Batch Optimization for DNA Synthesis

Konstantin Makarychev, Miklós Z. Rácz, Cyrus Rashtchian, and Sergey Yekhanin

***Abstract*— Large pools of synthetic DNA molecules have been recently used to reliably store significant volumes of digital data. While DNA as a storage medium has enormous potential because of its high storage density, its practical use is currently severely limited because of the high cost and low throughput of available DNA synthesis technologies. We study the role of batch optimization in reducing the cost of large scale DNA synthesis, which translates to the following algorithmic task. Given a large pool $\mathcal{S}$ of random quaternary strings of fixed length, partition $\mathcal{S}$ into batches in a way that minimizes the sum of the lengths of the shortest common supersequences across batches. We introduce two ideas for batch optimization that both improve (in different ways) upon a naive baseline: (1) using both $(ACGT)^*$ and its reverse $(TGCA)^*$ as reference strands, and batching appropriately, and (2) batching via the quantiles of an appropriate ordering of the strands. We also prove asymptotically matching lower bounds on the cost of DNA synthesis, showing that one cannot improve upon these two ideas. Our results uncover a surprising separation between two cases that naturally arise in the context of DNA data storage: the asymptotic cost savings of batch optimization are significantly greater in the case where strings in $\mathcal{S}$ do not contain repeats of the same character (homopolymers), as compared to the case where strings in $\mathcal{S}$ are unconstrained.**

***Index Terms*— DNA synthesis, DNA storage, shortest common supersequence, batch optimization.**

I. Introduction

STORING digital data in synthetic DNA molecules has received much attention in the past decade [1]–[11]. DNA data storage offers several orders of magnitude higher information density compared to conventional storage media, as well as the potential to store data reliably for hundreds or thousands of years. However, the prohibitively high cost and low throughput of modern DNA synthesis technologies present a key barrier that needs to be addressed in order to make DNA data storage a commonplace technology.

Manuscript received 6 November 2021; accepted 6 June 2022. Date of publication 21 June 2022; date of current version 21 October 2022. The work of Miklós Z. Rácz was supported in part by the NSF under Grant DMS 1811724 and in part by the Princeton SEAS Innovation Award. An earlier version of this paper was presented in part at the 2021 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT45174.2021.9517820]. *(Corresponding author: Miklós Z. Rácz.)*

Konstantin Makarychev is with the Department of Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: konstantin@northwestern.edu).

Miklós Z. Rácz is with Department of ORFE, Princeton University, Princeton, NJ 08544 USA (e-mail: mracz@princeton.edu).

Cyrus Rashtchian is with the Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: crashtchian@eng.ucsd.edu).

Sergey Yekhanin is with Microsoft Research, Redmond, WA 98052 USA (e-mail: yekhanin@microsoft.com).

Communicated by S. Boucheron, Associate Editor for Statistical Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3184903.

Digital Object Identifier 10.1109/TIT.2022.3184903

For the purposes of the current paper we can think of a DNA molecule as a string (strand) in the quaternary alphabet $\{A, C, G, T\}$. Today the dominant method for producing large quantities of DNA molecules is array-based DNA synthesis [12], [13]. With this technology the DNA synthesizer creates a large number of DNA strands in parallel, where each strand is grown by one DNA base (character) at a time. To append bases to strands, the synthesis machine follows a fixed supersequence of bases, called a reference strand. As the machine iterates through this supersequence, the next base is added to a select subset of the DNA strands. This process continues until the machine reaches the end of the supersequence. In particular, each synthesized DNA strand must be a subsequence of the reference strand. The cost of DNA synthesis is proportional to the length of the reference strand.

In applications to DNA data storage one typically needs to synthesize very large quantities of DNA molecules, significantly exceeding the capacity of any single DNA synthesizer. Therefore the pool of strands that one aims to synthesize needs to be partitioned into batches, where the size of each batch corresponds to the maximum load of the synthesizer. In this setting the total cost of DNA synthesis is proportional to the sum of the lengths of the shortest common supersequences of each batch. The focus of this paper is the algorithmic task of *batch optimization*, where the goal is to partition the strands into batches and assign every batch a reference strand in a way that minimizes this cost.

The encoding process that generates the list of DNA strands that need to be synthesized to store a given digital file varies with the specific system [2], [5], [9], [14] and is usually quite complex. The encoder adds redundancy to the data to allow for the correction of various types of errors that occur during DNA synthesis, storage, and sequencing, including insertions, deletions, and substitutions of individual bases, as well as missing DNA strands.

We now describe two aspects of encoding of digital data in DNA that are relevant to our work. Commonly, input digital data is randomized [9] using a seeded pseudorandom number generator or compressed and encrypted [14]; this is done in order to reduce the frequency of undesirable patterns that may occur in strands that are used to represent the data, for instance, patterns likely to cause the presence of DNA secondary structure [15]. Ensuring that strands look random also facilitates certain tasks that may be a part of the decoding process such as clustering [9], [16] and trace reconstruction [9], [17]–[21]. Another important aspect is as follows. Algorithms that encode digital data in DNA [3], [9] often ensure that the resulting strands do not contain long runs of the same character (i.e., homopolymers), since such runs are known to cause errors during the DNA sequencing

stage. The length of the longest allowed homopolymer run may be as low as one—that is, not allowing homopolymers—or unconstrained, depending on the scenario.

Motivated by the above considerations, in the current paper we model pools $\mathcal{S}$ of DNA strands that we aim to synthesise as large collections of random quaternary strings. We consider two key representative cases: the case where strings in $\mathcal{S}$ are unconstrained and the case where strings in $\mathcal{S}$ do not contain repeats of the same character.

A. Problem Statement

Fix a strand length n, and consider two different choices for the strand universe $\mathcal{U}$.

1) *Unconstrained strands:* $\mathcal{U} = \{A, C, G, T\}^n$.
2) *Strands without homopolymers:* $\mathcal{U}$ is the subset of $\{A, C, G, T\}^n$ that contains all strands with no consecutively repeated characters.

Let $\mathcal{S}$ be a subset of elements of $\mathcal{U}$, with $M := |\mathcal{S}|$; this is the pool of strands we wish to synthesize. Let k be an integer that divides M, and let π be a partition of $\mathcal{S}$ into k subsets (which we refer to as *batches*) $\mathcal{B}_1, \ldots, \mathcal{B}_k$ of size M/k.[1] We define $\mathsf{cost}(\mathcal{B}_i)$, the cost of synthesizing elements of the batch $\mathcal{B}_i$, as the length of the shortest common supersequence of all strands in $\mathcal{B}_i$. Using this notation, we define the cost of synthesizing the whole pool $\mathcal{S}$ as:

$$\mathsf{cost}(\mathcal{S}) := \min_{\pi} \sum_{i=1}^{k} \mathsf{cost}(\mathcal{B}_i). \tag{1}$$

We assume that elements of $\mathcal{S}$ are selected i.i.d. from $\mathcal{U}$ uniformly at random, and we are interested in upper and lower bounds for $\mathsf{cost}(\mathcal{S})$ that hold with high probability.

While in the practice of DNA synthesis the parameters n, M, and k are concrete numbers, to facilitate the asymptotic study of the problem we focus on the following relevant scenario: n is growing, M is significantly larger than but polynomial in n, and k is either a constant or a slowly growing function of n.

Example 1.1: Consider the setting of strands with no homopolymers. Let $n = 4$ and $M = 4$. Let the set of strands to be synthesized be $\mathcal{S} = \{AGCT, GCAT, CAGA, GAGC\}$. Assume that $k = 2$, that is, there are two batches and each batch contains two strands.

We can partition $\mathcal{S}$ into $\mathcal{B}_1 = \{AGCT, GCAT\}$ and $\mathcal{B}_2 = \{CAGA, GAGC\}$. The DNA synthesizer (printer) first prints $\mathcal{B}_1$. It starts with two empty strings $(\emptyset, \emptyset)$. Then, it appends A to the first strand and obtains strands $(A, \emptyset)$. It appends G to both strands and obtains (AG, G). Then, it appends the letters C, A, and T as follows:

$$(\emptyset, \emptyset) \xrightarrow{A} (A, \emptyset) \xrightarrow{G} (AG, G) \xrightarrow{C} (AGC, GC) \xrightarrow{A} (AGC, GCA) \xrightarrow{T} (AGCT, GCAT).$$

[1]The assumption that the batches are of equal size is made for simplicity. Indeed, our techniques extend to a more general setting where the batches are roughly the same size (e.g., up to a constant factor), and several results are phrased in this more general setting.

After the last step, we get the set $\mathcal{B}_1 = \{AGCT, GCAT\}$. The printer prints $\mathcal{B}_2$ as follows:

$$(\emptyset, \emptyset) \xrightarrow{C} (C, \emptyset) \xrightarrow{G} (C, G) \xrightarrow{A} (CA, GA) \xrightarrow{G} (CAG, GAG) \xrightarrow{A} (CAGA, GAG) \xrightarrow{C} (CAGA, GAGC).$$

In this example, we used the reference strand $AGCAT$ to print the set $\mathcal{B}_1$ in five steps and the reference strand $CGAGAC$ to print the set $\mathcal{B}_2$ in six steps. Therefore $\mathsf{cost}(\mathcal{S}) \leq 11$.

B. Main Results for Multiple Batches

Before describing our main results for multiple batches, we briefly and informally discuss the setting of a single batch—formal statements and proofs are in Section III. A natural reference strand to use to print a pool of strands $\mathcal{S}$ is the periodic strand $(ACGT)^*$, where $ACGT$ repeats indefinitely.

For this reference strand, the cost of printing a random strand can be written as $\sum_{i=1}^{n} X_i$, where $\{X_i\}_{i=1}^{n}$ are i.i.d. uniformly random on $\{1, 2, 3, 4\}$ in the case of unconstrained strands; in the case of strands without homopolymers, $\{X_i\}_{i=1}^{n}$ are independent, with X_1 uniformly random on $\{1, 2, 3, 4\}$ and X_i uniformly random on $\{1, 2, 3\}$ for $i \geq 2$. By using a standard concentration inequality we then obtain the upper bounds $\mathsf{cost}(\mathcal{S}) \leq 2.5n + 3\sqrt{n \log M}$ for unconstrained strands and $\mathsf{cost}(\mathcal{S}) \leq 2n + 3\sqrt{n \log M}$ for strands without homopolymers, with both bounds holding with probability $1 - o(1)$. Combining this with an appropriate stochastic domination argument that compares random walks, we also obtain matching lower bounds, for both choices of the strand universe $\mathcal{U}$. This shows that for a single batch no reference strand can do asymptotically better than the periodic strand $(ACGT)^*$.

The setting of multiple batches, which is the focus of our work, presents interesting challenges. As a simple baseline, we could consider randomly partitioning $\mathcal{S}$ into k batches. A direct application of the single batch upper bound would provide a cost of roughly $2.5nk + O(k\sqrt{n \log(M/k)})$ for unconstrained strands and $2nk + O(k\sqrt{n \log(M/k)})$ for strands without homopolymers. We provide improvements in both cases by using a slightly more sophisticated batching method.

We first observe a symmetry property: For any strand without homopolymers the cost of printing it using $(ACGT)^*$ and the cost of printing it using $(TGCA)^*$ add up to $4n + 1$, so the better choice of reference strand results in a cost of at most $2n$. This idea can be extended to a large set of strands, by choosing for each strand the better reference strand out of $(ACGT)^*$ and its reverse $(TGCA)^*$. We further improve upon the cost by leveraging a second idea. After partitioning strands based on which of the two reference strands is better, we then sort the strands based on their cost (with respect to the chosen reference strand). We then use a quantile-based batching process to group the first M/k lowest cost strands, then the next M/k, etc. We show that combining these two ideas reduces the total cost to $2nk - \Theta(k\sqrt{n})$ for $k \geq 3$ batches.

In the case of unrestricted strands, such an improvement is not possible, although we are able to show that with k batches a similar partitioning strategy, based on appropriately

ordering the strands and using quantiles, enables us to save a factor of k in the deviation term and obtain a total cost of $2.5nk + O(\sqrt{n \log M})$. We now formally state our results.

Theorem 1.2 (Upper Bounds): Let $\mathcal{S}$ be a set of M random strands in $\{A, C, G, T\}^n$, and let k be an integer satisfying $3 \leq k \leq \frac{1}{4}\sqrt{\frac{M}{\log M}}$. There exist absolute constants $C_1 > 0$ and $C_2 < \infty$ such that the following hold.

1) *(Strands Without Homopolymers):* There exists a way to efficiently partition $\mathcal{S}$ into k equal size batches $\mathcal{B}_1, \ldots, \mathcal{B}_k$ such that with probability at least $1 - 1/M$ we have that

$$\sum_{i=1}^{k} \mathsf{cost}(\mathcal{B}_i) \leq 2nk - C_1 k\sqrt{n}.$$

2) *(Unconstrained Strands):* There exists a way to efficiently partition $\mathcal{S}$ into k equal size batches $\mathcal{B}_1, \ldots, \mathcal{B}_k$ such that with probability at least $1 - 1/M$ we have that

$$\sum_{i=1}^{k} \mathsf{cost}(\mathcal{B}_i) \leq 2.5nk + C_2\sqrt{n \log M}.$$

We complement these results with almost tight lower bounds. Proving the following theorem is the most technically challenging part of our work.

