E}[\mathcal{D}]\big(\mathbf{E}[L'] - 4\mathbf{E}[f_A{}']\big)}{\big(4f_A - L\big)\big(\mathbf{E}[\mathcal{D}] + 4k\mathbf{E}[\mathcal{N}]\big)}.$$

Given S and S', we can observe L', $f_A{}'$, $\mathcal{N}$, and $\mathcal{D}$. By replacing the expectations above with these observations, we obtain our estimators for the parameters of mutation model. That is,

$$\hat{p}_s = 3\,\frac{k\,\mathcal{N}\big(L' - L + 4f_A - 4f_A{}'\big) + \mathcal{D}\big(L' - 4f_A{}'\big)}{\big(4f_A - L\big)\big(\mathcal{D} + 4k\mathcal{N}\big)}, \tag{10}$$

$$\hat{p}_d = \frac{\mathcal{D}\,[3(L' - 4f_A{}') + (L - 4f_A)]}{(L - 4f_A)\,(\mathcal{D} + 4k\mathcal{N})}, \text{and} \tag{11}$$

$$\hat{d} = \frac{L'}{L} - 1 + \frac{\mathcal{D}\,[3(L' - 4f_A{}') + (L - 4f_A)]}{(L - 4f_A)\,(\mathcal{D} + 4k\mathcal{N})}. \tag{12}$$

We briefly comment on our choice of estimators for the mutation model parameters, 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-mers with a single insertion, deletion, or substitution (i.e., $\mathcal{I}$, $\mathcal{S}$, and $\mathcal{D}$). These quantities contain enough information to estimate the mutation parameters (specifically, by solving the non-linear system of equations given by Lemma 4), and 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 7, 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 analysis.

▶ Remark 5. We briefly describe estimators for the case when $p_d = 0$, which corresponds to a strictly simpler mutation process without deletions. In this case, we would observe $\mathcal{D} = 0$. When this happens, (9) becomes undefined, and therefore, we can only use (7) and (8). Plugging in $p_d = 0$ into (7) and (8), we obtain the estimators:

$$\hat{p}_s = \frac{3(f_A{}' - f_A + L/4 - L'/4)}{L - 4f_A}, \tag{13}$$

$$\hat{p}_d = 0, \text{and} \tag{14}$$

$$\hat{d} = \frac{L'}{L} - 1. \tag{15}$$

We note that we did not encounter this case in our simulation experiments. Nevertheless, we include this case for completeness.

5 Theoretical results for $\hat{p}_s$, $\hat{p}_d$, and $\hat{d}$

In this section, we provide theoretical guarantees for the estimators from Section 4; i.e., (10), (11), and (12). In particular, we show that our estimates of p_s, p_d, and d are not only consistent estimators, but also that they are tightly concentrated in a symmetric interval around the true values of p_s, p_d, and d. We provide the bias analysis for $\hat{p}_s$ by proving asymptotically tight concentration bounds for $\mathcal{N}$, $\mathcal{D}$, and $L' - 4f_A{}'$. We then describe how to obtain analogous results for $\hat{p}_d$ and $\hat{d}$.

The concentration bounds for $\mathcal{N}$ and $\mathcal{D}$ stem from the fact that, from the perspective of the mutation process, k-mers located more than k positions apart are independent of each other.

▶ **Lemma 6.** *For any $\delta \in (0, 1)$:*

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

▶ **Lemma 7.** *For any $\delta \in (0, 1)$:*

$$\Pr\left[\left|\mathcal{D} - \mathbf{E}[\mathcal{D}]\right| \geq \delta\mathbf{E}[\mathcal{D}]\right] \leq 2k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{D}]}{3k}\right\}.$$

We prove a similar result for $P = L' - 4f_A{}'$ by noting that P can expressed as a sum of independent sub-exponential random variables. Our concentration bound for P is as follows.

▶ **Lemma 8.** *Let $K_1 = 1/\min\{\frac{1}{d+1}, \frac{1}{8}\}$. Then, there exist absolute constants $c_1, c_2 > 0$ such that the following holds for any $\delta > 0$:*

$$\Pr\left[|P - \mathbf{E}[P]| \geq 3\delta\right] \leq 2\exp\left\{-\frac{\delta^2}{25f_A}\right\} + 2\exp\left\{-\frac{\delta^2}{3(L - f_A)}\right\} + 2\exp\left\{-c_1\min\left(\frac{\delta^2}{c_2{}^2K_1{}^2}, \frac{\delta}{c_2K_1}\right)\right\}.$$

Finally, we prove that L' is also strongly concentrated around $\mathbf{E}[L']$.

▶ **Lemma 9.** *For any $\delta \in (0,1)$:*

$$\Pr\left[\left|L' - \mathbf{E}[L']\right| \leq \delta(Ld + Lp_d)\right] \;\geq\; 2\exp\left\{-\frac{L\delta^2 d^2}{2(d+1-\delta d)(d+1)}\right\} + 2\exp\left\{-\frac{Lp_d\delta^2}{3}\right\}. \quad (16)$$

We can piece the results of these lemmas together and prove the following result.

▶ **Theorem 10.** *Suppose $4f_A \leq L$ and $\frac{4}{3}p_s + p_d \leq 1$. Then, for sufficiently small $\delta > 0$, there exists constants $c_1, c_2 > 0$ such that*

$$\Pr[|\hat{p}_s - p_s| \geq 4\delta] \leq 6k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{N}]}{3k}\right\}4k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{D}]}{3k}\right\}+$$
$$+\,2\exp\left\{-\frac{\delta^2\mu_P^2}{225f_A}\right\} + 2\exp\left\{-\frac{\delta^2\mu_P^2}{27(L-f_A)}\right\} + 2\exp\left\{-c_1\min\left(\frac{\delta^2\mu_P^2}{{c_2}^2}, \frac{\delta\mu_P}{c_2}\right)\right\}.$$

The requirement that $4f_A \leq L$ in this theorem does not restrict generality: aside from equal nucleotide frequency (where the estimators are naturally invalid), at least one character $c \in \{A, C, T, G\}$ must satisfy $4f_c \leq L$. In addition, the assumption that $\frac{4}{3}p_s + p_d \leq 1$ holds when p_s and p_d are small (e.g., $p_s \leq 1/4$ and $p_d \leq 1/4$ suffices) which is the case most frequently encountered in practice.

We can obtain analogous results for $\hat{p}_d$ and $\hat{d}$ using the same approach. As such, Theorem 10 is our central theoretical result, establishing that the estimators developed in Section 4 are sound under the assumed conditions. Before discussing our implementation and presenting the experimental results, we comment on the error probability in Theorem 10. This probability is small when each of the terms in the sum are small. Since each of these terms decays (at least) exponentially with δ times an expectation, they will all generally be small. In particular:

- The probabilistic guarantees for the estimators are asymptotically formulated. Consequently, as the lengths of the strings S and S' increase, the probability of errors decreases, and the estimators become more accurate.
- Assuming $p_s + p_d < 1$, and assuming the strings S and S' are sufficiently long, $\mathbf{E}[\mathcal{N}]$ is always positive, and the first error term in the sum is sufficiently small.
- Assuming $p_s + p_d < 1$, and assuming the strings S and S' are sufficiently long, if $p_d = 0$, then we would observe $\mathcal{D} = 0$. In such a case, the theorem does not apply and we would need to use (13) to (15) to compute $\hat{p}_s$ and $\hat{d}$, as detailed in Remark 5. We note that in such a case, the fact that $\hat{p}_s$ is close to p_s with high probability follows easily from Lemma 8.

6 Implementation details

Our estimators for the three rates require counting the number of k-mers with single deletion, $\mathcal{D}$, and the number of k-mers with no mutation, $\mathcal{N}$. Counting these two quantities is not trivial and can be considered an interesting problem in and of itself. We therefore implemented an ad hoc solution to count $\mathcal{D}$ and $\mathcal{N}$ given the two strings S and S'. The steps to count $\mathcal{D}$ and $\mathcal{N}$ are as follows.

We start by extracting all k-mers in S, and building a de Bruijn graph using these k-mers. 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. We then use the alignment to look at all windows of length k, and count $\mathcal{D}$ and $\mathcal{N}$ 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 is to scan beyond the context of k characters – and unitigs are a natural way to do this.

In our implementation, we use semi-global alignment to allow part of one unitig to be aligned with the other. We build the de Bruijn graph and the unitigs using cuttlefish [10], and compute alignments using edlib [18]. The core goal of our implementation of these steps was not to make it efficient, but rather to obtain a working solution. As such, we leave an efficient way to compute $\mathcal{D}$ and $\mathcal{N}$ an open research question.

7 Experiments and results

In this section, we present a series of experiments to evaluate the performance of the estimators detailed in Section 4. 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 7.1 through 7.3 explore the sensitivity of the estimators with respect to these parameters. In Sections 7.4 and 7.5, we estimate mutation rates across a wide range of known rate combinations. And finally, in Section 7.6, we demonstrate that our estimated substitution rate outperforms estimates obtained under a substitution-only mutation model. For the experiments in Sections 7.1 through 7.4, the original sequence is a randomly generated synthetic sequence. In these cases, we compute the number of k-mers containing a single deletion and the number of unmutated k-mers directly from the known mutation process. For the experiments in Sections 7.5 and 7.6, we use real reference genomes as the original sequences. In these cases, the two types of k-mer counts are obtained using the steps described in Section 6.