Theorem 1.3 (Lower Bounds): Let $\mathcal{S}$ be a set of $M \geq 10n^2 \log n$ random strands in $\{A, C, G, T\}^n$, and let k be a positive integer satisfying $k \leq \frac{1}{10}\sqrt{\log M / \log\log M}$.

1) *(Strands Without Homopolymers):* There exists an absolute constant $c_1 < \infty$ such that the following holds with probability at least $1 - c_1/M$. For any partition of $\mathcal{S}$ into k equal size batches $\mathcal{B}_1, \ldots, \mathcal{B}_k$, we have that

$$\sum_{i=1}^{k} \mathsf{cost}(\mathcal{B}_i) \geq 2nk - c_1 k\sqrt{n \log k}.$$

2) *(Unconstrained Strands):* Suppose that $M \leq \exp(n)$. There exists an absolute constant $c_2 > 0$ such that the following holds with probability at least $1 - c_2^{-1}/M$. For any partition of $\mathcal{S}$ into k equal size batches $\mathcal{B}_1, \ldots, \mathcal{B}_k$, we have that

$$\sum_{i=1}^{k} \mathsf{cost}(\mathcal{B}_i) \geq 2.5nk + c_2\sqrt{n \log M}.$$

Comparing Theorems 1.2 and 1.3, we see that the upper and lower bounds match up to the absolute constants in the deviation term when k is small enough. As a consequence, this provides evidence that our batching method is nearly optimal, perhaps surprisingly.

Furthermore, Theorems 1.2 and 1.3 provide a clear separation between the two representative strand universes. On the one hand, for unconstrained strands we have, with probability $1 - o(1)$, that $\mathsf{cost}(\mathcal{S}) = 2.5nk + \Theta\left(\sqrt{n \log M}\right)$; that is, the cost *exceeds* the main term $2.5nk$ by the deviation term. On the other hand, for strands without homopolymers we have, with probability $1 - o(1)$, that $2nk - c_1 k\sqrt{n \log k} \leq \mathsf{cost}(\mathcal{S}) \leq 2nk - C_1 k\sqrt{n}$; that is, the cost is *smaller than* the main term $2nk$ by the deviation term.

C. Related Work

For an overview of the biochemical DNA synthesis process, we refer the interested reader to the surveys [10], [12]. Our work is motivated by several experimental papers that address the challenge of reducing the synthesis cost in both single and multi-batch settings [22]–[31]. Variants of the problem have also been studied that incorporate certain quality control measures [32]–[35]. Much of this previous work considers the $(ACGT)^*$ supersequence when analyzing the synthesis cost. Rahmann first observed that in this case the single batch cost of uniformly random strings is approximately Gaussian, but he did not provide a formal analysis nor any asymptotic or finite-size bounds [24]. In the multi-batch setting, previous work uses the same cost function as we do, namely the sum of the shortest common supersequence (SCS) lengths for each batch [26], [30]. In general, a wide array of algorithms have been proposed and empirically evaluated for selecting a short reference string given the set of DNA strands to synthesize. However, these heuristics do not come with provable guarantees, and many of them implicitly solve the SCS problem, which is known to be NP-hard for a collection of strings [36], [37].

From a theoretical point of view, a few recent works have considered minimizing the synthesis cost through coding-based approaches. Lenz et al. study reference strings that have a large number of subsequences, and they consider mappings to encode data by a set of strings while minimizing the single-batch synthesis cost [38]. A slightly different synthesis model has also been considered, where information is stored based on run-length patterns in the strings [39]–[41]. Our work also relates to combinatorial questions about the number of distinct subsequences [42]–[46].

There is also a large body of prior work on the longest common subsequence (LCS) of random strings [47]–[53]. The expected LCS length of two random length n strings is known to be $(\gamma + o(1))n$ for a value $\gamma > 0$ called the Chvátal-Sankoff constant. Despite decades of effort, the exact value of γ remains unknown for constant alphabet sizes. For two length n strings, the LCS and SCS are related via the equality $\mathsf{SCS}(S_1, S_2) = 2n - \mathsf{LCS}(S_1, S_2)$, but for larger sets, no analogous relationship is known. In particular, our results show that the average SCS length for a large collection of strings behaves very differently than for a pair of strings. While we are not aware of prior results on the SCS for multiple batches, our single batch results improve an existing bound on the expected SCS length in the special case of $M = n$ strings (see Remark 3.4).

D. Organization

The rest of the paper is organized as follows. We begin in Section II with a technical overview of our proofs. In Section III, we provide both upper and lower bounds for a single batch. In Section IV, we introduce the cost quantile preliminaries that we use for our multi-batch results. We prove the upper bounds for the multi-batch setting, Theorem 1.2, in Section V. Finally, we prove the lower bounds for the multi-batch setting, Theorem 1.3, in Section VI.

II. Proof Overview

In this section we give an overview of our results and the associated proofs. Suppose we want to synthesize a DNA strand S using a reference strand R. Denote the length of the prefix of R which we use for synthesis by $\mathsf{cost}_R(S)$. Then, the cost of printing a batch of strands $\mathcal{B}$ using R equals the maximum cost of printing S for $S \in \mathcal{B}$:

$$\mathsf{cost}_R(\mathcal{B}) = \max_{S \in \mathcal{B}} \mathsf{cost}_R(S).$$

We observe that the cost of printing any strand of length n using the periodic reference strand $(ACGT)^*$ is at most $4n$, since the i-th base of S can be printed using the corresponding base in the i-th quadruple of $(ACGT)^*$. Hence, the cost of synthesizing any batch of strands of length n is bounded from above by $4n$. As we discuss later, the cost of every strand without homopolymers with respect to the reference strand $(ACGT)^*$ is at most $3n+1$. So the cost of any batch of strands without homopolymers is also at most $3n+1$.

Since the cost of synthesizing every batch of strands is upper bounded by $4n$, we do not need to consider reference strands of length more than $4n$. However, for the sake of analysis, we shall assume that all reference strands R have an infinite length. The first $4n$ bases of these strands are arbitrary, while the remaining infinite suffix is a repetition of the pattern $ACGT$. We denote the set of all such strands by $\mathcal{R}^*$. Observe that every strand $\mathcal{S}$ can be synthesized using every $R \in \mathcal{R}^*$ because R contains the substring $(ACGT)^*$. Note that when we synthesize a batch $\mathcal{B}$ using a reference strand $R \in \mathcal{R}^*$, we truncate R after $\mathsf{cost}_R(\mathcal{B})$ bases, so effectively we use a reference strand of length $\mathsf{cost}_R(\mathcal{B})$.

A. Cost of a Single Batch

We first show how to estimate the cost of synthesizing a single batch of DNA strands. We prove that for a random strand S of length n and reference strand $\widetilde{R} = (ACGT)^*$, the expected $\mathsf{cost}_{\widetilde{R}}(S)$ equals $2.5n$. We then use concentration inequalities to argue that the maximum cost of strands in $\mathcal{B}$ is upper bounded by $2.5n + O(\sqrt{n \log M})$ with high probability, where M is the batch size. Similarly, we show that for every fixed strand R, we have that $\mathbb{E}[\mathsf{cost}_R(S)] \geq 2.5n$. Hence, for every fixed R the cost of $\mathcal{B}$ is also lower bounded by $2.5n + \Omega(\sqrt{n \log M})$ with high probability. We obtain a lower bound on the cost of a batch by taking the union bound over all $R \in \mathcal{R}^*$. Similarly, we get lower and upper bounds of $2n + \Omega(\sqrt{n \log M})$ and $2n + O(\sqrt{n \log M})$ for random strands without homopolymers.

We now discuss how to compute $\mathbb{E}[\mathsf{cost}_R(S)]$ for a given reference strand R and random S. Let $\tau_i(S,R)$ be the cost of the prefix $S_1, \ldots, S_i$. In other words, $\tau_i(S,R)$ is the index of the base in R that is used for synthesizing the i-th base in S. We let $\tau_0(S,R) = 0$. Observe that $\{\tau_i(S,R)\}_{i \geq 0}$ is a Markov chain: the value of $\tau_{i+1}(S,R)$ depends only on the current state $\tau_i(S,R)$ and the random value of S_{i+1}. We denote the increments of $\tau_i(S,R)$ by $X_i(S,R)$: for $i \in \{1, \ldots, n\}$, let

$$X_i(S,R) := \tau_i(S,R) - \tau_{i-1}(S,R).$$

Then, $\mathsf{cost}_R(S) = \tau_n(S,R) = \sum_{i=1}^{n} X_i(S,R)$. For the reference strand $\widetilde{R} = (ACGT)^*$, each increment $X_i(S,\widetilde{R})$ is a random variable uniformly distributed in $\{1,2,3,4\}$, and all $X_i(S,\widetilde{R})$ are mutually independent. Consequently, $\mathbb{E}[X_i(S,\widetilde{R})] = 2.5$ for all i and thus $\mathbb{E}[\mathsf{cost}_{\widetilde{R}}(S)] = 2.5n$. Furthermore, by the central limit theorem, the deviation of the cost from its expectation, $\mathsf{cost}_{\widetilde{R}}(S) - 2.5n$, is approximately Gaussian with mean 0 and variance $1.25n$. Thus, we can use Hoeffding's inequality and other concentration inequalities to obtain upper and lower bounds of on $\mathsf{cost}_{\widetilde{R}}(S)$. These bounds imply that the cost of a single batch of M strands equals $2.5n + \Theta(\sqrt{n \log M})$.

To show that $\mathbb{E}[X_i(S,R)] \geq 2.5$ for every $R \in \mathcal{R}^*$ and not only for $R = \widetilde{R}$, we observe that the sequence of random variables $X_1(S,R), \ldots, X_n(S,R)$ *stochastically dominates* a sequence of i.i.d random variables $Y_1, \ldots, Y_n$, where each Y_i is uniformly distributed in $\{1,2,3,4\}$. Hence,

$$\mathbb{E}[X_1(S,R) + \cdots + X_n(S,R)] \geq \mathbb{E}[Y_1 + \cdots + Y_n] = 2.5n.$$

For random strands without homopolymers, each jump $X_i(S,\widetilde{R})$ is uniformly distributed in $\{1,2,3\}$ for $i > 1$; and $X_1(S,\widetilde{R})$ is uniformly distributed in $\{1,2,3,4\}$. Hence, the expected cost $\mathsf{cost}_{\widetilde{R}}(S)$ is $2n + 1/2$. Also, note that the maximum possible value of $X_i(S,\widetilde{R})$ is 3 (for $i > 1$). Hence, the cost of every strand is upper bounded by $3n+1$.

B. Upper Bounds for Multiple Batches

We are going to use the same reference strand $\widetilde{R} = (ACGT)^*$ for synthesizing all batches of unconstrained strands and two different reference strands, $\widetilde{R} = (ACGT)^*$ and its reverse $\bar{R} = (TGCA)^*$, for synthesizing batches of strands without homopolymers.

1) Naïve Approach: Suppose we assign strands randomly to k batches. Then, each batch consists of M/k random strands sampled uniformly from $\{A,C,G,T\}^n$. Hence, the cost of every batch is $2.5n + \Theta(\sqrt{n \log M})$. Consequently, the total cost of synthesising k batches is $2.5nk + k \cdot \Theta(\sqrt{n \log M})$. We now show that by carefully assigning strands to batches we can improve this cost to $2.5nk + \Theta(\sqrt{n \log M})$ for unconstrained strands. Similarly, we show how to improve a naïve solution of cost $2nk + k \cdot \Theta(\sqrt{n \log M})$ for strands without homopolymers to a solution of cost $2nk - \Omega(k\sqrt{n})$.

2) Unconstrained Strands: Our strategy for splitting the set of unconstrained strands $\mathcal{S}$ into k batches is quite simple. For every strand S in $\mathcal{S}$, we compute $\mathsf{cost}_{\widetilde{R}}(S)$ and then sort strands by this cost. We put the first M/k strands in the first batch, the second M/k strands in the second batch, and so on. Then, the cost of the i-th batch is equal to the empirical i/k-th quantile of $\{\mathsf{cost}_{\widetilde{R}}(S)\}_{S \in \mathcal{S}}$. We formally define empirical quantiles in Section IV. In Section IV, we also show that, with high probability, empirical quantiles of $\{\mathsf{cost}_{\widetilde{R}}(S)\}_{S \in \mathcal{S}}$ are very close to the corresponding quantiles of the distribution of the random variable $\mathsf{cost}_{\widetilde{R}}(S)$, where S is randomly and uniformly drawn from $\{A,C,G,T\}^n$. The only exception is the empirical 1-quantile of the sample $\mathcal{S}$ which corresponds to the cost of the most expensive strand in $\mathcal{S}$. This cost is approximately equal to the $(1 - 1/M)$-quantile of the distribution of $\mathsf{cost}_{\widetilde{R}}(S)$, where M is the size of $\mathcal{S}$.

As we discussed above, $\mathsf{cost}_{\widetilde{R}}(S)$ can be approximated by the random variable $2.5n + g$, where g is a Gaussian random variable with mean 0 and variance $1.25n$. The sum of the $1/k, 2/k, \ldots, (k-1)/k$ quantiles of a symmetric Gaussian distribution equals 0, since the quantiles i/k and $(k-i)/k$ are symmetric around 0. However, the $(1 - 1/M)$-quantile of the distribution of g is relatively large and approximately equals $c\sqrt{n \log M}$. Hence, the total cost of synthesizing k batches approximately equals

$$2.5nk + c\sqrt{n \log M}.$$

We make this argument formal in Section V.

3) Strands Without Homopolymers: If we use the same batching strategy as we discussed above for strands without homopolymers, we obtain a solution of cost $2nk + c\sqrt{n \log M}$ with high probability. However, somewhat surprisingly, we can do better by utilizing two reference strands, $\widetilde{R} = (ACGT)^*$ and $\bar{R} = (TGCA)^*$, instead of just the single strand $\widetilde{R}$. We show that the random variables $\mathsf{cost}_{\widetilde{R}}(S)$ and $\mathsf{cost}_{\bar{R}}(S)$ are anticorrelated. Specifically, for every strand S without homopolymers, we (deterministically) have

$$\mathsf{cost}_{\widetilde{R}}(S) + \mathsf{cost}_{\bar{R}}(S) = 4n + 1.$$

This observation suggests the following strategy: We first sort all strands $\mathcal{S}$ by their cost when printed with $\widetilde{R}$. For the first $\lceil k/2 \rceil$ batches, we print them with $\widetilde{R}$, and we print the remaining batches with $\bar{R}$. Overall, we will argue that this batching process results in $k-2$ batches having a cost of at most $2n$, and a constant fraction of these batches having an additional savings of $\Omega(\sqrt{n})$, which results in the ultimate savings of $\Omega(k\sqrt{n})$. The only challenging batches are the "middle" two. We handle these by arguing that their costs are coupled so that together they do not exceed $4n + 1$. We next explain the intuition behind the main savings. See Section V-A for full details of the batching process and the analysis.

Since $(X_i(S, \widetilde{R}) + X_i(S, \bar{R}))/2 = 2$ for all $i > 1$ and S does not have homopolymers, the random variables $\mathsf{cost}_{\widetilde{R}}(S)$ and $\mathsf{cost}_{\bar{R}}(S)$ can be approximated by *correlated* random variables $2n - g$ and $2n + g$, where g is a Gaussian random variable with mean 0 and variance $2/3\, n$. The cost of every strand is thus approximately equal to $\min\{2n-g, 2n+g\} = 2n - |g|$, and the total cost of k batches is approximately equal to the sum of the i/k-quantiles of the random variable $2n - |g|$ for $i = 1, \ldots, k$. For sufficiently large k, this sum is approximately equal to

$$k \cdot (\mathbb{E}[2n - |g|]) = k \cdot (2n - \mathbb{E}[|g|]) = 2nk - k\sqrt{\frac{4}{3\pi}\, n}.$$

For small k ($k > 2$), the sum is upper bounded by $2nk - \Omega(k\sqrt{n})$. We prove this bound for $k \geq 3$ in Section V-A.

C. Lower Bounds for Multiple Batches

We now discuss how to obtain lower bounds on the cost of batch synthesis. We start with lower bounds that are based on the following observation: Every batch $\mathcal{B}$ must contain a $1/k$ fraction of all strands in $\mathcal{S}$. Consequently, its cost is lower bounded by the empirical $1/k$-quantile of $\{\mathsf{cost}_R(S)\}_{S \in \mathcal{S}}$, which, in turn, approximately equals the $1/k$-quantile of the distribution of the random variable $\mathsf{cost}_R(S)$, where S is a random strand. Here R is the reference strand used for synthesising $\mathcal{B}$. Using the notation (defined in Section IV) for empirical q-quantiles $\widetilde{Q}_{q,R}(\mathcal{S})$ and q-quantiles $Q_{q,R}(D)$ of a distribution D, we can lower bound the cost of $\mathcal{B}$ as follows:

$$\mathsf{cost}(\mathcal{B}) \geq \min_{R \in \mathcal{R}^*} \widetilde{Q}_{1/k,R}(\mathcal{S}) \gtrsim \min_{R \in \mathcal{R}^*} Q_{1/k,R}(D_{1/4}),$$

where $D_{1/4}$ is the uniform distribution of strands of length n. Using Hoeffding's inequality for $\mathsf{cost}_R(S)$ along with bounds on $\widetilde{Q}_{1/k,R}(\mathcal{S})$ and $Q_{1/k,R}(D_{1/4})$ from Section IV, we then show that $Q_{1/k,R}(D_{1/4}) \geq 2.5n - O(\sqrt{n \log k})$ which yields a lower bound of $k \cdot (2.5n - O(\sqrt{n \log k}))$ on the total cost of synthesizing k batches. For strands without homopolymers, the same argument gives a bound of $k \cdot (2n - O(\sqrt{n \log k}))$.

1) Improved Lower Bound for Unconstrained Strands: We then improve the lower bound on the cost of batch synthesis of unconstrained strands by showing that while the cost of all batches are lower bounded by $2.5n - O(\sqrt{n \log k})$, the cost of the most expensive batch is at least $2.5n + \Omega(\sqrt{n \log M})$. Note that a similar statement does not hold for strands without homopolymers. To prove that the cost of the most expensive batch is $2.5n + \Omega(\sqrt{n \log M})$, we consider a subset $\mathcal{S}''$ of strands that have disproportionately many (roughly, $n/4 + c\sqrt{n \log M}$) repeated bases. We show that a random set $\mathcal{S}$ contains many such strands (approximately $\sqrt{M}$) and then prove that for random strands S from $\mathcal{S}''$, the expected cost $\mathsf{cost}_R(S)$ is at least $2.5n + c\sqrt{n \log M}$. This gives us a lower bound of $2.5nk + c\sqrt{n \log M} - O(k\sqrt{n \log k})$ on the total cost of synthesising k batches (note, typically $M \gg k$).

III. Single Batch Analysis

As a warm-up to the multiple batch setting, in this section we analyze the single batch setting, that is, the setting where $k = 1$. As discussed in Section I-B, by using the periodic strand $(ACGT)^*$ as a reference strand, we obtain (assuming $M \geq n$) the upper bounds $\mathsf{cost}(\mathcal{S}) \leq 2.5n + 3\sqrt{n \log M}$ for unconstrained strands and $\mathsf{cost}(\mathcal{S}) \leq 2n + 3\sqrt{n \log M}$ for strands without homopolymers, with both bounds holding with probability $1 - o(1)$. The formal statement and its short proof are in Section III-A.

We also obtain matching lower bounds, for both choices of the strand universe $\mathcal{U}$, by an appropriate stochastic domination argument that compares random walks. The formal statements are in Section III-B and their proofs are in Section III-C.

A. Upper Bounds for a Single Batch

Theorem 3.1: Consider the problem setup in Section I-A with $k = 1$. Let $M' := \max\{M, n\}$.

1) *(Unconstrained Strands):* With probability at least $1 - 1/n$ we have that
$$\mathsf{cost}(\mathcal{S}) \leq 2.5n + 3\sqrt{n \log M'}.$$
2) *(Strands Without Homopolymers):* With probability at least $1 - 1/n$ we have that
$$\mathsf{cost}(\mathcal{S}) \leq 2n + 3\sqrt{n \log M'}.$$

Proof: Consider first the case of unconstrained strands. We fix $R := (ACGT)^*$ as the reference strand with which we print the strands in $\mathcal{S}$. For a strand $S \in \mathcal{U}$, let $\tau_i(S)$ denote the index of the base of R that is used to print the ith base of S. With this notation, R needs $\max_{S\in\mathcal{S}} \tau_n(S)$ bases to print all strands in $\mathcal{S}$. This shows that $\mathsf{cost}(\mathcal{S}) \le \mathsf{cost}_R(\mathcal{S}) = \max_{S\in\mathcal{S}} \tau_n(S)$. Let $\lambda := 3\sqrt{n \log M'}$ and $m := 2.5n + \lambda$. Combining the previous observation with a union bound we thus have that

$$\begin{aligned}\mathbb{P}\{\mathsf{cost}(\mathcal{S}) > m\} &\le \mathbb{P}\{\mathsf{cost}_R(\mathcal{S}) > m\} \\ &= \mathbb{P}\left\{\max_{S\in\mathcal{S}} \tau_n(S) > m\right\} \le M\mathbb{P}\{\tau_n(S) > m\}.\end{aligned}$$

The key observation is that if S is a uniformly random strand, then $\tau_n(S) \stackrel{d}{=} \sum_{i=1}^n X_i$, where $\{X_i\}_{i=1}^n$ are i.i.d. random variables that are uniform on $\{1,2,3,4\}$; here $\stackrel{d}{=}$ denotes equality in distribution. Therefore, noting that $\mathbb{E}[X_1] = 2.5$, by Hoeffding's inequality (Theorem 1.1) we have that

$$\begin{aligned}\mathbb{P}\{\tau_n(S) > m\} &= \mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > \lambda\right\} \\ &\le \exp\left(-\frac{2\lambda^2}{9n}\right) = \frac{1}{(M')^2}.\end{aligned}$$

Combining the two displays above we have that $\mathbb{P}\{\mathsf{cost}(\mathcal{S}) > m\} \le M/(M')^2 \le 1/n$, as desired.

The case of strands without homopolymers is analogous, the only change is that now $\{X_i\}_{i=1}^n$ are independent random variables with X_1 uniformly random on $\{1,2,3,4\}$ and X_i uniformly random on $\{1,2,3\}$ for $i \ge 2$. □

B. Lower Bounds for a Single Batch

The lower bounds for the two strand universes are analogous, but we state them separately for clarity. In both cases we present two bounds: one focusing on the constant of the linear term, with weak assumptions on M, the other focusing on the second order term, assuming slightly more about M.

1) Unconstrained Strands: Recall the upper bound of $2.5n + 3\sqrt{n \log M}$ from Theorem 3.1 (assuming $M \ge n$). Our first lower bound result says that the constant 2.5 cannot be improved, even if the batch only consists of a (large enough) constant amount of strands.

Theorem 3.2: Consider the problem setup in Section I-A with $k = 1$ and unconstrained strands. Fix $\varepsilon > 0$ and let $M \ge 21/\varepsilon^2$. Then, with probability at least $1 - \exp(-n)$ we have that

$$\mathsf{cost}(\mathcal{S}) \ge (2.5 - \varepsilon)\, n.$$

This result can be significantly sharpened: if the number of strands in $\mathcal{S}$ is at least linear in n (and at most exponential in n), then not only is the main term $2.5n$ required in the cost, but even the additional $\sqrt{n \log M}$ term is necessary.

Theorem 3.3: Consider the problem setup in Section I-A with $k = 1$ and unconstrained strands. Let M satisfy $(5\exp(45))\, n \le M \le 5n\exp(4n/25)$. Then, with probability at least $1 - \exp(-n)$ we have that

$$\mathsf{cost}(\mathcal{S}) \ge 2.5n + \frac{1}{5}\sqrt{n \log\left(\frac{M}{5n}\right)}.$$

Remark 3.4 (Expected Shortest Common Supersequence (SCS)): Since $\mathsf{cost}(\mathcal{S}) = \mathsf{SCS}(\mathcal{S})$, our results also provide bounds on the SCS of a set of strings. Jiang and Li consider the case when $M = n$, and they prove that $\mathbb{E}_{\mathcal{S}}[\mathsf{SCS}(\mathcal{S})] = 2.5n \pm O(n^{0.707})$ for a set $\mathcal{S}$ of n uniformly random length n quaternary strings [37, Corollary 4.11]. Combining Theorem 3.1 and Theorem 3.2 with $\epsilon = \sqrt{21/n}$ we obtain $2.5n - \sqrt{21n} \le \mathbb{E}_{\mathcal{S}}[\mathsf{SCS}(\mathcal{S})] \le 2.5n + 3\sqrt{n \log n}$, tightening the prior result.

2) Strands Without Homopolymers: Next, we state theorems analogous to those in Section III-B.1, but in the constrained setting where $\mathcal{U}$ contains strands without homopolymers. (We allow the reference strand to potentially have homopolymers.) Recall the upper bound of $2n + 3\sqrt{n \log M}$ from Theorem 3.1 (assuming $M \ge n$). Our first lower bound result says that the constant 2 cannot be improved, even if the batch only consists of a (large enough) constant amount of strands.

Theorem 3.5: Consider the problem setup in Section I-A with $k = 1$ and strands without homopolymers. Fix $\varepsilon > 0$ and let $M \ge 9/\varepsilon^2$. Then, with probability at least $1 - \exp(-n)$ we have that

$$\mathsf{cost}(\mathcal{S}) \ge (2 - \varepsilon)\, n.$$

This result can be significantly sharpened: if the number of strands in $\mathcal{S}$ is at least linear in n (and at most exponential in n), then not only is the main term $2n$ required in the cost, but even the additional $\sqrt{n \log(M)}$ term is necessary.