7.1 Sensitivity of the estimators to k-mer lengths

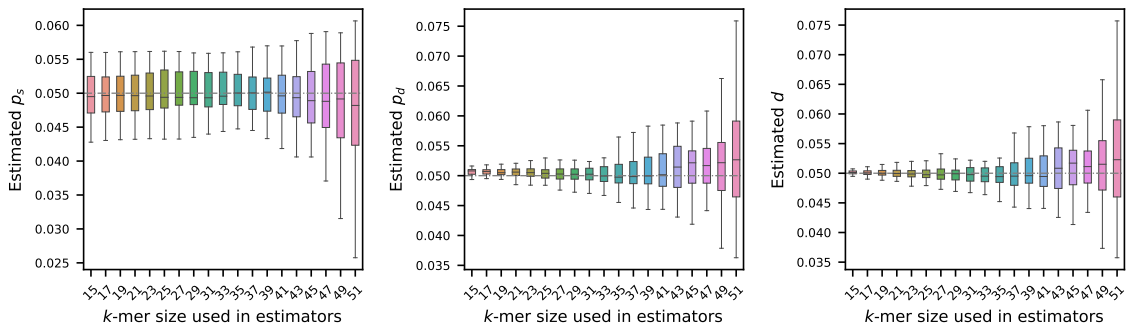

Figure 1 *Effect of k-mer length on mutation rate estimation (true rates set to 0.05).* A synthetic genome of 1 million nucleotides, mutated genomes were generated by setting p_s, p_d, and $d = 0.05$ – shown by the gray dashed horizontal line. Estimated rates were then computed using a range of k-mer 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.

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', 20% 'C', 20% 'G', and 20% 'T'. From this reference, we simulated 30 mutated sequences, independently from each other, using the mutation model described in Section 2. For each of these mutated sequences, we

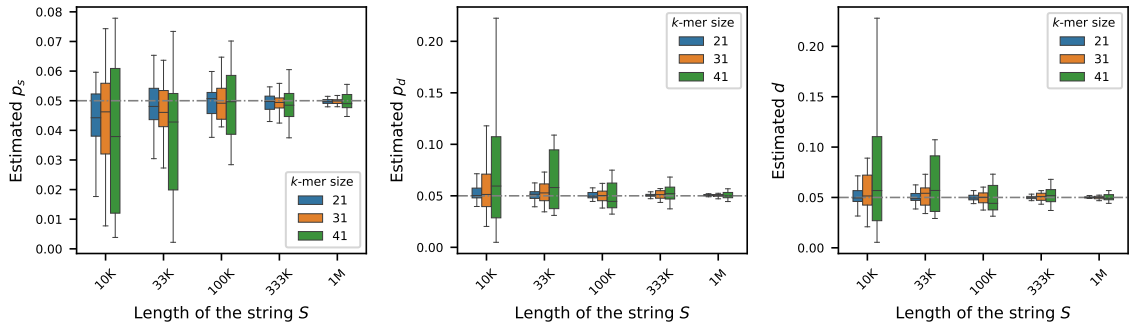

Figure 2 *Effect of sequence length on mutation rate estimation (true rates set to 0.05).* For synthetic genomes of varying lengths, mutated genomes were generated by setting p_s, p_d, and $d = 0.05$ – shown by the gray dashed horizontal line. Estimated rates were computed using three k-mer sizes: 21, 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.

estimated mutation rates using the estimators defined in Section 4 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 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 10 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-mers with a single substitution $\mathcal{S}$ in the estimators may correct this behavior, but additional analyses are required to substantiate this hypothesis.

7.2 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 sampling. The nucleotide composition of each sequences was fixed, with the frequencies of 'A', 'C', 'G', and 'T' set to 0.3, 0.2, 0.2, and 0.2, respectively. 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 d to 0.05. We then estimated the mutation rates using the estimators outlined in Section 4 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 10, which states that the associated error is asymptotically vanishing in L, the length of the string S: as L increases, the number of k-mers $\mathcal{K}$ increases, and therefore the probability of error decreases, leading to a more precise estimation.

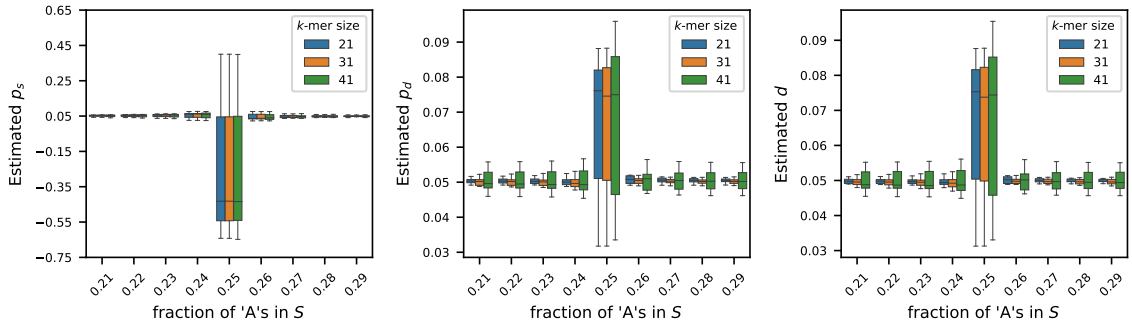

Figure 3 *Effect of nucleotide composition on mutation rate estimation (true rates set to 0.05).* For synthetic genomes of length $L = 1$M, the fraction of 'A's is varied from 0.21 to 0.29, and the frequencies of 'G's, 'C's, and 'T's are set equally. For each setting, the mutated string S' was generated by setting $p_s, p_d, d = 0.05$. 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 $f_A \approx L/4$, in which case the estimators become unstable – as predicted by Theorem 10.

7.3 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 d to 0.05. We then estimated the mutation rates using the estimators described in Section 4 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 $\hat{d}$ work reasonably well to estimate the true rates when the frequency of 'A' characters, f_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 10: when $L - 4f_A = 0$, $\mu_P = \mathbf{E}[P] = 0$, and the probabilistic guarantees become unbounded. We only get meaningful guarantees of consistency when $4f_A$ is strictly smaller than L. While Theorem 10 does not guarantee consistency when $4f_A > L$, this does not restrict generality, as explained in Section 4, and as demonstrated by the estimators' performances when the fraction of 'A's is larger than 1/4.

7.4 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 20% each for 'C', 'G', and 'T'. 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 d 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 4 for each of these mutated genomes.

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

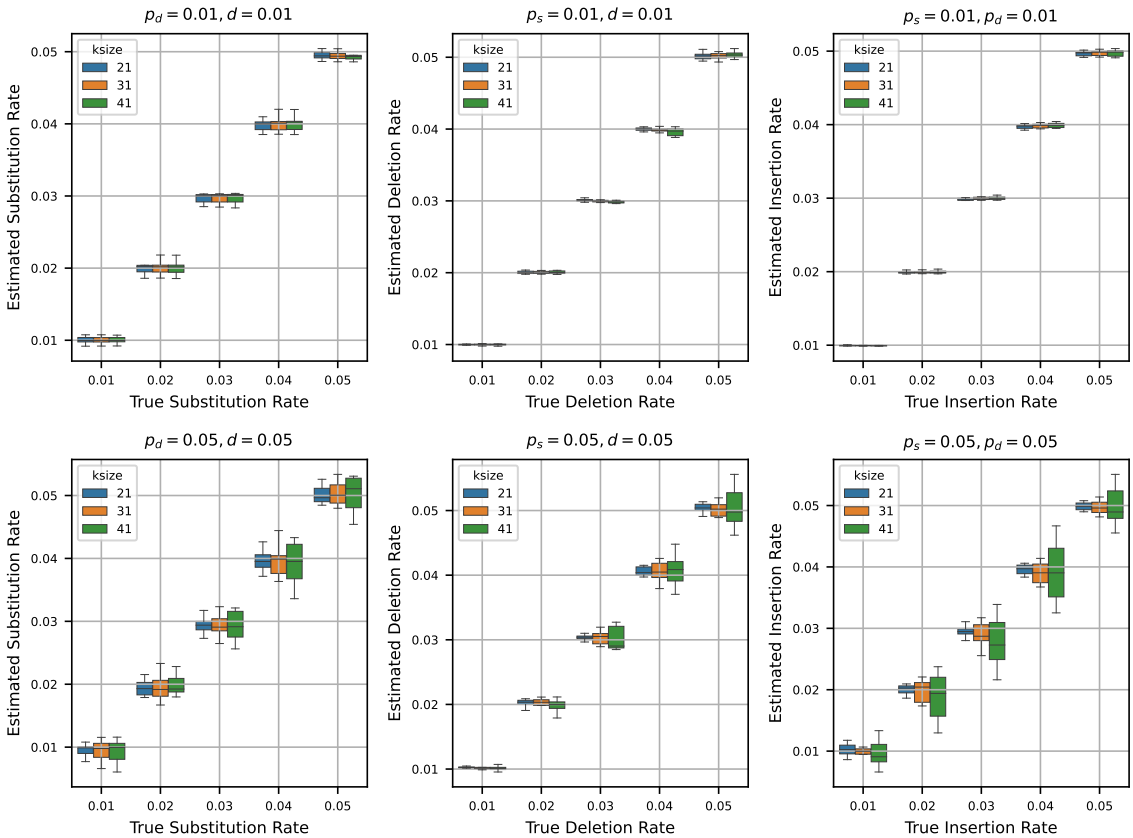

Figure 4 *Estimated mutation rates versus true values, where the original string is a synthetic sequence*. Each subplot corresponds to a case where two mutation rates are fixed (either at 0.01 or 0.05) and the third is varied from 0.01 to 0.05. Each boxplot shows the variability in estimation across 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.

the true rate varies from 0.01 to 0.05. We repeat this process independently for p_s, p_d, and d.

- **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 7.1, where longer k-mers resulted in decreased precision of the estimators.

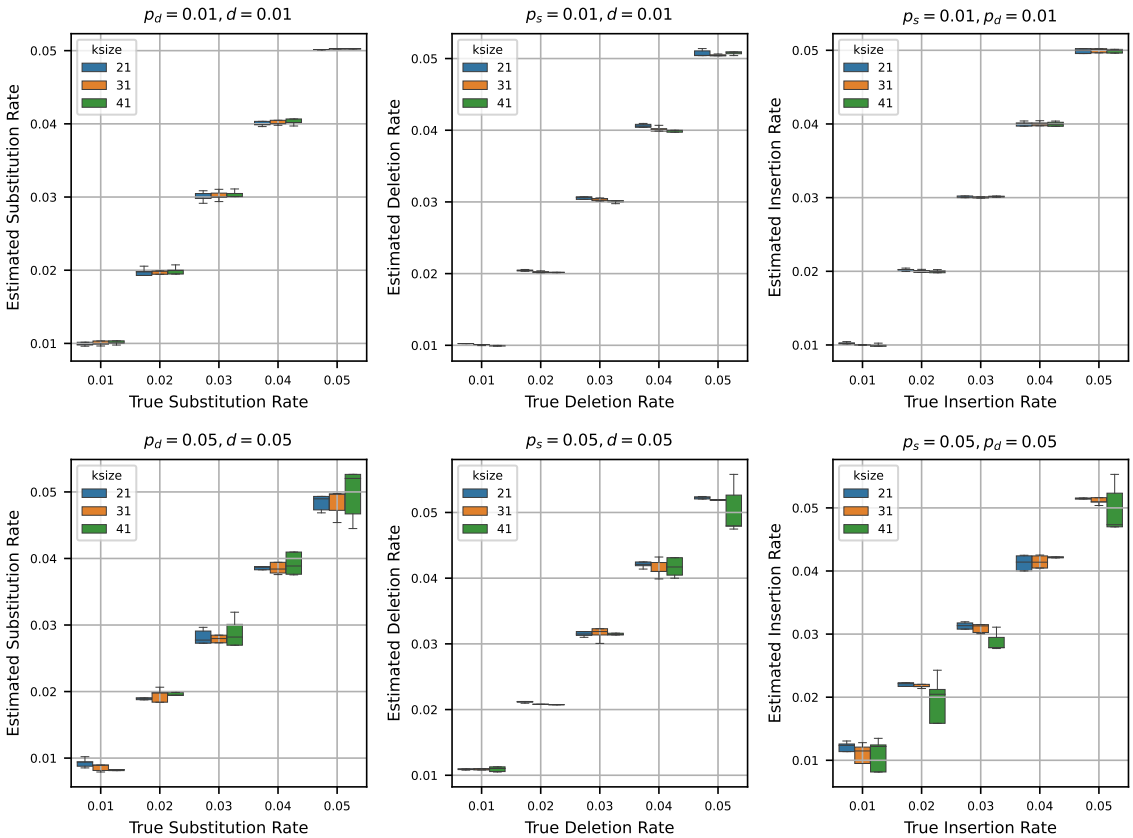

Figure 5 *Estimated mutation rates versus true values, where the original string is the reference genome of Staphylococcus (length 2.8 million).* Each subplot corresponds to a case where two mutation rates are fixed (either at 0.01 or 0.05) and the third varies from 0.01 to 0.05. Each boxplot shows the 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.

7.5 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 d from 0.01 to 0.05. Similar to Section 7.4, we generated 10 independent mutated sequences for each combination of p_s, p_d, and d to capture stochastic variability. We then estimated the mutation rates using the estimators outlined in Section 4 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 counting k-mers with a single deletion ($\mathcal{D}$) or no mutation ($\mathcal{N}$) in a real

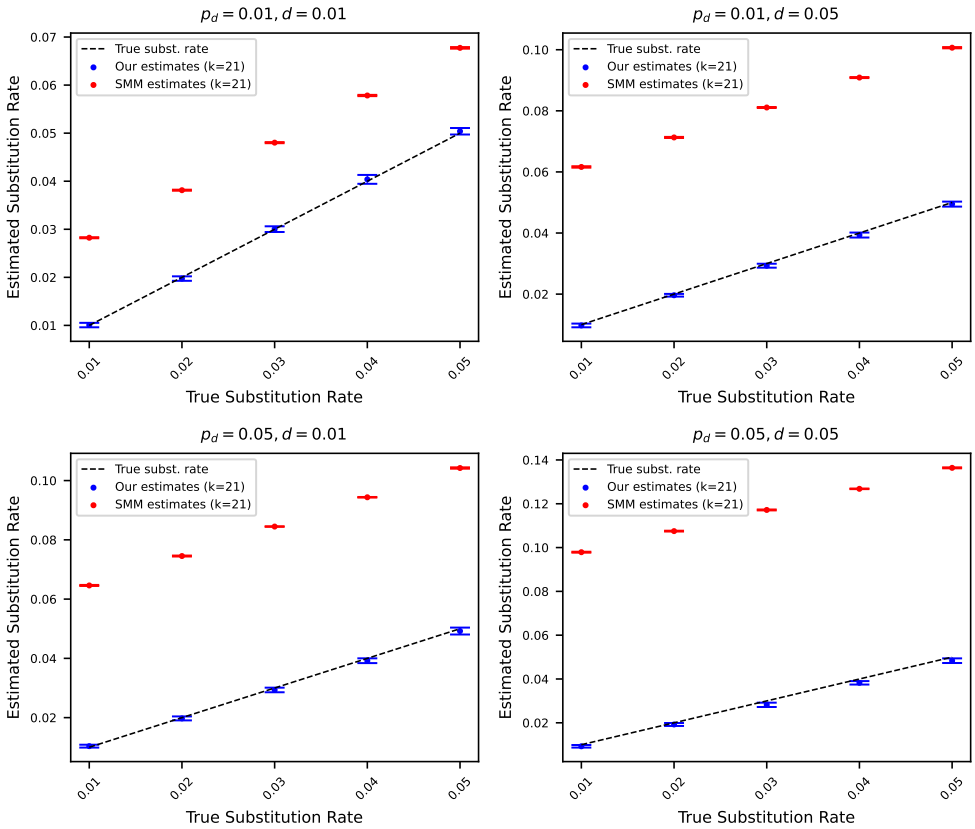

Figure 6 *Estimated substitution rates versus true substitution rates, where the original string is the reference genome of Staphylococcus (length 2.8 million). Rates were estimated using Equation* (10) *and using a simple mutation model (SMM) that only considers substitution.* Each subplot corresponds to a case where p_d and d 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.

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.

7.6 Comparison with substitution rates estimated using simple mutation model

We conclude the experiments section by contrasting the substitution rates estimated using Equation (10) with substitution rates estimated considering a simple mutation model. We use the statistics of k-mers developed in a recent work [1] 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 7.5. In Figure 6, we show the estimated substitution rates using our estimators

(Equation (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.

8 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 unbiased, purely algebraic estimators for the three elementary mutation rates: p_s, p_d, and d; proved relatively tight sub-exponential concentration bounds that cumulatively guarantee consistency as sequence length grows. 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 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 unmutated and single-deletion k-mers sufficed to show practical accuracy of our estimators, this raises and interesting open problem: estimating the number of these unmutated and single-deletion k-mers in efficiently for large scale data sets.

Several directions invite further investigation. First, incorporating the count $\mathcal{S}$ of single-substitution k-mers 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 [8]) 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 a relatively complex mutation process parameters 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. 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.
2. Pafnuty Lvovich Chebyshev. Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées*, 12:177–184, 1867.
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. 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.
8. 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.
9. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
10. 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.
11. 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.
12. 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.
13. 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.
14. Tizian Schulz and Paul Medvedev. ESKEMAP: exact sketch-based read mapping. *Algorithms for Molecular Biology*, 19(1):19, 2024.
15. Jim Shaw and Yun William Yu. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods*, 20(11):1661–1665, 2023.
16. Jim Shaw and Yun William Yu. Rapid species-level metagenome profiling and containment estimation with sylph. *Nature Biotechnology*, pages 1–12, 2024.
17. 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.
18. Martin Šošić and Mile Šikić. Edlib: a c/c++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 2017.
19. 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.
20. Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
21. 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.
22. Andrzej Zielezinski, Susana Vinga, Jonas Almeida, and Wojciech M Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology*, 18:1–17, 2017.

Supplementary Materials

The full version of this manuscript, including proofs in an extended abstract, is available on the BioRxiv: `https://tinyurl.com/5d8fw2y2`.

A List of notations used in the manuscript

In this section, we list all the notations used in the manuscript.