Theorem 3.6: Consider the problem setup in Section I-A with $k = 1$ and strands without homopolymers. Let M satisfy $(5\exp(45))\, n \le M \le 5n\exp(4n/25)$. Then, with probability at least $1 - \exp(-n)$ we have that

$$\mathsf{cost}(\mathcal{S}) \ge 2n + \frac{3}{20}\sqrt{n \log\left(\frac{M}{5n}\right)}.$$

C. Lower Bound Proofs for a Single Batch

We start with some definitions. Let $\Sigma := \{A, C, G, T\}$. For two strands $R, S \in \Sigma^*$, let $\mathcal{E}_R(S)$ denote the event that R is a superstring of S. Let $\mathcal{E}_R(\mathcal{S})$ denote the event that R is a superstring of all strands in $\mathcal{S}$, that is, $\mathcal{E}_R(\mathcal{S}) := \cap_{S\in\mathcal{S}} \mathcal{E}_R(S)$. For an integer m, let $\mathcal{E}_m(\mathcal{S})$ denote the event that there exists a strand $R \in \Sigma^{\le m}$ such that the event $\mathcal{E}_R(\mathcal{S})$ holds. That is,

$$\{\mathsf{cost}(S) \le m\} = \mathcal{E}_m(\mathcal{S}) = \bigcup_{R\in\Sigma^{\le m}} \mathcal{E}_R(\mathcal{S}) = \bigcup_{R\in\Sigma^m} \mathcal{E}_R(\mathcal{S}),$$

where the second equality holds because if R is a superstring of S, then all superstrings of R are also a superstring of S.

1) Proofs for Unconstrained Strands:

Proof of Theorem 3.2: Fix $m := (2.5 - \varepsilon)\, n$. We start with a union bound:

$$\begin{aligned}\mathbb{P}\{\mathsf{cost}(\mathcal{S}) \le m\} &= \mathbb{P}\{\mathcal{E}_m(\mathcal{S})\} \le \sum_{R\in\Sigma^m} \mathbb{P}\{\mathcal{E}_R(\mathcal{S})\} \\ &= \sum_{R\in\Sigma^m} (\mathbb{P}\{\mathcal{E}_R(S)\})^M, \qquad (2)\end{aligned}$$

where the equality is due to the fact that the strands in $\mathcal{S}$ are i.i.d. Our goal now is to bound $\mathbb{P}\{\mathcal{E}_R(S)\}$, where $R \in \Sigma^m$ is fixed and $S \in \Sigma^n$ is uniformly random.

To understand the probability of this event we introduce some notation. First, we extend the reference strand R indefinitely, by concatenating the strand $(ACGT)^*$ to the end of R—this is done just so that everything in the following is well-defined—we refer to this extended strand as R'. Note that R' is a superstring of S, and the original reference strand R is a superstring of S if and only if the printing of S using R' succeeds in at most m steps. Let $\tau_i(S,R')$ denote the index of the base of R' that is used to print the ith base of S. We also define $X_1(S,R') := \tau_1(S,R')$, and $X_i(S,R') := \tau_i(S,R') - \tau_{i-1}(S,R')$ for $i > 1$. With this notation we have that

$$\mathcal{E}_R(S) = \{\tau_n(S,R') \le m\} = \left\{\sum_{i=1}^{n} X_i(S,R') \le m\right\}.$$

Let $\{Y_i\}_{i=1}^n$ be i.i.d. random variables that are uniform on $\{1,2,3,4\}$. Note that for an arbitrary R the random variables $\{X_i(S,R')\}_{i=1}^n$ are not i.i.d.—in fact, they are not even necessarily independent. Specifically, given R' and $\{X_j(S,R')\}_{j=1}^{i-1}$, the *support* of $X_i(S,R')$ is determined. However, no matter what, this support always consists of four distinct positive integers—the distances to the next occurrences of the four bases A, C, G, and T in R. Moreover, the *distribution* on this support is always uniform—this is because, given R' and $\{X_j(S,R')\}_{j=1}^{i-1}$, the value of $X_i(S,R')$ is determined by S_i, which is uniformly random on Σ. Since at best (in terms of minimization) the four distinct positive integers in the support of $X_i(S,R')$ are 1, 2, 3, and 4, the random variable $X_i(S,R')$ *stochastically dominates* Y_i. Moreover, since the bases $\{S_i\}_{i=1}^n$ are independent, we also have that $\sum_{i=1}^n X_i(S,R')$ stochastically dominates $\sum_{i=1}^n Y_i$, and so we have that

$$\mathbb{P}\{\mathcal{E}_R(S)\} = \mathbb{P}\left\{\sum_{i=1}^{n} X_i(S,R') \le m\right\} \le \mathbb{P}\left\{\sum_{i=1}^{n} Y_i \le m\right\}. \tag{3}$$

We can now bound this latter quantity using standard estimates. In particular, using Hoeffding's inequality (Theorem 1.1) we obtain that

$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i \le m\right\} \le \exp\left\{-\frac{2}{9}\varepsilon^2 n\right\}.$$

Plugging this back into (2) and (3), we have obtained the bound

$$\begin{aligned}\mathbb{P}\{\mathsf{cost}(\mathcal{S}) \le m\} &\le 4^m \exp\left(-\frac{2}{9}\varepsilon^2 M n\right) \\ &\le \exp\left(\left\{2.5\log 4 - \frac{2}{9}\varepsilon^2 M\right\} n\right). \end{aligned} \tag{4}$$

If $M \ge 21/\varepsilon^2$, then this is at most $\exp(-n)$. □

Proof of Theorem 3.3: Let $m := 2.5n + \lambda$, where $\lambda = \frac{1}{5}\sqrt{n\log\left(\frac{M}{5n}\right)}$. The proof is identical to the proof of Theorem 3.2 until (3). At this point in the proof, we bound this probability differently. Specifically, by applying Lemma 1.3 with $\ell = 4$, we have that

$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i > m\right\} \ge \exp\left(-25\lambda^2/n\right) = \frac{5n}{M}. \tag{5}$$

In order to obtain this bound via Lemma 1.3 we must have that $(4/3)\sqrt{n} \le \lambda \le 2n/25$; these inequalities hold due to the condition $(5\exp(45))\, n \le M \le 5n\exp(4n/25)$ assumed in Theorem 3.3. Note in particular that we thus have $m \le 2.58n$. From (5) we thus have that

$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i \le m\right\} \le 1 - \frac{5n}{M}.$$

Plugging this into (2) and (3), we have obtained—analogously to (4)—the bound

$$\begin{aligned}\mathbb{P}\{\mathsf{cost}(\mathcal{S}) \le m\} &\le 4^m \left(1 - \frac{5n}{M}\right)^M \\ &\le \exp\left(\left(2.58\log 4\right) n - 5n\right) \le \exp(-n).\end{aligned}$$

□

2) Proofs for Strands Without Homopolymers:

Proof of Theorem 3.5: The proof is almost identical to the proof of Theorem 3.2, so we only highlight the changes. First, we set $m := (2-\varepsilon)n$. Second, $\{Y_i\}_{i=1}^n$ are now i.i.d. random variables that are uniform on $\{1,2,3\}$. Thus Hoeffding's bound gives that

$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i \le m\right\} \le \exp\left(-\frac{1}{2}\varepsilon^2 n\right).$$

The rest of the proof is identical. □

Proof of Theorem 3.6: Let $m := 2n + \lambda$, where $\lambda = \frac{3}{20}\sqrt{n\log\left(\frac{M}{5n}\right)}$. The proof is identical to the proof of Theorem 3.3, except $\{Y_i\}_{i=1}^n$ are now i.i.d. random variables that are uniform on $\{1,2,3\}$, and thus Lemma 1.3 is applied with $\ell = 3$. □

IV. Empirical Quantiles

Consider a distribution D on strands, and a reference strand R. For $q \in (0,1)$, we define the q-quantile of the distribution of the printing cost $\mathsf{cost}_R(S)$, where S is distributed according to D, as the minimum t such that $\mathbb{P}\{\mathsf{cost}_R(S) \le t\} \ge q$:

$$Q_{q,R}(D) := \min\left\{t : \mathbb{P}\{\mathsf{cost}_R(S) \le t\} \ge q\right\}.$$

Similarly, for a set of strands $\mathcal{S}$, we define an *empirical* variant of Q as:

$$\widetilde{Q}_{q,R}(\mathcal{S}) := \min\left\{t : \frac{|\{S \in \mathcal{S} : \mathsf{cost}_R(S) \le t)\}|}{|\mathcal{S}|} \ge q\right\}.$$

For a family of reference strands $\mathcal{R}$, we let

$$Q_{q,\mathcal{R}}(D) := \min_{R\in\mathcal{R}} Q_{q,R}(D)$$

and

$$\widetilde{Q}_{q,\mathcal{R}}(\mathcal{S}) := \min_{R\in\mathcal{R}} \widetilde{Q}_{q,R}(\mathcal{S}).$$

Consider a random set $\mathcal{S}_{D,M}$ that contains M i.i.d. samples from the distribution D. By the Glivenko-Cantelli theorem, we have $\widetilde{Q}_{q,R}(\mathcal{S}_{D,M}) \approx Q_{q,R}(D)$ for every R and a sufficiently large M. We use the Dvoretzky-Kiefer-Wolfowitz inequality to get a quantitative bound on $\widetilde{Q}_{q,R}(\mathcal{S}_{D,M})$.

Lemma 4.1: For every distribution D on strands, every reference strand R, and every positive ε, we have

$$\mathbb{P}\Big\{Q_{q-\varepsilon,R}(D) \le \widetilde{Q}_{q,R}(\mathcal{S}_{D,M}) \le Q_{q+\varepsilon,R}(D) \text{ for all } q \in (\varepsilon, 1-\varepsilon)\Big\} > 1 - 2e^{-2M\varepsilon^2}. \quad (6)$$

Furthermore, for every family of reference strands $\mathcal{R}$, we have

$$\mathbb{P}\Big\{Q_{q-\varepsilon,\mathcal{R}}(D) \le \widetilde{Q}_{q,\mathcal{R}}(\mathcal{S}_{D,M}) \le Q_{q+\varepsilon,\mathcal{R}}(D) \text{ for all } q \in (\varepsilon, 1-\varepsilon)\Big\} > 1 - 2|\mathcal{R}|e^{-2M\varepsilon^2}. \quad (7)$$

We prove this lemma in Appendix B.

V. Upper Bounds for Multiple Batches

A. Batching for Strands Without Homopolymers

Our approach is to define a quantile-based batching process and to then split the strands into two groups based on whether we use the reference strand $\widetilde{R} = (ACGT)^*$ or its reverse $\bar{R} = (TGCA)^*$. We first observe that one of these two options will lead to a cost of at most $2n$ if the strand does not contain homopolymers (and this result holds deterministically).

Lemma 5.1: Let $\widetilde{R} = (ACGT)^*$ and $\bar{R} = (TGCA)^*$ be the alternating sequence and its reverse. For a strand $S \in \{A,C,G,T\}^n$ without homopolymers, we have that

$$\mathsf{cost}_{\widetilde{R}}(S) + \mathsf{cost}_{\bar{R}}(S) = 4n + 1. \quad (8)$$

Moreover, there exists a bijection $\varphi : \{A,C,G,T\}^n \to \{A,C,G,T\}^n$ such that for every strand $S \in \{A,C,G,T\}^n$ we have that

$$\mathsf{cost}_{\widetilde{R}}(S) + \mathsf{cost}_{\widetilde{R}}(\varphi(S)) = 4n + 1. \quad (9)$$

Proof: Let $\tau_i(S,\widetilde{R})$ and $\tau_i(S,\bar{R})$ be the time that the i^{th} character of S is printed using $\widetilde{R}$ or $\bar{R}$, respectively. Consider the per-character costs $X_i(S,\widetilde{R}) = \tau_i(S,\widetilde{R}) - \tau_{i-1}(S,\widetilde{R})$ and $X_i(S,\bar{R}) = \tau_i(S,\bar{R}) - \tau_{i-1}(S,\bar{R})$. Observe that $X_i(S,\widetilde{R}) + X_i(S,\bar{R}) = 4$ for $i > 1$ and $X_1(S,\widetilde{R}) + X_1(S,\bar{R}) = 5$. Hence,

$$\mathsf{cost}_{\widetilde{R}}(S) + \mathsf{cost}_{\bar{R}}(S) = \sum_{i=1}^{n} (X_i(S,\widetilde{R}) + X_i(S,\bar{R})) = 4n + 1.$$

We now map every strand S to its compliment by replacing each base A with T, C with G, G with C, and T with A. Observe that if we renamed each base as above both in S and the reference strand $\widetilde{R}$, then the cost would not change. That is, $\mathsf{cost}_{\widetilde{R}}(S) = \mathsf{cost}_{\bar{R}}(\varphi(S))$. Using (8) we thus obtain (9). □

The key ideas in the following proof are to (i) use a quantile-based batching process to group the strands, and then (ii) decide whether they will be printed with $\widetilde{R}$ or $\bar{R}$ to reduce the overall cost. By using the above lemma, we know that for strands that have cost larger than $2n$ under $\widetilde{R}$, they will have cost at most $2n$ under $\bar{R}$. Using this and properties of the batching process construction, we argue that for $k-2$ of the batches the cost is at most $2n$, with a constant fraction of these batches having cost at most $2n - \Omega(\sqrt{n})$. For the two "middle" batches, we only show that the sum of their costs is at most $4n+1$. Overall, we achieve a total cost of $2nk - \Omega(k\sqrt{n})$ when summing over the k batches.