Table 1 List of notations and their meanings

Notation	Meaning
E	Expectation
Var	Variance
Pr	Probability of an event
inf	Infimum
exp	Exponent to e
min	Minimum
max	Maximum
p_s	Substitution rate
p_d	Deletion rate
d	Average insertion length
S	The original string
S'	The mutated string
L	Length of (number of characters in) S
L'	Length of (number of characters in) S'
f_A, f_C, f_G, f_T	Number of 'A', 'C', 'G', 'T' in S, respectively. Here, $f_A + f_C + f_G + f_T = L$.
f_A', f_C', f_G', f_T'	Number of 'A', 'C', 'G', 'T' in S', respectively. Here, $f_A' + f_C' + f_G' + f_T' = L'$.
P	$P = L' - 4f_A'$
$\mathcal{K}$	Number of k-mers in S. Here, $\mathcal{K} = L - k + 1$.
$\mathcal{K}'$	Number of k-mers in S'. Here, $\mathcal{K}' = L' - k + 1$.
i	A position in the string
S_i	The character in S at position i
$Q^{(i)}$	The string inserted after S_i, by the mutation process
l_i	Length of $Q^{(i)}$
D_i	Indicator that S_i was deleted by the mutation process
A_i	Indicator that stay action was chosen at position i
B_i	Indicator that S'_i = 'A', where $S_i \neq$ 'A'
C_i	Number of 'A's in $Q^{(i)}$
$\mathcal{K}_i$	The k-mer in S starting at position i, ending at position $i + k - 1$
N_i	Indicator that the mutation process did not mutate $\mathcal{K}_i$
E_i	Indicator that a single substitution, and no other mutation was introduced in $\mathcal{K}_i$
Δ_i	Indicator that a single deletion, and no other mutation was introduced in $\mathcal{K}_i$
$\mathcal{N}$	Number of k-mers that were not mutated by the mutation process
$\mathcal{S}$	Number of k-mers where a single substitution and no other mutation was introduced by the mutation process
$\mathcal{D}$	Number of k-mers where a single deletion and no other mutation was introduced by the mutation process

Notation	Meaning
$\mathcal{I}$	Number of k-mers where a single insertion and no other mutation was introduced by the mutation process
$\mathcal{L}_{\text{ins}}$	Total number of characters inserted by the mutation process.
$\mathcal{L}_{\text{del}}$	Total number of characters deleted by the mutation process.
$\mathcal{A}$	$\sum_{i:S_i=\text{'A'}} A_i$
$\mathcal{B}$	$\sum_{i:S_i\neq\text{'A'}} B_i$
C	$\sum_{i=1}^{L} C_i$
M_i	$\sum_{j=0}^{c-1} N_{i+kj}$, where $\mathcal{K} = ck$ (assuming $\mathcal{K}$ is a multiple of k)

B Preliminaries

This section gathers several definitions and standard concentration inequalities that we use in our proofs. For a random variable X with finite non-zero variance σ^2, Chebyshev's inequality [2] says that for any real value r

$$\Pr\Big[|X - \mathbf{E}[X]| \geq r\sigma\Big] \leq \frac{1}{r^2}.$$

Sums of independent random variables exhibit sharper concentration, as captured by the following standard inequalities. The first one is the Chernoff bound for sums of independent and identically distributed Bernoulli random variables [4].

▶ **Lemma 11.** *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:*

$$\Pr(X \geq (1+\delta)\mu) \leq e^{-\delta^2\mu/(2+\delta)}, \qquad 0 \leq \delta, \tag{17}$$

$$\Pr(X \leq (1-\delta)\mu) \leq e^{-\delta^2\mu/2}, \qquad 0 < \delta < 1, \tag{18}$$

$$\Pr(|X-\mu| \geq \delta\mu) \leq 2e^{-\delta^2\mu/3}, \qquad 0 < \delta < 1. \tag{19}$$

A more general inequality that applies to sums of independent and bounded random variables is Hoeffding's inequality [9].

▶ **Lemma 12.** *Let $X_1, \ldots, X_n$ be a sequence of independently and identically distributed random variables taking values in $[a, b]$. Let $X = \sum_{i=1}^{n} X_i$, and $\mathbf{E}[X] = \mu$. Then, for any $\delta > 0$:*

$$\Pr\left(|X-\mu| \geq \delta\right) \leq 2e^{-\frac{2\delta^2}{n(b-a)^2}}. \tag{20}$$

We will analyze sums of *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 [20, Chapter 2.7].

▶ **Definition 13.** *A random variable X such that $\mathbf{E}[X] = 0$ is called sub-exponential if either of the following equivalent properties holds:*

1. *There exists $K_1 > 0$ such that for all $t \geq 0$:*

$$\Pr\Big[|X| \geq t\Big] \leq 2\exp\{-t/K_1\}.$$

2. *There exists $K_2 > 0$ such that*

$$\mathbf{E}\Big[e^{|X|/K_2}\Big] \leq 2. \tag{21}$$

The constants K_1 and K_2 differ by an absolute multiplicative constant. The smallest K_2 such that Equation (21) holds is called the *sub-exponential norm of* X and is denoted by $\|X\|_{\psi_1}$.

For sums of sub-exponential random variables, the following version of Bernstein's inequality holds; see [20, Theorem 2.8.1].

▶ **Lemma 14.** *Let $X_1, \ldots, X_n$ be a sequence of independent, mean-zero, sub-exponential random variables, and let $X = \sum_{i=1}^{n} X_i$. Then, there exists an absolute constant $c > 0$ such that for every $t \geq 0$:*

$$\Pr\left[|X| \geq t\right] \leq 2\exp\left[-c\min\left(\frac{t^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right].$$

For the reader's convenience, our notations are summarized in Table 1.

C Proofs of theorems

▶ **Lemma 1.** *Let L' be the length of S'. Then*

$$\mathbf{E}[L'] = L\,(1 + d - p_d), \tag{1}$$
$$\mathbf{Var}[L'] = L\Big(d(d+1) + p_d(1 - p_d)\Big).$$

Proof. We note that the mutation process goes through every position in S, and removes the character S_i with probability p_d, and inserts a string $Q^{(i)}$ of length l_i. Only these two operations change the number of characters in S – substitute and stay actions do not change the number of total characters.

Let D_i be an indicator variable defined as follows:

$$D_i = \begin{cases} 1 & \text{if } S_i \text{ was removed,} \\ 0 & \text{otherwise;} \end{cases}$$

for $1 \leq i \leq L$. Using the D_i's, we can express L' as follows:

$$L' = L + \sum_{i=1}^{L} l_i - \sum_{i=1}^{L} D_i \ . \tag{22}$$

We can compute the expected length of S' by using the linearity of expectation. We note that l_i is a random variable that follows the geometric distribution with a mean of d, and therefore, $\mathbf{E}[l_i] = d$. The indicator variables D_i's take a value 1 with probability p_d, and therefore, $\mathbf{E}[D_i] = p_d$. Using these facts and the linearity of expectation, Equation (1) is proved.

We next compute the variance of L'. Note that D_i's are indicator variables, and therefore, $\mathbf{Var}[D_i] = p_d(1 - p_d)$. We also observe that l_i's are geometrically distributed with mean d where $d \geq 0$, and therefore, $\mathbf{Var}[l_i] = d(d+1)$. Note that all l_i's and all D_i's are mutually independent. Using the fact that D_i's and l_i's are independent proves the variance and completes the proof. ◀

▶ **Lemma 2.** *Let f_A and $f_A{}'$ be the number of 'A's in S and S', respectively. Then,*

$$\mathbf{E}[f_A{}'] = f_A(1 - p_s - p_d) + \frac{p_s(L - f_A)}{3} + \frac{dL}{4}. \tag{2}$$

Proof. We can quantify $f_A{}'$ by introducing a few random variables.

Let A_i be an indicator defined as follows:

$$A_i = \begin{cases} 1 & \text{if } a_i \text{ was a stay operation,} \\ 0 & \text{otherwise;} \end{cases}$$

defined only for the positions i such that S_i = 'A'. We note that the probability of $A_i = 1$ is $1 - p_s - p_d$, and therefore, $\mathbf{E}[A_i] = 1 - p_s - p_d$.

Let us also define B_i as follows:

$$B_i = \begin{cases} 1 & \text{if } a_i \text{ was a substitute operation, and } S'_i = \text{'A'}, \\ 0 & \text{otherwise;} \end{cases}$$

defined only for the positions i such that $S_i \neq$ 'A'. It is straightforward to argue that the probability of $B_i = 1$ is $p_s/3$, and therefore, $\mathbf{E}[B_i] = p_s/3$.

And finally, let C_i be the number of As in $Q^{(i)}$, defined for $i = 1$ to L. Note that every character in $Q^{(i)}$ can be an 'A' with probability 1/4. Also, note that the total number of characters in $Q^{(i)}$ is l_i, which is drawn from a geometric distribution with mean d. Therefore, it is straightforward to derive that $\mathbf{E}[C_i] = d/4$.

Using these indicators, we can quantify $f_A{}'$ as follows:

$$f_A{}' = \sum_{i:S_i=\text{'A'}} A_i + \sum_{i:S_i\neq\text{'A'}} B_i + \sum_{i=1}^{L} C_i. \tag{23}$$

Plugging the expected values of A_i's, B_i's and C_i's in Equation (23) completes the proof.

◀

▶ **Lemma 3.** $\mathbf{E}[\mathcal{N}] = \mathcal{K}\,(1 - p_s - p_d)^{\,k}\,\frac{1}{(d+1)^{\,k-1}}$.

Proof. For a k-mer to have no mutation, the mutation model would have to choose the stay operation for all of its k positions, the probability of which is $(1 - p_s - p_d)^k$. Additionally, for all the $k - 1$ insert positions, the mutation model would need to choose strings of length zero. The probability of selecting 0 from a geometric distribution with mean $d \geq 0$ is $1/(d + 1)$. Therefore, the probability that 0 was sampled for all these $k - 1$ positions is $1/(d + 1)^{k-1}$.

Using these facts, we can write that the probability that a k-mer does not have a mutation is $(1 - p_s - p_d)^k/(d + 1)^{k-1}$. Let N_i be the indicator variable defined as follows:

$$N_i = \begin{cases} 1 & \text{if } k\text{-mer starting at position } i \text{ has no mutation,} \\ 0 & \text{otherwise;} \end{cases}$$

for $1 \leq i \leq \mathcal{K}$. Let $\mathcal{N}$ be the number of k-mers that have no mutation. Then, we can write the following:

$$\mathcal{N} = \sum_{i=1}^{\mathcal{K}} N_i \ . \tag{24}$$

Using the fact that $\mathbf{E}[N_i] = (1 - p_s - p_d)^k/(d + 1)^{k-1}$, we apply linearity of expectation in the above equation, which completes the proof.