Proof of Theorem 1.2(1): For the proof, we assume that k divides M; otherwise, we could set the batches to have size within $M/k \pm 1$. Let $\mathcal{D}$ denote the uniform distribution of length n strands without homopolymers.

We start by defining the batching process and the assignment of reference strands (i.e., choosing between $\widetilde{R}$ and $\bar{R}$ for each batch). To group the strands, we first sort the strands in $\mathcal{S}$ in a nondecreasing order according to their cost with respect to $\widetilde{R}$. Then, we let $\mathcal{B}_i$ be the subset of strands placed between positions $(i-1)M/k+1$ and iM/k (inclusive) in the ordering (e.g., $\mathcal{B}_1$ contains the M/k lowest cost strands). After the batches have been defined, we use $\widetilde{R}$ to print the first $\ell := \lceil k/2 \rceil$ batches and we use $\bar{R}$ for the remainder.

By Lemma 5.1, the batches with higher cost under $\widetilde{R}$ have lower cost under $\bar{R}$. Furthermore, with high probability, all batches have cost at most $2n$ except perhaps the "middle" two batches $\mathcal{B}_\ell$ and $\mathcal{B}_{\ell+1}$. More precisely, we utilize the empirical quantiles i/k for $i \in [k]$, where we recall that $\widetilde{Q}_{i/k,\widetilde{R}}(\mathcal{S})$ denotes the minimum value t_i such that an i/k fraction of strands in $\mathcal{S}$ have cost at most t_i with respect to $\widetilde{R}$. In particular, the bound

$$\mathsf{cost}_{\widetilde{R}}(\mathcal{B}_i) \le \widetilde{Q}_{i/k,\widetilde{R}}(\mathcal{S}) \quad (10)$$

for $i \in \{1,\dots,\ell-1\}$ follows from the batch construction process. Similarly, we have that

$$\mathsf{cost}_{\bar{R}}(\mathcal{B}_{k-i}) \le 4n + 1 - \widetilde{Q}_{(k-i-1)/k,\widetilde{R}}(\mathcal{S}) \quad (11)$$

for $i \in \{0,1,\dots,k-\ell-2\}$, due to Lemma 5.1 and the batch construction process.

The bounds in (10) and (11) both involve empirical quantiles, for which Lemma 4.1 provides uniform bounds. We apply Lemma 4.1 with $\varepsilon = 1/(4k)$. This implies that, with probability at least $1 - 2\exp\left(-M/(8k^2)\right)$, we have that

$$\widetilde{Q}_{i/k,\widetilde{R}}(\mathcal{S}) \le Q_{(i+1/4)/k,\widetilde{R}}(\mathcal{D})$$

for all $i \in \{1,\dots,\ell-1\}$, and furthermore that

$$\widetilde{Q}_{(k-i-1)/k,\widetilde{R}}(\mathcal{S}) \ge Q_{(k-i-5/4)/k,\widetilde{R}}(\mathcal{D})$$

for every $i \in \{0,1,\dots,k-\ell-2\}$. Plugging these bounds into (10) and (11) we obtain that, with probability at least $1 - 2\exp\left(-M/(8k^2)\right)$, the total cost is bounded above by

$$\begin{aligned}
&\sum_{i=1}^{\ell} \mathsf{cost}_{\widetilde{R}}(\mathcal{B}_i) + \sum_{i=\ell+1}^{k} \mathsf{cost}_{\bar{R}}(\mathcal{B}_i) \\
&\le \sum_{i=1}^{\ell-1} Q_{(i+1/4)/k,\widetilde{R}}(\mathcal{D}) + \left\{\mathsf{cost}_{\widetilde{R}}(\mathcal{B}_\ell) + \mathsf{cost}_{\bar{R}}(\mathcal{B}_{\ell+1})\right\} \\
&\quad + \sum_{i=0}^{k-\ell-2} \left(4n + 1 - Q_{(k-i-5/4)/k,\widetilde{R}}(\mathcal{D})\right). \quad (12)
\end{aligned}$$

In the rest of the proof we bound from above these two sums, as well as the term in the middle.

First, we claim that for every fixed $\delta > 0$ there exists $\alpha = \alpha(\delta) > 0$ such that

$$Q_{1/2-\delta,\widetilde{R}}(\mathcal{D}) \le 2n - \alpha\sqrt{n}. \quad (13)$$

To see this, recall that the cost of printing a strand S using $\widetilde{R}$ is a sum of independent random variables: the cost of the first character is uniform in $\{1,2,3,4\}$ and the cost of the remaining characters are uniform in $\{1,2,3\}$ (since the strand S does not have homopolymers). Therefore $\mathsf{cost}_{\widetilde{R}}(S)$ is approximately Gaussian with mean $2n+1/2$ and variance on the order of n. This means that if we consider a quantile that is bounded away from the median, then the cost is smaller than the mean by at least a constant factor of the standard deviation. Since the standard deviation is on the order of $\sqrt{n}$, this implies (13).

Turning back to (12), consider the first $\lceil k/3 \rceil$ terms of the first sum in (12) (note that $\lceil k/3 \rceil \geq 1$ since $k \geq 3$). Note that $(\lceil k/3 \rceil + 1/4)/k \leq 5/12 = 1/2 - 1/12$ for all $k \geq 3$. Let $\alpha_* := \alpha(1/12) > 0$. Then (13) implies that for every $i \in \{1, \ldots, \lceil k/3 \rceil\}$ we have that

$$\begin{aligned} Q_{(i+1/4)/k,\widetilde{R}}(\mathcal{D}) &\leq Q_{(\lceil k/3 \rceil + 1/4)/k,\widetilde{R}}(\mathcal{D}) \\ &\leq Q_{5/12,\widetilde{R}}(\mathcal{D}) \leq 2n - \alpha_* \sqrt{n}. \end{aligned}$$

Note also that for every $i \in \{1, \ldots, \ell - 1\}$ we have that $(i + 1/4)/k \leq (\ell - 3/4)/k \leq (k/2 - 1/4)/k = 1/2 - 1/(4k) < 1/2$. Together with (13), this implies that for every $i \in \{1, \ldots, \ell - 1\}$ we have that

$$Q_{(i+1/4)/k,\widetilde{R}}(\mathcal{D}) \leq 2n.$$

Putting the bounds in the previous two displays together, we obtain that

$$\begin{aligned} &\sum_{i=1}^{\ell-1} Q_{(i+1/4)/k,\widetilde{R}}(\mathcal{D}) \\ &= \sum_{i=1}^{\lceil k/3 \rceil} Q_{(i+1/4)/k,\widetilde{R}}(\mathcal{D}) + \sum_{\lceil k/3 \rceil + 1}^{\ell-1} Q_{(i+1/4)/k,\widetilde{R}}(\mathcal{D}) \\ &\leq \sum_{i=1}^{\lceil k/3 \rceil} \left(2n - \alpha_* \sqrt{n}\right) + \sum_{\lceil k/3 \rceil + 1}^{\ell-1} 2n \\ &\leq 2(\ell - 1)n - (\alpha_*/3)k\sqrt{n}. \end{aligned} \tag{14}$$

Similarly, note that for every $i \in \{0, 1, \ldots, k - \ell - 2\}$ we have that $(k - i - 5/4)/k \geq 1/2 + 1/(4k) > 1/2$. Therefore, due to the symmetry of the distribution of $\mathsf{cost}_{\widetilde{R}}(S)$, we have that $Q_{(k-i-5/4)/k,\widetilde{R}}(\mathcal{D}) \geq 2n+1$ for every $i \in \{0, 1, \ldots, k-\ell-2\}$. Using this bound in the second sum of (12), together with (14), we obtain that the quantity in (12) is bounded from above by

$$2(k-2)n - (\alpha_*/3)k\sqrt{n} + \left\{\mathsf{cost}_{\widetilde{R}}(\mathcal{B}_\ell) + \mathsf{cost}_{\overline{R}}(\mathcal{B}_{\ell+1})\right\}. \tag{15}$$

Finally, we bound the sum in the curly brackets in (15). Let S' be a strand in $\mathcal{B}_\ell$ such that $\mathsf{cost}_{\widetilde{R}}(\mathcal{B}_\ell) = \mathsf{cost}_{\widetilde{R}}(S')$, and let S'' be a strand in $\mathcal{B}_{\ell+1}$ such that $\mathsf{cost}_{\overline{R}}(\mathcal{B}_{\ell+1}) = \mathsf{cost}_{\overline{R}}(S'')$. By the construction of the batching process we have that $\mathsf{cost}_{\widetilde{R}}(S') \leq \mathsf{cost}_{\widetilde{R}}(S'')$. On the other hand, by Lemma 5.1 we have that $\mathsf{cost}_{\widetilde{R}}(S'') + \mathsf{cost}_{\overline{R}}(S'') = 4n + 1$. Putting all this together we have that

$$\begin{aligned} \mathsf{cost}_{\widetilde{R}}(\mathcal{B}_\ell) + \mathsf{cost}_{\overline{R}}(\mathcal{B}_{\ell+1}) &= \mathsf{cost}_{\widetilde{R}}(S') + \mathsf{cost}_{\overline{R}}(S'') \\ &\leq \mathsf{cost}_{\widetilde{R}}(S'') + \mathsf{cost}_{\overline{R}}(S'') = 4n + 1. \end{aligned}$$

Plugging this back into (15), we obtain that the total cost is bounded from above by $2kn + 1 - (\alpha_*/3)k\sqrt{n}$, as desired, with this bound holding with probability at least $1 - 2\exp\left(-M/(8k^2)\right)$. □

B. Batching for Unrestricted Strands

The main idea of the proof is to analyze the cost quantiles after splitting into batches. In this case, the optimal splitting is a bit easier to define, as we will always use the reference strand $\widetilde{R} = (ACGT)^*$. Once the reference strand is fixed, it is easy to see that for arbitrary strands, the optimal batching process involves first taking M/k lowest cost strands and then the next M/k lowest cost and so on.

Proof of Theorem 1.2(2): In this argument, we consider a set $\mathcal{S}$ of M strands sampled i.i.d. from the uniform distribution $\mathcal{U}$ over $\{A, C, G, T\}^n$. For simplicity of notation, we let $\widetilde{Q}_q = \widetilde{Q}_{q,\widetilde{R}}(\mathcal{S})$ and $Q_q = Q_{q,\widetilde{R}}(\mathcal{U})$ denote the empirical and distributional q-quantiles of the cost, respectively. We utilize the empirical quantiles i/k for $i \in [k]$, where we recall that $\widetilde{Q}_{i/k}$ denotes the minimum value t_i such that an i/k fraction of strands in $\mathcal{S}$ have cost at most t_i. In particular, the bound $\mathsf{cost}_{\widetilde{R}}(\mathcal{B}_i) \leq \widetilde{Q}_{i/k}$ follows from the batch construction process.

Lemma 4.1 provides a uniform bound on each empirical quantile, and we apply it with $\varepsilon = 1/k$; specifically, this implies that, with probability at least $1 - 2\exp(-\frac{2M}{k^2})$, we have for all $i \in [k-1]$ that $\widetilde{Q}_{i/k} \leq Q_{(i+1)/k}$. In the following we assume that we are on this event; in particular, on this event we have that $\mathsf{cost}_{\widetilde{R}}(\mathcal{B}_i) \leq Q_{(i+1)/k}$ holds for all $i \in [k-1]$.

We next claim that for any even $k' \leq k - 3$ we have that

$$\sum_{i=1}^{k'} Q_{(i+1)/k} \leq 2.5nk'. \tag{16}$$

Before showing (16), we conclude the proof assuming that it holds. When k is even, we set $k' = k - 4$, and otherwise, $k' = k - 3$. Then, for i in the range $k' < i \leq k$, we simply use the bound from Theorem 3.1 that shows that $\mathsf{cost}_{\widetilde{R}}(\mathcal{B}_i) \leq 2.5n + 3\sqrt{n \ln M}$. Combining this with (16) implies the desired bound with $C_2 = 12$.

We now turn to proving (16). We have already seen in Lemma 5.1 that the cost distribution is symmetric around $2.5n$ for strands uniform over $\{A, C, G, T\}^n$. This implies that

$$Q_{(i+1)/k} + Q_{(k-i-2)/k} \leq 2 \cdot 2.5n \tag{17}$$

for each $i \leq k'/2$. Here we have paired up $(i+1)/k$ with $(k - i - 2)/k$, and hence, we have chosen a cost quantile on either side of the mean, but with a slight asymmetry, shifting the larger one over by one. The key observation is that $Q_{(k-i-2)/k}$ deviates from the mean less than $Q_{(i+1)/k}$. More precisely, we have that $Q_{(i+1)/k} \leq 2.5n \leq Q_{(k-i-2)/k}$, and furthermore, $Q_{(k-i-2)/k} - 2.5n \leq 2.5n - Q_{(i+1)/k}$. As this holds for each $i \leq k'/2$, the inequality in (16) follows. □

VI. Lower Bounds

In this section we prove Theorem 1.3, which provides lower bounds on the cost of batched DNA synthesis for random

strands, for both unconstrained strands and strands without homopolymers.

We first introduce some notation and terminology. For a strand S, we say that a base S_i is a repetition of the previous base if this base is the same as the previous base, that is, if $S_i = S_{i-1}$. Let $d(S)$ be the number of bases that are repetitions of the previous base in S (that is, $d(S) := |\{i : S_i = S_{i+1}\}|$). For every $p \in [0,1]$ and positive integer n, define a distribution $D_{p,n}$ on DNA strands of length n by letting the probability of a strand S of length n according to $D_{p,n}$ be equal to

$$D_{p,n}(S) := \frac{1}{4} \cdot \Big(\frac{1-p}{3}\Big)^{n-d(S)-1} p^{d(S)}.$$

We can think of this distribution as follows. The first base of the strand is chosen uniformly at random from the alphabet $\{A, C, G, T\}$. Then, every consecutive base is a repetition of the previous base with probability p and another base with probability $(1-p)$. If a base is not a repetition of the previous base, then it is chosen uniformly at random among the remaining three bases.

In this paper, the two most important distributions on DNA strands are $D_{0,n}$ and $D_{1/4,n}$. The former is the uniform distribution on strands of length n without homopolymers; the latter is the uniform distribution on strands of length n (unconstrained, i.e., allowing homopolymers). In this section, we will also crucially use another distribution: $D_{1/4+\delta,n}$ with $\delta \approx \sqrt{\frac{\log M}{n}}$. To simplify notation, we will omit the second parameter of the distribution D and write $D_p \equiv D_{p,n}$, since the length of all strands we consider is n.

We will prove the following theorem, from which Theorem 1.3 readily follows.

Theorem 6.1: Let $\mathcal{S} \equiv \mathcal{S}_{D_p,M}$ be a set of M i.i.d. strands from the distribution D_p, where p is 0 or $1/4$. Suppose that M is divisible by k, and $k \le \sqrt{M/(24n)}$. Then, with probability at least $1 - 2\exp(-M/(4k^2)) \ge 1 - 2\exp(-6n)$, the optimal cost of printing the strands in $\mathcal{S}$ using k batches of equal size is at least

$$k \cdot (2n - \sqrt{5n\log(2k)}) \qquad \text{for strands w/o homopolymers;} \tag{18}$$

$$k \cdot (2.5n - \sqrt{5n\log(2k)}) \qquad \text{for unconstrained strands.} \tag{19}$$

Furthermore, there exist positive absolute constants c_1, c_2, and c_3 such that the following holds in the setting of unconstrained strands (that is, when $p = 1/4$). Suppose that the number of batches satisfies $k \le c_1 \min\{\sqrt{\log M/\log\log M}, \sqrt{n/\log n}, \sqrt[4]{M}/\sqrt{n}\}$. Then, with probability at least $1 - \exp(-c_3\sqrt{M}/k^2)$, the total cost of printing the strands in $\mathcal{S}_{D_{1/4},M}$ using k batches of equal size is at least

$$k \cdot \Big(2.5n + c_2 \frac{\sqrt{n \cdot \min\{\log M, n\}}}{k}\Big). \tag{20}$$

Proof: We use the approach outlined in the proof overview (see Section II). We first prove the bounds (18) and (19). To do so, we show that with high probability the cost of every batch that contains at least n/k strands is greater than $2n - \sqrt{5n\log(2k)}$ for random strands without homopolymers and $2.5n - \sqrt{5n\log(2k)}$ for unconstrained random strands. Consequently, the total cost of printing k batches is at least $k \cdot (2n - \sqrt{5n\log(2k)})$ and $k \cdot (2.5n - \sqrt{5n\log(2k)})$ for random strands without and with homopolymers, respectively.

Let $\mathcal{R}^*$ be the set of all *reasonable* reference strands for printing strands of lengths n, that is, $\mathcal{R}^*$ is the set of all possible strands of length $4n$ appended with the infinite repeating sequence $(ACGT)^*$ (see Section II). The size of this set is $|\mathcal{R}^*| = 4^{4n} < e^{6n}$. Recall the notions of the q-quantile $Q_{q,\mathcal{R}}(D)$ of the distribution D and the empirical q-quantile $\widetilde{Q}_{q,\mathcal{R}}(\mathcal{S})$ of a sample $\mathcal{S}$ that we introduced in Section IV. Since every batch in the optimal partitioning of the set $\mathcal{S}$ contains a $1/k$ fraction of all strands, its cost is at least $\widetilde{Q}_{1/k,\mathcal{R}^*}(\mathcal{S})$. In Lemma 4.1 we showed that $\widetilde{Q}_{q,\mathcal{R}}(\mathcal{S}) \ge Q_{q-\varepsilon,\mathcal{R}}(D)$ with probability close to 1 for sufficiently large M. Specifically, using Lemma 4.1 with parameters $q = 1/k$, $\varepsilon = q/2$, and $\mathcal{R} = \mathcal{R}^*$, we obtain the following bound:

$$\mathbb{P}\Big\{\widetilde{Q}_{1/k,\mathcal{R}^*}(\mathcal{S}_{D,M}) \ge Q_{1/(2k),\mathcal{R}^*}(D)\Big\} \ge 1 - 2|\mathcal{R}^*|e^{-M/(2k^2)} > 1 - 2e^{-M/(4k^2)}, \tag{21}$$

where in the second inequality we used that $|\mathcal{R}^*| < \exp(6n) \le e^{M/(4k^2)}$; this holds due to the condition $k \le \sqrt{M/(24n)}$ that is assumed in the statement of Theorem 6.1.

We now obtain lower bounds on $Q_{q,\mathcal{R}^*}(D_0)$, $Q_{q,\mathcal{R}^*}(D_{1/4})$, and $Q_{q,\mathcal{R}^*}(D_{1/4+\delta})$.

Lemma 6.2: For $q \in (0,1)$ we have that

$$Q_{q,\mathcal{R}^*}(D_0) \ge 2n - \sqrt{5n\log 1/q}; \tag{22}$$

$$Q_{q,\mathcal{R}^*}(D_{1/4}) \ge 2.5n - \sqrt{5n\log 1/q}. \tag{23}$$

Lemma 6.3: For $q \in (0,1)$ and $\delta \in (0, 1/600]$ we have that

$$Q_{q,\mathcal{R}^*}(D_{1/4+\delta}) \ge 2.5n + 2/3\ \delta n - 5\sqrt{n\log 1/q}. \tag{24}$$

We prove Lemmas 6.2 and 6.3 in Section VI-A. Lemma 6.2, combined with inequality (21), immediately yield the lower bounds in (18) and (19) on the total cost of printing random strands without and with homopolymers.

We now show how to strengthen the lower bound for unconstrained strands, obtaining the desired inequality (20). We will prove that, with probability at least $1 - \exp(-c'\sqrt{M}/k^2)$ for some absolute constant $c' > 0$, there exists a batch with cost at least

$$2.5n + 2/3\ \delta n - 5\sqrt{n\log(6k)}, \tag{25}$$

where $\delta = \min\Big\{\sqrt{\frac{\log(M/16)}{16n}}, 1/600\Big\}$. Thus, using the lower bound from the first part of the theorem for all the other $k-1$ batches, with probability at least

$$1 - \exp(-c'\sqrt{M}/k^2) - 2\exp(-M/(4k^2)) \ge 1 - \exp(-c_3\sqrt{M}/k^2),$$

the total cost of printing $\mathcal{S}_{D_{1/4},M}$ using k batches is at least

$$\begin{aligned}&\left(2.5n + {}^2\!/_3\,\delta n - 5\sqrt{n\log(6k)}\right)\\&\quad + (k-1)\cdot\left(2.5n - \sqrt{5n\log(2k)}\right)\\&\geq k\cdot 2.5n + {}^2\!/_3\,\delta n - 5\ k\sqrt{n\log(6k)}\\&\geq k\cdot 2.5n\\&\quad + \sqrt{n}\left[\frac{1}{900}\min\left\{\sqrt{\log(M/16)},\sqrt{n}\right\} - 5\ k\sqrt{\log(6k)}\right].\end{aligned}$$

Now if $k \leq c_1 \min\{\sqrt{\log M/\log\log M}, \sqrt{n/\log n}\}$ for a small enough $c_1 > 0$, then the quantity in the display above is bounded from below by $k\cdot 2.5n + c_2\sqrt{n\min\{\log M, n\}}$ for some $c_2 > 0$. Thus we have obtained (20).

What remains is to prove the claim in (25). The proof of this relies on the following lemma, the proof of which we defer to Section VI-B.

Lemma 6.4: For every $0 \leq \delta \leq \min\left\{\sqrt{\frac{\log(M/16)}{16n}}, 0.1\right\}$ there exists a coupling $(\mathcal{S},\mathcal{S}')$ of random multisets containing strands of length n that satisfies the following:

1) $\mathcal{S}$ has the same distribution as $\mathcal{S}_{D_{1/4},M}$;
2) $\mathcal{S}'$ has the same distribution as $\mathcal{S}_{D_{1/4+\delta},\lfloor\sqrt{M}\rfloor}$; and
3) there exists an absolute constant $c > 0$ such that

$$\mathbb{P}\{|\mathcal{S}\cap\mathcal{S}'| \geq |\mathcal{S}'|/3\} \geq 1 - \exp(-c\sqrt{M}).$$

Consider the pair of random multisets $(\mathcal{S},\mathcal{S}')$ from Lemma 6.4. Since $\mathcal{S}$ has the same distribution as $\mathcal{S}_{D_{1/4},M}$, it suffices to show that, with probability at least $1 - \exp(-c'\sqrt{M}/k^2)$, in every partitioning of $\mathcal{S}$ into k batches there exists at least one batch with cost at least $2.5n + {}^2\!/_3\,\delta n - 5\sqrt{n\log(6k)}$. Let $\mathcal{S}'' := \mathcal{S}\cap\mathcal{S}'$. Since $\mathcal{S}''\subseteq\mathcal{S}$, one of the batches must contain at least a $1/k$ fraction of all strands in $\mathcal{S}''$. Denote this batch by $\mathcal{B}$ and let $\mathcal{B}'' := \mathcal{B}\cap\mathcal{S}''$. If $|\mathcal{S}''| \geq |\mathcal{S}'|/3$, then $|\mathcal{B}''| \geq |\mathcal{S}''|/k \geq |\mathcal{S}'|/(3k)$. Consequently, the cost of printing $\mathcal{B}''$ is at least $\widetilde{Q}_{1/(3k),\mathcal{R}^*}(\mathcal{S}')$. Since $\mathcal{B}''\subseteq\mathcal{B}$, the cost of printing $\mathcal{B}$ is also at least $\widetilde{Q}_{1/(3k),\mathcal{R}^*}(\mathcal{S}')$. By Lemma 6.4, we have that $|\mathcal{S}''| \geq |\mathcal{S}'|/3$ with probability at least $1 - \exp(-c\sqrt{M})$. Thus, with this probability, the cost of printing $\mathcal{B}$ is at least $\widetilde{Q}_{1/(3k),\mathcal{R}^*}(\mathcal{S}')$.

Now by the inequality (21) (replacing $1/k$ with $1/(3k)$ and M with $\lfloor\sqrt{M}\rfloor$) we have that

$$\begin{aligned}&\mathbb{P}\left\{\widetilde{Q}_{1/(3k),\mathcal{R}^*}(\mathcal{S}') \geq Q_{1/(6k),\mathcal{R}^*}(D_{1/4+\delta})\right\}\\&\qquad\geq 1 - 2|\mathcal{R}^*|e^{-\sqrt{M}/(20k^2)} > 1 - 2\ e^{-\sqrt{M}/(40k^2)},\end{aligned}$$

where in the second inequality we used that $|\mathcal{R}^*| < \exp(6n) \leq \exp(\sqrt{M}/(40k^2))$; this holds due to the assumption that $k \leq \sqrt[4]{M}/\sqrt{240n}$. Finally, by Lemma 6.3 we have that

$$Q_{1/(6k),\mathcal{R}^*}(D_{1/4+\delta}) \geq 2.5n + {}^2\!/_3\,\delta n - 5\sqrt{n\log(6k)}.$$

Putting everything together we obtain that, with probability at least $1 - 2\exp(-\sqrt{M}/(40k^2)) - \exp(-c\sqrt{M})$, the cost of printing $\mathcal{B}$ is at least $2.5n + {}^2\!/_3\,\delta n - 5\sqrt{n\log(6k)}$. This concludes the proof of the claim in (25).