◀

▶ **Lemma 4.** *The expectations of $\mathcal{S}$, $\mathcal{D}$, and $\mathcal{I}$ are given by:*

$$\mathbf{E}[\mathcal{S}] = \mathcal{K}\,k\,(1 - p_s - p_d)^{k-1}\,p_s\,\frac{1}{(d+1)^{\,k-1}}. \tag{3}$$

$$\mathbf{E}[\mathcal{D}] = \mathcal{K}\,k\,(1 - p_s - p_d)^{k-1}\,p_d\,\frac{1}{(d+1)^{\,k-1}}. \tag{4}$$

$$\mathbf{E}[\mathcal{I}] = \mathcal{K}\,(k-1)\,(1 - p_s - p_d)^{k}\,\frac{d}{(d+1)^{\,k}}. \tag{5}$$

Proof. Let E_i be an indicator random variable defined as follows for $1 \leq i \leq \mathcal{K}$:

$$E_i = \begin{cases} 1 & \text{if } k\text{-mer starting at position } i \text{ has a single substitution and no other mutation,} \\ 0 & \text{otherwise.} \end{cases}$$

We note that for a k-mer to have introduced with only a single substitution and no other mutations, the mutation process needs to insert 0 length strings in all the $k-1$ insert positions, and other than a single substitution position, for all other $k-1$ characters in the k-mer, the mutation process needs to select a stay operation. The single substitution operation could be any of the k positions. Using these facts, we can write that $\mathbf{E}[S_i] = k\,(1-p_s-p_d)^{k-1}p_s/(d+1)^{k-1}$. Let $\mathcal{S}$ be the number of total k-mers with a single substitution and no other mutation. Then, we have the following:

$$\mathcal{S} = \sum_{i=1}^{\mathcal{K}} E_i \ .$$

Equation (3) follows from the above using linearity of expectation. The expectations of $\mathcal{D}$ and $\mathcal{I}$ are also derived following a similar argument.

◀

▶ **Lemma 6.** *For any $\delta \in (0,1)$:*

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

Proof. Recall the definition of $\mathcal{N}$ in terms of N_i indicator random variables, introduced in the proof of Lemma 3. We showed that that $\Pr[N_i = 1] = (1-p_s-p_d)^k/(d+1)^{k-1}$. We observe that N_i's are not independent of each other. However, the dependency only exists within a window of k k-mers in each direction. In other words, N_i is only dependent on N_j' s, where $j = \max\{0, i-k+1\}$ to $\min\{i+k-1, L\}$.

For ease of exposition, we treat only the case $\mathcal{K} \equiv 0 \mod k$. When $\mathcal{K} = ck + r$ with $0 < r < k$, we either pad the sequence with $k-r$ additional positions (alternatively trim it), or group the indices by residue class modulo k allowing the last block(s) to have size $c+1$ (so then $\mathbf{E}[M_i]$ takes value of cq or $(c+1)q$). In either case, the resulting concentration inequality remains unchanged. So without the loss of generality, we assume that $\mathcal{K}$, the number of k-mers in S, is a multiple of k. That is, $\mathcal{K} = ck$, where $c \in \mathbb{N}$. Using these notations, we define a set of random variables M_i as follows:

$$M_i = \sum_{j=0}^{c-1} N_{i+kj},$$

where $i = 1, 2, 3, \cdots, k$. We defined M_i's in a way such that each M_i is a sum of c independent Bernoulli random variables, each with success probability $\Pr[N_i = 1] = (1-p_s-p_d)^k/(d+1)^{k-1}$. For notational simplicity, let $q = \Pr[N_i = 1]$.

It follows that $\mathbf{E}[\mathcal{N}] = \mathcal{K}q$. Also, the expected value of M_i is given by the following.

$$\mathbf{E}[M_i] = cq = \frac{\mathcal{K}}{k}q,$$

where $i = 1, 2, 3, \cdots, k$. Additionally, as each M_i is a sum of iid Bernoulli random variables, we can use the Chernoff inequality in Equation (19) for M_i. Doing so gives us the following result:

$$\Pr\left[\left|M_i - \frac{\mathcal{K}}{k}q\right| \geq \delta\frac{\mathcal{K}}{k}q\right] \leq 2\exp\left\{-\frac{\mathcal{K}\delta^2 q}{3k}\right\},$$

which holds for $i = 1, 2, 3, \cdots, k$. Taking a union bound of these k inequalities completes the proof. ◀

▶ **Lemma 7.** *For any $\delta \in (0, 1)$:*

$$\Pr\left[\left|\mathcal{D} - \mathbf{E}[\mathcal{D}]\right| \geq \delta\mathbf{E}[\mathcal{D}]\right] \leq 2k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{D}]}{3k}\right\}.$$

Proof. Using the same terminology used to prove Equation (4), we observe that every k-mer can be mutated with a single deletion and no other mutation with a probability of $k\left(1-p_s-p_d\right)^{k-1} p_d \frac{1}{\left(d+1\right)^{k-1}}$. Let us define the following sequence of indicator variables:

$$\Delta_i = \begin{cases} 1 & \text{if the mutation process introduces only a single deletion to the } k\text{-mer starting at position } i, \\ 0 & \text{otherwise;} \end{cases}$$

for $i = 1$ to L. We note that the following holds:

$$\mathbf{E}[\Delta_i] = \Pr[\Delta_i = 1] = k\left(1 - p_s - p_d\right)^{k-1} p_d \frac{1}{\left(d+1\right)^{k-1}},$$

for $i = 1$ to L. Using these indicator variables, we have the following:

$$\mathcal{D} = \sum_{i=1}^{L} \Delta_i.$$

We observe that Δ_i's have the same dependency nature as N_i's as we observed in the proof of Lemma 6: Δ_i only depends on Δ_j' s, where $j = \max\{0, i-k+1\}$ to $\min\{i+k-1, L\}$. Using the same k-way split completes the proof.

◀

▶ **Lemma 8.** *Let $K_1 = 1/\min\{\frac{1}{d+1}, \frac{1}{8}\}$. Then, there exist absolute constants $c_1, c_2 > 0$ such that the following holds for any $\delta > 0$:*

$$\Pr\left[|P - \mathbf{E}[P]| \geq 3\delta\right] \leq 2\exp\left\{-\frac{\delta^2}{25 f_A}\right\} + 2\exp\left\{-\frac{\delta^2}{3(L-f_A)}\right\} + 2\exp\left\{-c_1\min\left(\frac{\delta^2}{{c_2}^2{K_1}^2}, \frac{\delta}{c_2 K_1}\right)\right\}.$$

Proof. We characterize P by considering three separate groups of characters: the 'A' characters in S, the non-'A' characters in S, and the characters that are inserted into S (by the mutation process).

First, we observe that an 'A' in S contributes -3 to P if it was untouched by the mutation process, contributes 1 to P if it was substituted into a non-'A' character by the mutation process, and contributes 0 to P if it was deleted by the mutation process. To capture these cases, we define X_i as follows:

$$X_i = \begin{cases} -3 & \text{with probability } 1 - p_s - p_d, \\ 1 & \text{with probability } p_s, \text{ and} \\ 0 & \text{with probability } p_d; \end{cases}$$

defined for the i's such that $S_i =$ 'A'. We recall that there are f_A such characters in S.

Next, we observe that a non-'A' character in S contributes 1 to P if it was untouched by the mutation process or substituted into another non-'A' character, contributes -3 to P if it is substituted into an 'A' character by the mutation process, and contributes 0 to P if it was deleted by the mutation process. To capture these cases, we define Y_i as follows:

$$Y_i = \begin{cases} 1 & \text{with probability } 1 - p_s/3 - p_d, \\ -3 & \text{with probability } p_s/3, \text{ and} \\ 0 & \text{with probability } p_d; \end{cases}$$

defined for the i's such that $S_i \neq$ 'A'. We recall that there are $L - f_A$ such characters in S.

And finally, we define Z_i for $i = 1$ to L as follows:

$$Z_i = l_i - 4C_i. \tag{25}$$

We can therefore characterize P in terms of X_i, Y_i, and Z_i as follows:

$$P = \sum_{i:S_i=\text{'A'}} X_i + \sum_{i:S_i\neq\text{'A'}} Y_i + \sum_{i=1}^{L} Z_i.$$

For notational simplicity, let us define the following.

$$X = \sum_{i:S_i=\text{'A'}} X_i,$$
$$Y = \sum_{i:S_i\neq\text{'A'}} Y_i, \quad \text{and}$$
$$Z = \sum_{i=1}^{L} Z_i.$$

Then, we have

$$P = X + Y + Z.$$

We first observe that the expectations of X_i, Y_i, and Z_i are easily obtained.

$$\mathbf{E}[X_i] = p_s - 3(1 - p_s - p_d) = -3 + 4p_s + 3p_d,$$
$$\mathbf{E}[Y_i] = 1 - p_s/3 - p_d - p_s = 1 - \frac{4}{3}p_s - p_d, \quad \text{and}$$
$$\mathbf{E}[Z_i] = 0.$$

Here, we have used the fact that $\mathbf{E}[l_i] = d$ and $\mathbf{E}[C_i] = d/4$. Using these three expected values, we can easily obtain the expectations of X, Y, Z, and P.

$$\mu_X = \mathbf{E}[X] = f_A(-3 + 4p_s + 3p_d),$$
$$\mu_Y = \mathbf{E}[Y] = (L - f_A)\Big(1 - \frac{4}{3}p_s - p_d\Big),$$
$$\mu_Z = \mathbf{E}[Z] = 0, \quad \text{and}$$
$$\mu_P = \mathbf{E}[P] = (L - 4f_A)\Big(1 - \frac{4}{3}p_s - p_d\Big).$$

We obtain a strong concentration of X around μ_X by applying Hoeffding's inequality shown in (20). Doing this gives us the following result, which holds for any $\delta > 0$:

$$\Pr(|X - \mu_X| \geq \delta) \leq 2e^{-\frac{\delta^2}{25 f_A}}. \tag{26}$$

A similar application of Hoeffding's inequality for Y gives us the following result, which also holds for any $\delta > 0$.

$$\Pr(|Y - \mu_Y| \geq \delta) \leq 2e^{-\frac{\delta^2}{3(L - f_A)}}. \tag{27}$$

Recall that Z is the sum of Z_i's. We prove that Z is strongly concentrated around its mean 0 by first proving that the Z_i's are sub-exponential random variables.