We complete the proof of Theorem 6.1 in Sections VI-A and VI-B, where we prove Lemmas 6.2, 6.3, and 6.4. □

A. Proofs of Lemma 6.2 and Lemma 6.3

Proof of Lemma 6.2: We show that $Q_{q,\mathcal{R}^*}(D_0) \geq 2n - \Delta_{n,q}$ and $Q_{q,\mathcal{R}^*}(D_{1/4}) \geq 2.5n - \Delta_{n,q}$, where $\Delta_{n,q} := \sqrt{5n\log 1/q}$. To establish this inequality, it suffices to prove that for all $R\in\mathcal{R}^*$,

$$\begin{aligned}\mathbb{P}_{S\sim D_0}\{\mathsf{cost}_R(S) \ \leq 2n - \Delta_{n,q}\} < q;\\ \mathbb{P}_{S\sim D_{1/4}}\{\mathsf{cost}_R(S) \ \leq 2.5n - \Delta_{n,q}\} < q.\end{aligned}$$

To do this, we use the stochastic domination argument that we previously used in Section III-C. Let $Y_1,\ldots,Y_n$ be i.i.d. random variables that are uniformly distributed on $\{1,2,3\}$, and let $Z_1,\ldots,Z_n$ be i.i.d. random variables that are uniformly distributed on $\{1,2,3,4\}$. Then, by the same arguments as in Section III-C, we have that

$$\mathbb{P}_{S\sim D_0}\{\mathsf{cost}_R(S) \ \leq 2n-\Delta_{n,q}\} \leq \mathbb{P}\left\{\sum_{i=1}^n Y_i \leq 2n - \Delta_{n,q}\right\}$$

and

$$\begin{aligned}&\mathbb{P}_{S\sim D_{1/4}}\{\mathsf{cost}_R(S) \ \leq 2.5n - \Delta_{n,q}\}\\&\qquad\leq \mathbb{P}\left\{\sum_{i=1}^n Z_i \leq 2.5n - \Delta_{n,q}\right\}.\end{aligned}$$

Note that $\mathbb{E}[Y_1] = 2$ and $\mathbb{E}[Z_1] = 2.5$. Thus by Hoeffding's inequality (Theorem 1.1) we have that

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^n Y_i \leq 2n - \Delta_{n,q}\right\} \leq \exp\left(-\frac{2\Delta_{n,q}^2}{4n}\right) < q;\\ \mathbb{P}\left\{\sum_{i=1}^n Z_i \leq 2.5n - \Delta_{n,q}\right\} \leq \exp\left(-\frac{2\Delta_{n,q}^2}{9n}\right) < q.\end{aligned}$$

This completes the proof of (22) and (23). □

Proof of Lemma 6.3: To prove the claim it suffices to show that for all $R\in\mathcal{R}^*$ we have that

$$\mathbb{P}_{S\sim D_{1/4+\delta}}\left\{\mathsf{cost}_R(S) \leq 2.5n + {}^2\!/_3\,\delta n - 5\sqrt{n\log 1/q}\right\} < q. \tag{26}$$

Accordingly, we fix $R\in\mathcal{R}^*$ for the rest of the proof and show (26).

As in Section III-C, for $i \geq 1$ let $\tau_i(S,R)$ denote the index of the base of R that is used to print the ith base of S, and let $\tau_0(S,R) = 0$ for notational convenience. For $i\geq 1$, define $X_i(S,R) := \tau_i(S,R) - \tau_{i-1}(S,R)$, and also let $Y_i := \min\{X_i(S,R), 5\}$. When $S\sim D_{1/4}$, we have seen (see Section III-C) that the distribution of Y_i stochastically dominates the uniform distribution on $\{1,2,3,4\}$, and in particular we have that $\mathbb{E}_{S\sim D_{1/4}}[Y_i \mid \tau_{i-1}(S,R)] \geq 2.5$. When $S\sim D_{1/4+\delta}$, this inequality does not necessarily hold for all $i\geq 1$. However, we still have the following.

Claim 6.5: For all $i\geq 1$ we have that

$$\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_i \mid \tau_{i-1}(S,R)] \geq 2.5 - 2\delta. \tag{27}$$

Proof: Given $\tau_{i-1}(S,R)$ (and the knowledge of the fixed reference strand R), we know the *support* of the random variable $X_i(S,R)$, which consists of four distinct positive integers. We also know that the conditional probabilities of taking on each particular value are given by $1/4+\delta$, $1/4-\delta/3$, $1/4-\delta/3$, and $1/4-\delta/3$, with some particular permutation. The conditional expectation is thus minimized when the support is $\{1,2,3,4\}$ and the largest probability $1/4+\delta$ is assigned to 1. Hence,

$$\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_i \mid \tau_{i-1}(S,R)] \geq (1/4+\delta)\cdot 1 + (1/4-\delta/3)\cdot 2 + (1/4-\delta/3)\cdot 3 + (1/4-\delta/3)\cdot 4 = 2.5-2\delta.$$

□

We now show that, by averaging over three consecutive terms, we can obtain a better lower bound on the conditional expectation that is strictly greater than 2.5 on average.

Lemma 6.6: For every $i \in \{1,\dots,n-2\}$ and $\delta \in (0,1/600]$, we have that

$$\mathbb{E}_{S\sim D_{1/4+\delta}}\left[\frac{Y_i+Y_{i+1}+Y_{i+2}}{3}\,\middle|\,\tau_{i-1}(S,R)\right] \geq 2.5+2/3\,\delta. \tag{28}$$

Proof: For $i\geq 1$ let $\tau^* := \tau_{i-1}(S,R)$ and note that for $i\geq 2$, by definition, R_{τ^*} is the base of the reference strand R which is used for printing base $i-1$ of strand S. Consider the next 12 bases of the reference strand, that is, the substrand $R_{\tau^*+1},\cdots,R_{\tau^*+12}$ of R. We examine two cases.

First, suppose that this substrand is a triple repetition of some permutation of the bases A, C, G, T. For example,

$$R_{\tau^*+1},\cdots,R_{\tau^*+12} = ACGT\,ACGT\,ACGT.$$

Then, for $j\in\{i+1,i+2\}$, the conditional distribution of Y_j given $\tau_{i-1}(S,R)$ is given by: $Y_j=1$, $Y_j=2$, and $Y_j=3$ all with probability $1/4-\delta/3$, and $Y_j=4$ with probability $1/4+\delta$. Thus,

$$\begin{aligned}\mathbb{E}_{S\sim D_{1/4+\delta}}&\left[Y_{i+1}+Y_{i+2}\,|\,\tau_{i-1}(S,R)\right]\\ &= 2\Big((1/4-\delta/3)\cdot(1+2+3)+(1/4+\delta)\cdot 4\Big)\\ &= 2(2.5+2\delta).\end{aligned}$$

By Claim 6.5 we have that $\mathbb{E}_{S\sim D_{1/4+\delta}}\big[Y_i \mid \tau_{i-1}(S,R)\big] \geq 2.5-2\delta$. By adding up these bounds, we obtain the desired bound (28).

Now consider the second case: that the substrand $R_{\tau^*+1},\cdots,R_{\tau^*+12}$ is not a triple repetition of the same permutation of A, C, G, T. Then, this substrand must contain four consecutive bases that are not a permutation of A, C, G, T. Let τ' be the index of the base just before the first such quadruple of bases. By construction we must have $\tau'\in\{\tau^*,\tau^*+1,\dots,\tau^*+8\}$. Since $R_{\tau'+1},R_{\tau'+2},R_{\tau'+3},R_{\tau'+4}$ is not a permutation of A, C, G, T, at least one letter from the alphabet $\{A,C,G,T\}$ is absent in this quadruple. Hence, if $\tau_j(S,R)=\tau'$ for some j, then the expected next jump is at least

$$\begin{aligned}\mathbb{E}_{S\sim D_{1/4+\delta}}&[Y_{j+1} \mid \tau_j(S,R)=\tau']\\ &\geq (1/4+\delta)\cdot 1 + (1/4-\delta/3)\cdot 2\\ &\quad + (1/4-\delta/3)\cdot 3 + (1/4-\delta/3)\cdot 5\\ &= 2.5+(0.25-7/3\,\delta).\end{aligned} \tag{29}$$

We now distinguish three cases depending on whether $\tau'=\tau^*$ or $\tau'\in\{\tau^*+1,\dots,\tau^*+4\}$ or $\tau'\in\{\tau^*+5,\dots,\tau^*+8\}$. First, if $\tau'=\tau^*$, then by (29) we have that $\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_i \mid \tau_{i-1}(S,R)]\geq 2.5+(0.25-7/3\,\delta)$. We also have, by Claim 6.5 and the tower rule, that $\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_j \mid \tau_{i-1}(S,R)]\geq 2.5-2\delta$ for $j\in\{i+1,i+2\}$. Putting these bounds together we have, if $\tau'=\tau^*$, that

$$\begin{aligned}\mathbb{E}_{S\sim D_{1/4+\delta}}&\left[Y_i+Y_{i+1}+Y_{i+2}\,|\,\tau_{i-1}(S,R)\right]\\ &\geq 2.5+(0.25-7/3\,\delta)+2(2.5-2\delta)\\ &= 3\cdot 2.5+0.25-19/3\,\delta \geq 3(2.5+2/3\,\delta),\end{aligned}$$

where the last inequality holds when $\delta\leq 0.03$. This shows (28) in this case.

Next, suppose that $\tau'\in\{\tau^*+1,\dots,\tau^*+4\}$. By Claim 6.5 we have that

$$\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_i+Y_{i+2} \mid \tau_{i-1}(S,R)] \geq 2(2.5-2\delta). \tag{30}$$

In order to bound $\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_{i+1} \mid \tau_{i-1}(S,R)]$, we first condition on $\tau_i(S,R)$. Note that since $\tau'>\tau^*$, the bases $R_{\tau^*+1},\dots,R_{\tau^*+4}$ are all different. Therefore we must have that $\tau_i(S,R)\in\{\tau^*+1,\dots,\tau^*+4\}$. Conditioning on the value of $\tau_i(S,R)$ we thus have that

$$\begin{aligned}&\mathbb{E}[Y_{i+1} \mid \tau_{i-1}(S,R)]\\ &= \sum_{\ell=1}^{4}\mathbb{E}[Y_{i+1} \mid \tau_i(S,R)=\tau^*+\ell]\times\\ &\qquad\times\mathbb{P}\left\{\tau_i(S,R)=\tau^*+\ell\,|\,\tau_{i-1}(S,R)\right\}\\ &= \mathbb{E}[Y_{i+1} \mid \tau_i(S,R)=\tau']\mathbb{P}\left\{\tau_i(S,R)=\tau'\,|\,\tau_{i-1}(S,R)\right\}\\ &\quad+\mathbb{E}[Y_{i+1} \mid \tau_i(S,R)\neq\tau']\times\\ &\qquad\times\left(1-\mathbb{P}\left\{\tau_i(S,R)=\tau'\,|\,\tau_{i-1}(S,R)\right\}\right),\end{aligned}$$

where all expectations and probabilities are under $S\sim D_{1/4+\delta}$. We then use (29) to bound from below the first expectation in the sum above, and we use Claim 6.5 to bound from below the second expectation. Plugging in these bounds we obtain that

$$\begin{aligned}&\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_{i+1} \mid \tau_{i-1}(S,R)]\\ &\geq 2.5-2\delta\\ &\quad+(0.25-\delta/3)\mathbb{P}_{S\sim D_{1/4+\delta}}\left\{\tau_i(S,R)=\tau'\,|\,\tau_{i-1}(S,R)\right\}\\ &\geq 2.5-2\delta+(0.25-\delta/3)^2 \geq 2.5+6\delta,\end{aligned} \tag{31}$$

where the last inequality holds when $\delta\leq 1/200$. Putting together (30) and (31) we obtain (28) in this case as well.

Finally, suppose that $\tau'\in\{\tau^*+5,\dots,\tau^*+8\}$. By Claim 6.5 we have that

$$\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_i+Y_{i+1} \mid \tau_{i-1}(S,R)] \geq 2(2.5-2\delta). \tag{32}$$

In order to bound $\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_{i+2} \mid \tau_{i-1}(S,R)]$, we first condition on $\tau_{i+1}(S,R)$. Note that since $\tau' > \tau^* + 4$, the bases $R_{\tau^*+1},\ldots,R_{\tau^*+4}$ are all different, and the bases $R_{\tau^*+5},\ldots,R_{\tau^*+8}$ are a repetition of the bases $R_{\tau^*+1},\ldots,R_{\tau^*+4}$. This implies that $\tau_i(S,R) \in \{\tau^*+1,\ldots,\tau^*+4\}$, and, moreover, that given $\tau_i(S,R)$, we have that $\tau_{i+1}(S,R) \in \{\tau_i(S,R)+1,\ldots,\tau_i(S,R)+4\}$. In particular, note that if $\tau_i(S,R) = \tau' - 4$ and $S_{i+1} = S_i$, then $\tau_{i+1}(S,R) = \tau'$. Thus conditioning on whether or not $\tau_{i+1}(S,R) = \tau'$ occurs, we have that

$$\begin{aligned}&\mathbb{E}[Y_{i+2} \mid \tau_{i-1}(S,R)]\\ &= \mathbb{E}[Y_{i+2} \mid \tau_{i+1}(S,R)=\tau']\mathbb{P}\left\{\tau_{i+1}(S,R)=\tau' \mid \tau_{i-1}(S,R)\right\}\\ &\quad + \mathbb{E}[Y_{i+2} \mid \tau_{i+1}(S,R)\neq\tau']\times\\ &\quad\quad \times\left(1-\mathbb{P}\left\{\tau_{i+1}(S,R)=\tau' \mid \tau_{i-1}(S,R)\right\}\right).\end{aligned}$$