▶ **Lemma 15.** *For each Z_i, $i = 1$ to L, the following holds for all $t \geq 0$:*

$$\Pr[|Z_i| \geq t] \leq 2e^{-t/K_1},$$

where $\frac{1}{K_1} = \min\{\frac{1}{d+1}, \frac{1}{8}\}$.

Proof. Recall the definition of Z_i's in Equation (25). We observe that l_i follows the Geometric distribution with mean d, and C_i follows Binomial distribution, where the success probability is $1/4$, and the number of trials is l_i. Therefore, we have the following:

$$l_i \sim \text{Geom}(p),$$

and

$$C_i \sim \text{Bin}(l_i, 1/4).$$

Here, $p = 1/(d + 1)$, and the Geometric distribution supports 0, 1, 2, etc.

We also observe that $\mathbf{E}[Z_i] = d - d = 0$. We now calculate the probability that $|Z_i|$ deviates from its mean 0 by more than t as follows:

$$\begin{aligned}
\Pr[|Z_i| \geq t] &= \sum_{y \geq 0} \Pr[l_i = y] \Pr[|\text{Bin}(y, 1/4) - y/4| \geq t/4] \\
&\leq 2p \sum_{y \geq 0} (1 - p)^y e^{\frac{-t^2}{8y}} \\
&\leq 2p \left(\sum_{y=0}^{\lfloor x \rfloor} (1 - p)^y e^{-\frac{t^2}{8y}} + \sum_{y=\lfloor x \rfloor + 1}^{\infty} (1 - p)^y e^{-\frac{t^2}{8y}} \right)
\end{aligned}$$

For $y \leq t$, we have $\frac{t^2}{8y} \geq \frac{t}{8}$, so $e^{-\frac{t^2}{8y}} \leq e^{-t/8}$ and we have the following.

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

For $y > t$, we have that the following holds.

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

Putting the bounds together, we have that the following holds.

$$\Pr[|Z_i| \geq t] \leq 2\left(e^{-t/8} + e^{-pt}\right) \leq 2e^{-t/K_1}$$

where $1/K_1 = \min\{p, 1/8\}$. Setting $p = 1/(d+1)$ completes the proof.

◀

Lemma 15 shows that Z_i's are sub-exponential random variables. As Z_i's are mean-zero sub-exponential random variables, we can apply Bernstein's inequality for a sum of Z_i's, which gives us the following result.

▶ **Corollary 16.** *There exist absolute constants $c_1, c_2 > 0$ such that the following holds for any $\delta > 0$*

$$\Pr\left[|Z| \geq \delta\right] \leq 2\exp\left\{-c_1 \min\left(\frac{\delta^2}{c_2{}^2 K_1{}^2}, \frac{\delta}{c_2 K_1}\right)\right\} \tag{28}$$

Taking a union bound of Equations 26, 27 and 28 completes the proof. ◀

▶ **Lemma 9.** *For any $\delta \in (0,1)$:*

$$\Pr\left[\left|L' - \mathbf{E}[L']\right| \leq \delta(Ld + Lp_d)\right] \geq 2\exp\left\{-\frac{L\delta^2 d^2}{2(d+1-\delta d)(d+1)}\right\} + 2\exp\left\{-\frac{Lp_d\delta^2}{3}\right\}. \tag{16}$$

Proof. We prove that L' is strongly concentrated around its expected value by proving concentration bounds for the individual pieces of Equation (22). Specifically, we prove that $\sum l_i$ and $\sum D_i$ are strongly concentrated around their respective expected values.

For the sake of notational simplicity, let us define $\mathcal{L}_{\text{ins}} = \sum_{i=1}^{L} l_i$, and $\mathcal{L}_{\text{del}} = \sum_{i=1}^{L} D_i$. Then, L' is simply

$$L' = L + \mathcal{L}_{\text{ins}} - \mathcal{L}_{\text{del}}. \tag{29}$$

We first prove a strong concentration of $\mathcal{L}_{\text{ins}}$ around its expected value, and later prove the same result for $\mathcal{L}_{\text{del}}$. We finally use these results to show that L' is strongly concentrated around its mean.

We observe that every l_i is sampled from a geometric distribution with mean $d \geq 0$. This geometric distribution has support for values 0, 1, 2, 3, etc. (opposed to supporting positive values only). Another interpretation of this Geometric distribution is that it assigns probability to the number of failures before encountering a success – where each trial is a Bernoulli trial with success probability $1/(d+1)$. We also observe that the expected value of l_i is d, and therefore, the expected value of $\mathcal{L}_{\text{ins}}$ is Ld.

We aim to determine $\Pr[\mathcal{L}_{\text{ins}} \geq (1+\delta)Ld]$. We notice that $\mathcal{L}_{\text{ins}} \geq (1+\delta)Ld$ is equivalent to observing at least $(1+\delta)Ld$ failures before observing L successes from a series of Bernoulli trials

with success probability $1/(d+1)$. Using the same argument, this is equivalent to observing at most L successes in $(1+\delta)Ld + L$ Bernoulli trials with success probability $1/(d+1)$. Let Z_i be a Bernoulli random variable with success probability $1/(d+1)$ for $i \geq 1$. Then, we have the following.

$$\Pr[\mathcal{L}_{\text{ins}} \geq (1+\delta)Ld] = \Pr[\sum_{i=1}^{(1+\delta)Ld+L} Z_i \leq L] \tag{30}$$

The expected number of success events in this collection of Bernoulli trials is $\mu = \frac{(1+\delta)Ld+L}{d+1}$. By reorganizing Equation (30), we get the following result.

$$\Pr[\mathcal{L}_{\text{ins}} \geq (1+\delta)Ld] = \Pr[\sum_{i=1}^{(1+\delta)Ld+L} Z_i \leq (1-\epsilon)\mu] \tag{31}$$

where $\epsilon = \frac{\delta d}{\delta d+d+1}$. Using Chernoff inequality for the sum of independent Bernoulli random variables from Equation (18), the probability in Equation (31) is at most $\exp\{-\epsilon^2\mu/2\}$. Using the values of μ and ϵ as defined before, we have the following result.

$$\Pr[\mathcal{L}_{\text{ins}} \geq (1+\delta)Ld] \leq \exp\{-\frac{L\delta^2 d^2}{2(d+1+\delta d)(d+1)}\} \tag{32}$$

Proceeding in an analogous manner, we can prove the probability for the other tail.

$$\Pr[\mathcal{L}_{\text{ins}} \leq (1-\delta)Ld] \leq \exp\{-\frac{L\delta^2 d^2}{2(d+1-\delta d)(d+1)}\} \tag{33}$$

Taking a union bound using the probabilities in Equation (32) and Equation (33), we have the following.

$$\Pr[|\mathcal{L}_{\text{ins}} - Ld| \leq \delta Ld] \geq 1 - 2\exp\{-\frac{L\delta^2 d^2}{2(d+1-\delta d)(d+1)}\} \tag{34}$$

We next prove that $\mathcal{L}_{\text{del}}$ is strongly concentrated around its mean. We recall that $\mathcal{L}_{\text{del}} = \sum_{i=1}^{L} D_i$, where each D_i for $i = 1$ to L is a Bernoulli trial with success probability p_d. Here, all D_i's are independent and identically distributed, and $\mathbf{E}[\mathcal{L}_{\text{del}}] = Lp_d$. Therefore, we can use the Chernoff bound in Equation (19) for $\mathcal{L}_{\text{del}}$, which gives us the following result.

$$\Pr[|\mathcal{L}_{\text{del}} - Lp_d| \leq \delta Lp_d] \quad \geq \quad 1 - 2\exp\{-\frac{Lp_d\delta^2}{3}\}. \tag{35}$$

Recall that $L' = L + \mathcal{L}_{\text{ins}} - \mathcal{L}_{\text{del}}$, as defined in Equation (29). Therefore, by using the strong concentration of $\mathcal{L}_{\text{ins}}$ and $\mathcal{L}_{\text{del}}$ around their respective expected values in Equation (34) and Equation (35), the proof is completed. ◀

▶ **Theorem 10.** *Suppose $4f_A \leq L$ and $\frac{4}{3}p_s + p_d \leq 1$. Then, for sufficiently small $\delta > 0$, there exists constants $c_1, c_2 > 0$ such that*

$$\Pr[|\hat{p}_s - p_s| \geq 4\delta] \leq 6k\exp\Big\{-\frac{\delta^2\mathbf{E}[\mathcal{N}]}{3k}\Big\}4k\exp\Big\{-\frac{\delta^2\mathbf{E}[\mathcal{D}]}{3k}\Big\}+$$
$$+ 2\exp\Big\{-\frac{\delta^2\mu_P^2}{225f_A}\Big\} + 2\exp\Big\{-\frac{\delta^2\mu_P^2}{27(L-f_A)}\Big\} + 2\exp\Big\{-c_1\min\Big(\frac{\delta^2\mu_P^2}{{c_2}^2}, \frac{\delta\mu_P}{c_2}\Big)\Big\}.$$

Proof. Assume $4f_A \leq L$ and that $\frac{4}{3}p_s + p_d \leq 1$. Let $Q = k\mathcal{N} + \mathcal{D}$, $R = k(L-4f_A)\mathcal{N}$, and $T = 4k\mathcal{N} + \mathcal{D}$. Observe that

$$\hat{p}_s = -3 \cdot \frac{PQ - R}{(L - 4f_A)T};$$

our proof will transfer concentration bounds for these quantities to one for $\hat{p}_s$.

First, by Lemma 6

$$\Pr\left[|\mathcal{N} - \mathbf{E}[\mathcal{N}]| \geq \delta\mathbf{E}[\mathcal{N}]\right] \leq 2k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{N}]}{3k}\right\}. \tag{36}$$

So if $4f_A \leq L$, we can multiply the inequality inside the probability by $k(L - 4f_A)$ to obtain that:

$$\Pr\left[|R - \mathbf{E}[R]| \geq \delta\mathbf{E}[R]\right] \leq 2k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{N}]}{3k}\right\}.$$

Similarly, from Lemmas 6 and 7 and a union bound, we have that for any $\delta \geq 0$

$$\Pr\left[|Q - \mathbf{E}[Q]| \geq \delta\mathbf{E}[Q]\right] \leq 2k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{N}]}{3k}\right\} + 2k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{D}]}{3k}\right\},$$

$$\Pr\left[|T - \mathbf{E}[T]| \geq \delta\mathbf{E}[T]\right] \leq 2k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{N}]}{3k}\right\} + 2k\exp\left\{-\frac{\delta^2\mathbf{E}[\mathcal{D}]}{3k}\right\}, \tag{37}$$

Moreover, since under the assumptions $\mathbf{E}[P] \geq 0$, by Lemma 8, there exists constants c_1 and c_2 such that

$$\Pr\Big[|P - \mathbf{E}[P]| \geq \delta\mathbf{E}[P]\Big] \leq 2\exp\left\{-\frac{\delta^2\mathbf{E}[P]^2}{225f_A}\right\} + 2\exp\left\{-\frac{\delta^2\mathbf{E}[P]^2}{27(L - f_A)}\right\} + 2\exp\left\{-c_1\min\left(\frac{\delta^2\mathbf{E}[P]^2}{{c_2}^2}, \frac{\delta\mathbf{E}[P]}{c_2}\right)\right\}.$$

Now, let us assume that the concentration bounds above do hold and that indeed:

$$(1-\delta)\,\mathbf{E}[P] \leq P \leq (1+\delta)\,\mathbf{E}[P], \tag{38}$$

$$(1-\delta)\,\mathbf{E}[Q] \leq Q \leq (1+\delta)\,\mathbf{E}[Q], \tag{39}$$

$$(1-\delta)\,\mathbf{E}[R] \leq R \leq (1+\delta)\,\mathbf{E}[R], \tag{40}$$

$$(1-\delta)\,\mathbf{E}[T] \leq T \leq (1+\delta)\,\mathbf{E}[T]. \tag{41}$$

Then,

$$\frac{(1-\delta)^2}{1+\delta}\frac{\mathbf{E}[P]\mathbf{E}[Q]}{\mathbf{E}[T]} \leq \frac{PQ}{T} \leq \frac{(1+\delta)^2}{1-\delta}\frac{\mathbf{E}[P]\mathbf{E}[Q]}{\mathbf{E}[T]}.$$

Note that here we rely on the assumption that $\mathcal{D} > 0$, which guarantees that $T > 0$.

We observe that $\frac{(1-\delta)^2}{1+\delta} \geq 1 - 4\delta$ and $\frac{(1+\delta)^2}{1-\delta} \leq 1 + 4\delta$ for small enough δ. Then, we have the following:

$$(1-4\delta)\frac{\mathbf{E}[P]\mathbf{E}[Q]}{\mathbf{E}[T]} \leq \frac{PQ}{T} \leq (1+4\delta)\frac{\mathbf{E}[P]\mathbf{E}[Q]}{\mathbf{E}[T]} \tag{42}$$

Similarly, by dividing Equation (40) by Equation (41), and noting that that $\frac{1-\delta}{1+\delta} \geq 1 - 3\delta$, and $\frac{1+\delta}{1-\delta} \leq 1 + 3\delta$ for small enough δ, we obtain:

$$(1-3\delta)\frac{\mathbf{E}[R]}{\mathbf{E}[T]} \leq \frac{1-\delta}{1+\delta}\cdot\frac{\mathbf{E}[R]}{\mathbf{E}[T]} \leq \frac{R}{T} \leq \frac{1+\delta}{1-\delta}\cdot\frac{\mathbf{E}[R]}{\mathbf{E}[T]} \leq (1+3\delta)\frac{\mathbf{E}[R]}{\mathbf{E}[T]}.$$

Combined with Equation (42) this yields that

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

Multiplying by $\frac{-3}{L-4f_A}$ we obtain that $p_s - 4\delta W \leq \hat{p}_s \leq p_s + 4\delta W$ with $W = 3\frac{\mathbf{E}[P]\mathbf{E}[Q]+\mathbf{E}[R]}{(L-4f_A)E[T]}$. It can be checked that $W \leq 2$, and the result then follows from a union bound.

◀

D Interesting but unused results

Herein, we describe some interesting observations and results that were uncovered in the process of finding our main results, but which ultimately were not used in any proofs or the like. These include additional statistics, concentration inequalities, and an alternate, polynomial-based mutation parameter estimation approach.

D.1 Estimating the mutation rates using a polynomial

Unlike the estimators shown in Section 4, we can estimate the mutation rates by using a different combination of observations. In this section, we assume that $\mathbf{E}[\mathcal{S}]$, $\mathbf{E}[\mathcal{D}]$, and $\mathbf{E}[\mathcal{I}]$ are all positive. This is equivalent to assuming that p_s, p_d, and d are all non-zero. We later handle the cases for $p_s = 0$, and/or $p_d = 0$, and/or $d = 0$ separately.

For the ease of algebraic manipulations, we first define the following.

$$\mathcal{S}_{\text{norm}} = \frac{\mathcal{S}}{k\,\mathcal{K}} \tag{43}$$

$$\mathcal{I}_{\text{norm}} = \frac{\mathcal{I}}{(k-1)\mathcal{K}} \tag{44}$$

$$\mathcal{D}_{\text{norm}} = \frac{\mathcal{D}}{k\,\mathcal{K}} \tag{45}$$

We first take the ratio of Equations 3 and 4 to get the following.

$$p_s = \frac{\mathbf{E}[\mathcal{S}_{\text{norm}}]}{\mathbf{E}[\mathcal{D}_{\text{norm}}]} p_d \tag{46}$$

We also take the ratio of Equations 4 and 5 to derive the following.

$$\frac{\mathbf{E}[\mathcal{D}_{\text{norm}}]}{\mathbf{E}[\mathcal{I}_{\text{norm}}]} = \frac{d+1}{d} \times \frac{p_d}{1 - \mathbf{E}[\mathcal{S}_{\text{norm}}]p_d/\mathbf{E}[\mathcal{D}_{\text{norm}}] - p_d} \tag{47}$$

By reorganizing, we get the following.

$$d + 1 = \frac{\mathbf{E}[\mathcal{D}_{\text{norm}}] - (\mathbf{E}[\mathcal{S}_{\text{norm}}] + \mathbf{E}[\mathcal{D}_{\text{norm}}])p_d}{\mathbf{E}[\mathcal{D}_{\text{norm}}] - (\mathbf{E}[\mathcal{S}_{\text{norm}}] + \mathbf{E}[\mathcal{D}_{\text{norm}}] + \mathbf{E}[\mathcal{I}_{\text{norm}}])p_d} \tag{48}$$

We next plug the values of $d + 1$ and p_s into Equation (4), and get the following polynomial after reorganizing.

$$\mathbf{E}[\mathcal{D}_{\text{norm}}]^k = p_d\Big(\mathbf{E}[\mathcal{D}_{\text{norm}}] - (\mathbf{E}[\mathcal{S}_{\text{norm}}] + \mathbf{E}[\mathcal{D}_{\text{norm}}] + \mathbf{E}[\mathcal{I}_{\text{norm}}])p_d\Big)^{k-1} \tag{49}$$

Given S and S', we can only observe $\mathcal{I}_{\text{norm}}$, $\mathcal{D}_{\text{norm}}$, and $\mathcal{I}_{\text{norm}}$. Using these observations, we can estimate p_d from Equation (49).

$$\mathcal{D}_{\text{norm}}{}^k = \hat{p}_d\Big(\mathcal{D}_{\text{norm}} - (\mathcal{S}_{\text{norm}} + \mathcal{D}_{\text{norm}} + \mathcal{I}_{\text{norm}})\hat{p}_d\Big)^{k-1} \tag{50}$$

Let $y = \mathcal{D}_{\text{norm}} - (\mathcal{S}_{\text{norm}} + \mathcal{D}_{\text{norm}} + \mathcal{I}_{\text{norm}})\hat{p}_d$. This results in the following polynomial:

$$y^k - \mathcal{D}_{\text{norm}}\, y^{k-1} + (\mathcal{S}_{\text{norm}} + \mathcal{D}_{\text{norm}} + \mathcal{I}_{\text{norm}})\mathcal{D}_{\text{norm}}{}^k = 0, \tag{51}$$

where

$$\hat{p_d} = \frac{\mathcal{D}_{\text{norm}} - y}{\mathcal{S}_{\text{norm}} + \mathcal{D}_{\text{norm}} + \mathcal{I}_{\text{norm}}}. \tag{52}$$

After solving for $\hat{p_d}$, we estimate the other two rates as follows.

$$\hat{p_s} = \frac{\mathcal{S}_{\text{norm}}}{\mathcal{D}_{\text{norm}}}\hat{p_d} \tag{53}$$

$$\hat{d} = \frac{(\mathcal{D}_{\text{norm}})\hat{p_d}}{\mathcal{D}_{\text{norm}} - (\mathcal{S}_{\text{norm}} + \mathcal{D}_{\text{norm}} + \mathcal{I}_{\text{norm}})\hat{p_d}} \tag{54}$$

We next handle the edge cases in which these equations do not work. Note that we assume that $1 - p_s - p_d > 0$. From Equation (4), we see that $p_d = 0$ if and only if $\mathbf{E}[\mathcal{D}] = 0$. Therefore, if we observe $\mathcal{D} = 0$, we estimate $\hat{p_d} = 0$. Using a similar argument, if we observe $\mathcal{S} = 0$, then we estimate $\hat{p_s} = 0$. And finally, if we observe $\mathcal{I} = 0$, then we estimate $\hat{d} = 0$.

D.2 Concentration of $f_A{}'$ around $\mathbf{E}[f_A{}']$

We next prove that $f_A{}'$ is strongly concentrated around its expected value. Recall the definition of $f_A{}'$ defined in Equation (23). For the sake of notational simplicity, we define the following.

$$\mathcal{A} = \sum_{i:S_i = \text{'A'}} A_i\ , \tag{55}$$

$$\mathcal{B} = \sum_{i:S_i \neq \text{'A'}} B_i\ , \text{and} \tag{56}$$

$$C = \sum_{i=1}^{L} C_i, \tag{57}$$

where A_i, B_i, and C_i are the same as introduced in Section 3. Using these notations, $f_A{}'$ is simply

$$f_A{}' = \mathcal{A} + \mathcal{B} + C. \tag{58}$$

D.2.1 Strong concentration of $\mathcal{A}$ around $\mathbf{E}[\mathcal{A}]$

We note that the A_i's in Equation (55) are iid Bernoulli random variables with success probability $1 - p_s - p_d$. We also note that number of 'A' characters in S is f_A. Therefore, we have the following.

$$\mathbf{E}[\mathcal{A}] = f_A(1 - p_s - p_d) \tag{59}$$

As the $\mathcal{A}$ is a sum of iid Bernoulli random variables, we can use the Chernoff bound in Equation (19) for $\mathcal{A}$, which gives us the following result.

$$\Pr\left[\left|\mathcal{A} - \mathbf{E}[\mathcal{A}]\right| \leq \delta f_A(1 - p_s - p_d)\right] \ \geq\ 1 - 2\exp\{-\frac{\delta^2 f_A(1 - p_s - p_d)}{3}\} \tag{60}$$

D.2.2 Strong concentration of $\mathcal{B}$ around $\mathbf{E}[\mathcal{B}]$

We next prove a concentration bound for the sum of B_i's, as defined in Equation (23). We note that B_i's here are iid Bernoulli random variables, each with success probability of $p_s/3$. We also note that the number of characters in S that are not 'A' is $L - f_A$. Therefore, we have the following result.

$$\mathbf{E}[\mathcal{B}] = (L - f_A)\frac{p_s}{3} \tag{61}$$

As $\mathcal{B}$ is simply a sum of iid Bernoulli random variables, we can use the same Chernoff inequality in Equation (19) for $\mathcal{B}$, which gives us the following result.

$$\Pr\left[\left|\mathcal{B} - \mathbf{E}[\mathcal{B}]\right| \leq \delta(L - f_A)\frac{p_s}{3}\right] \;\geq\; 1 - 2\exp\{-\frac{\delta^2(L - f_A)p_s}{9}\} \tag{62}$$

D.2.3 Strong concentration of C around $\mathbf{E}[C]$

We finally prove a concentration inequality for $\sum_{i=1}^{L} C_i$, as defined in Equation (23). Recall the result in Equation (34) – developed for a sum of iid geometric random variables, each with mean d. We prove a similar result for C by showing that each C_i for $i = 1$ to L is geometrically distributed with mean $d/4$.

We first recall that the probability mass function (PMF) for a geometric distribution with parameter p. If a random variable X follows a geometric distribution with probability parameter p where $0 < p \leq 1$, then the PMF is as follows:

$$\Pr[X = j] = p(1 - p)^j, \tag{63}$$

where $j = 0, 1, 2, 3, \cdots$, etc. The expected value of X is given by

$$\mathbf{E}[X] = \frac{1 - p}{p}. \tag{64}$$

We can express the PMF of X using its mean. If the mean is m, then we need $p = 1/(m + 1)$, and therefore, the PMF of a geometric random variable with mean m is given by the following:

$$\Pr[X = j] = \frac{m^j}{(m + 1)^{j+1}}, \tag{65}$$

where $j = 0, 1, 2, 3, \cdots$, etc.

We next observe that C_i is the number of 'A' characters in $Q^{(i)}$, where $Q^{(i)}$ is the string inserted after S_i and before S_{i+1}. The number of characters in $Q^{(i)}$ is l_i, which follows a geometric distribution with mean d, and each character can be an 'A' with probability 1/4. Note that it immediately follows that $\mathbf{E}[C_i] = d/4$. We can therefore compute the probability that $C_i = j$ as follows:

$$\Pr[C_i = j] = \sum_{x=j}^{\infty} \Pr[C_i = j \mid l_i = x] \Pr[l_i = x] \tag{66}$$

$$= \sum_{x=j}^{\infty} \binom{x}{j} \left(\frac{1}{4}\right)^j \left(\frac{3}{4}\right)^{x-j} \Pr[l_i = x] \tag{67}$$

$$= \sum_{x=j}^{\infty} \binom{x}{j} \left(\frac{1}{4}\right)^j \left(\frac{3}{4}\right)^{x-j} \frac{d^x}{(d+1)^{x+1}} \tag{68}$$

$$= \left(\frac{1}{4}\right)^j \left(\frac{4}{3}\right)^j \frac{1}{d+1} \sum_{x=j}^{\infty} \binom{x}{j} \left(\frac{3}{4}\right)^x \frac{d^x}{(d+1)^x} \tag{69}$$

$$= \left(\frac{1}{4}\right)^j \left(\frac{4}{3}\right)^j \frac{1}{d+1} \sum_{x=j}^{\infty} \binom{x}{j} \left(\frac{3d}{4d+4}\right)^x \tag{70}$$

We can now use the following summation for Binoamial coefficients:

$$\sum_{x=0}^{\infty} \binom{x}{j} r^x = \frac{r^j}{(1-r)^{j+1}}, \tag{71}$$

where $|r| < 1$. In our case, $r = 3d/(4d+4)$. Our mutation process assumes $d \geq 0$, which means $4d + 4 > 3d$, and therefore, this summation can be used to simplify Equation (70).

$$\Pr[C_i = j] = \left(\frac{1}{4}\right)^j \left(\frac{4}{3}\right)^j \frac{1}{d+1} \left(\frac{3d}{4d+4}\right)^j \frac{1}{\left(1 - \frac{3d}{4d+4}\right)^{j+1}} \tag{72}$$

$$= \left(\frac{1}{4}\right)^j \left(\frac{4}{3}\right)^j \frac{1}{d+1} \times \frac{3^j d^j}{4^j (d+1)^j} \times \frac{4^{j+1}(d+1)^{j+1}}{(d+4)^{j+1}} \tag{73}$$

$$= \frac{4d^j}{(d+1)^{j+1}} \tag{74}$$

$$= \frac{(d/4)^j}{\left(d/4+1\right)^{j+1}} \tag{75}$$

This is the same form as the PMF of a geometric distribution shown in Equation (65). Since $\mathbf{E}[C_i] = d/4$, it follows that each C_i is a geometric random variable with mean $d/4$.

We can now use the result proved for sum of geometric random variables with mean d, as shown in Equation (34). Note that $C = \sum_{i=1}^{L} C_i$, which means $\mathbf{E}[C] = Ld/4$, where each C_i follows a geometric distribution with mean $d/4$. Using the result in Equation (34) gives us the following:

$$\Pr\left[\left|C - \frac{Ld}{4}\right| \leq \delta \frac{Ld}{4}\right] \geq 1 - 2\exp\left\{-\frac{L\delta^2 d^2}{2(d+4-\delta d)(d+4)}\right\}. \tag{76}$$

D.2.4 Strong concentration of $f_A{}'$ around $\mathbf{E}[f_A{}']$

We can now use the concentration bounds in Equations 60, 62, and 76 to take a union bound and prove the following.

$$\Pr\left[\left|f_A' - \mathbf{E}[f_A']\right| \leq \delta \mathbf{E}[f_A']\right] \leq$$
$$1 - 2\exp\left\{-\frac{\delta^2 f_A(1 - p_s - p_d)}{3}\right\} - 2\exp\left\{-\frac{\delta^2 (L - f_A) p_s}{9}\right\} - 2\exp\left\{-\frac{L\delta^2 d^2}{2(d + 4 - \delta d)(d + 4)}\right\}. \tag{77}$$