As before, we use (29) to bound from below the first expectation in the sum above, and we use Claim 6.5 to bound from below the second expectation. Plugging in these bounds we obtain that

$$\begin{aligned}&\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_{i+2} \mid \tau_{i-1}(S,R)] \geq 2.5-2\delta\\ &+(0.25-\delta/3)\mathbb{P}_{S\sim D_{1/4+\delta}}\left\{\tau_{i+1}(S,R)=\tau' \mid \tau_{i-1}(S,R)\right\}.\end{aligned}$$

Since $\tau_i(S,R) = \tau' - 4$ and $S_{i+1} = S_i$ together imply that $\tau_{i+1}(S,R) = \tau'$, we have that

$$\begin{aligned}&\mathbb{P}_{S\sim D_{1/4+\delta}}\left\{\tau_{i+1}(S,R)=\tau' \mid \tau_{i-1}(S,R)\right\}\\ &\geq \mathbb{P}_{S\sim D_{1/4+\delta}}\left\{\tau_i(S,R)=\tau'-4 \text{ and } S_{i+1}=S_i \mid \tau_{i-1}(S,R)\right\}\\ &\geq (1/4-\delta/3)\,(1/4+\delta) > 1/16.\end{aligned}$$

Putting together the two previous displays we obtain that

$$\begin{aligned}\mathbb{E}_{S\sim D_{1/4+\delta}}[Y_{i+2} \mid \tau_{i-1}(S,R)] &\geq 2.5+1/64-(2+1/48)\delta\\ &\geq 2.5+6\delta, \qquad (33)\end{aligned}$$

where the last inequality holds whenever $\delta \leq 1/600$. Putting together (32) and (33) we obtain (28) in this case as well. This concludes all cases and hence concludes the proof. □

We now conclude the proof of (26). For simplicity, we assume in the following that n is divisible by 3; the case of n not being divisible by 3 can be handled by padding. Let $Z_0 := 0$ and for $t \in \{1,\ldots,n/3\}$ let

$$Z_t := \sum_{i=1}^{3t} Y_i - 3(2.5+2/3\,\delta)t.$$

By definition we have that

$$\mathsf{cost}_R(S) \;= \sum_{i=1}^{n} X_i(S,R) \geq \sum_{i=1}^{n} Y_i = Z_{n/3} + (2.5+2/3\,\delta)n.$$

and so

$$\begin{aligned}&\mathbb{P}_{S\sim D_{1/4+\delta}}\left\{\mathsf{cost}_R(S) \leq 2.5n+2/3\;\delta n - 5\sqrt{n\log 1/q}\right\}\\ &\qquad \leq \mathbb{P}_{S\sim D_{1/4+\delta}}\left\{Z_{n/3} \leq -5\sqrt{n\log 1/q}\right\}. \qquad (34)\end{aligned}$$

Lemma 6.6 implies that $\{Z_t\}_{t=0}^{n/3}$ is a submartingale with respect to its natural filtration. Since $Y_i \in \{1,\ldots,5\}$ for every $i \geq 1$, the submartingale differences satisfy $Z_{t+1} - Z_t = (Y_{3t+1}+Y_{3t+2}+Y_{3t+3}) - 3(2.5+2/3\delta) \in [-4.5-2\delta, 7.5-2\delta]$, that is, they take values in an interval of length 12. Thus by the Azuma-Hoeffding inequality (see, e.g., [54]) we have that

$$\begin{aligned}&\mathbb{P}_{S\sim D_{1/4+\delta}}\left\{Z_{n/3} \leq -5\sqrt{n\log 1/q}\right\}\\ &\qquad \leq \exp\left(-\frac{2\cdot 5^2 n\log 1/q}{(n/3)\cdot 12^2}\right) = q^{25/24} < q. \qquad (35)\end{aligned}$$

Putting together (34) and (35) we obtain (26). □

B. Proof of Lemma 6.4

Proof of Lemma 6.4: For simplicity of exposition, assume that $\sqrt{M}$ is an integer number (if it is not, we can replace M with $M' = \lfloor\sqrt{M}\rfloor^2$). Let $\Sigma := \{A,C,G,T\}$ and denote by $\mathcal{L}$ the set of all strands S of length n with the number of repeating bases, $d(S)$, being at most $\lceil(1/4+\delta)n\rceil$:

$$\mathcal{L} = \{S \in \Sigma^n : d(S) \leq \lceil(1/4+\delta)n\rceil\}.$$

Let $\phi := 1/\sqrt{M}$. We first show the following claim.

Claim 6.7: Suppose that $0 \leq \delta \leq \min\left\{\sqrt{\frac{\log(M/16)}{16n}}, 0.1\right\}$. For all strands $S \in \mathcal{L}$ we have that

$$D_{1/4}(S) \geq 2\phi \cdot D_{1/4+\delta}(S). \qquad (36)$$

Proof: First observe that for all $S \in \Sigma^n$ we have that

$$\begin{aligned}\frac{D_{1/4+\delta}(S)}{D_{1/4}(S)} &= \frac{\frac{1}{4}\cdot\left(\frac{1}{4}-\frac{\delta}{3}\right)^{n-d(S)-1}\left(\frac{1}{4}+\delta\right)^{d(S)}}{\frac{1}{4^n}}\\ &= \left(1-4/3\,\delta\right)^{n-d(S)-1}\left(1+4\delta\right)^{d(S)}.\end{aligned}$$

For $S \in \mathcal{L}$ we have that $d(S) \leq (1/4+\delta)n+1$ by definition. Plugging this bound into the display above we obtain for $S \in \mathcal{L}$ that

$$\begin{aligned}\frac{D_{1/4+\delta}(S)}{D_{1/4}(S)} &\leq \left(1-4/3\,\delta\right)^{(3/4-\delta)n-2}\left(1+4\delta\right)^{(1/4+\delta)n+1}\\ &= \left((1-4/3\;\delta)^3(1+4\delta)\right)^{n/4}\cdot\left(\frac{1+4\delta}{1-4/3\;\delta}\right)^{\delta n}\cdot\\ &\quad\cdot\left(\frac{(1+4\delta)}{(1-4/3\;\delta)^2}\right).\end{aligned}$$

We now bound from above the three factors on the right hand side. The first factor is at most 1, since $(1-4/3\;\delta)^3(1+4\delta) \leq 1$ by the AM-GM inequality. The third factor is bounded from above by 2 for $\delta \in [0, 1/10]$. Finally, to bound the second factor, observe that the quadratic function

$$(1+4x)-(1-4x/3)(1+8x) = \frac{32\,x(x-1/4)}{3}$$

is negative for $x \in (0,1/4)$. Hence, $(1+4\delta)/(1-4/3\;\delta) \leq (1+8\delta)$ for $\delta \in [0,1/4]$ and so

$$\left(\frac{1+4\delta}{1-4/3\;\delta}\right)^{\delta n} \leq (1+8\delta)^{\delta n} = e^{\delta n\ln(1+8\delta)} \leq e^{8\delta^2 n}.$$

Combining these bounds we obtain that

$$D_{1/4}(S) \geq 1/2\; e^{-8\delta^2 n}\cdot D_{1/4+\delta}(S) \geq 2\phi\cdot D_{1/4+\delta}(S),$$

where in the last inequality we used that $0 \leq \delta \leq \sqrt{\frac{\log(M/16)}{16n}}$ and that $\phi = 1/\sqrt{M}$. □

We also need to estimate the probability of the set $\mathcal{L}$.

Claim 6.8: We have that $D_{1/4+\delta}(\mathcal{L}) \geq 1/2$.

Proof: The number of repetitions $d(S)$ for random strands $S \sim D_p$ has a binomial distribution $B(n-1, p)$. The median of the binomial distribution $B(n-1, p)$ is at most $\lceil (n-1)p \rceil$. Thus, $\mathbb{P}_{S \sim D_p}\{d(S) \leq \lceil (n-1)p \rceil\} \geq 1/2$ and hence $D_{1/4+\delta}(\mathcal{L}) \geq 1/2$. □

We now return to the proof of Lemma 6.4. Define two probability measures A and B on Σ^n:

$$A(S) := \begin{cases} \frac{D_{1/4+\delta}(S)}{D_{1/4+\delta}(\mathcal{L})}, & \text{for } S \in \mathcal{L}; \\ 0, & \text{for } S \notin \mathcal{L}; \end{cases}$$

and

$$B(S) := \frac{D_{1/4}(S) - \phi\, A(S)}{1-\phi}.$$

Claim 6.9: A and B are probability measures on Σ^n. That is, A and B are nonnegative and $A(\Sigma^n) = B(\Sigma^n) = 1$.

Proof: If $S \sim D_{1/4+\delta}$, then the conditional distribution of S given that $S \in \mathcal{L}$ is exactly given by A. In other words, for every $S' \in \Sigma^n$ we have that $\mathbb{P}_{S \sim D_{1/4+\delta}}(S = S' \mid S \in \mathcal{L}) = A(S')$. Therefore A is a probability measure. It is easy to verify that $B(\Sigma^n) = 1$:

$$B(\Sigma^n) = \frac{D_{1/4}(\Sigma^n) - \phi\, A(\Sigma^n)}{1-\phi} = \frac{1-\phi}{1-\phi} = 1.$$

Thus, it remains to show that B is nonnegative. For $S \notin \mathcal{L}$ we have that $A(S) = 0$, so $B(S) = D_{1/4}(S)/(1-\phi) \geq 0$. By Claim 6.8 we know that $D_{1/4+\delta}(\mathcal{L}) \geq 1/2$, so for every $S \in \mathcal{L}$ we have that $A(S) \leq 2\, D_{1/4+\delta}(S)$. Therefore for every $S \in \mathcal{L}$ we have that

$$B(S) = \frac{D_{1/4}(S) - \phi\, A(S)}{1-\phi} \geq \frac{D_{1/4}(S) - 2\phi\, D_{1/4+\delta}(S)}{1-\phi} \geq 0,$$

where the last inequality is a consequence of Claim 6.7. □

We now show how to define the coupling $(\mathcal{S}, \mathcal{S}')$ satisfying the three conditions of Lemma 6.4. Let $S_1, S_2, \ldots,$ be an infinite sequence of i.i.d. random strands, where each strand S_i is distributed according to the distribution $D_{1/4+\delta}$. Let $L_1, L_2, \ldots,$ be the subsequence of strands that belong to the set $\mathcal{L}$. Note in particular that, by construction, $L_1, L_2, \ldots,$ are i.i.d. strands distributed according to the distribution A.

First, let $\mathcal{S}' := \{S_1, S_2, \ldots, S_{\sqrt{M}}\}$. By construction we have that $\mathcal{S}'$ has the same distribution as $\mathcal{S}_{D_{1/4+\delta}, \sqrt{M}}$, which shows part (a) of the claim.

Next, we define $\mathcal{S}$. We generate the strands of $\mathcal{S}$ one by one. For each $i = 1, \ldots, M$, we flip a coin: with probability ϕ, we add the first not yet selected strand from the sequence $\{L_j\}_{j \geq 1}$ to $\mathcal{S}$; with the remaining probability $(1-\phi)$, we add an independent random strand drawn from the distribution B to $\mathcal{S}$. By construction the strands in $\mathcal{S}$ are i.i.d. Moreover, by construction, each strand is distributed according to $\phi A + (1-\phi)\, B = D_{1/4}$, where the equality follows from the definition of B. So $\mathcal{S}$ has the same distribution as $\mathcal{S}_{D_{1/4}, M}$, which shows part (b) of the claim.

Finally, we verify part (c) of the claim. The intersection of the sets $\mathcal{S}$ and $\mathcal{S}'$ is a prefix of the sequence $L_1, L_2, \ldots$. In expectation, the set $\mathcal{S}$ contains $\phi M = \sqrt{M}$ elements from the sequence $L_1, L_2, \ldots$. Thus, by the Chernoff bound, it contains the first $\sqrt{M}/2$ elements of this sequence with probability at least $1 - \exp(-c_1\sqrt{M})$. Similarly, in expectation, the set $\mathcal{S}'$ contains at least $\sqrt{M}/2$ elements from $L_1, L_2, \ldots$. Hence, by the Chernoff bound, it contains the first $\sqrt{M}/3$ elements from the sequence $L_1, L_2, \ldots$ with probability at least $1 - \exp(-c_2\sqrt{M})$. Thus, by a union bound,

$$\mathbb{P}\left\{\left\{L_1, L_2, \ldots, L_{\sqrt{M}/3}\right\} \subseteq \mathcal{S} \cap \mathcal{S}'\right\} \geq 1 - \exp(-c_1\sqrt{M}) - \exp(-c_2\sqrt{M}).$$

□

Appendix A
Standard Tail Bounds

We recall Hoeffding's inequality (see, e.g., [54]), which we use throughout our proofs.

Theorem 1.1 (Hoeffding's Inequality): Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ almost surely for all $i \leq n$. Let $S_n := X_1 + \ldots + X_n$. Then for every $t > 0$ we have that

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

We also recall the Paley-Zygmund inequality (see, e.g., [54]).

Lemma 1.2 (Paley-Zygmund Inequality): Let Z be a nonnegative random variable and let $\theta \in [0,1]$. Then

$$\mathbb{P}\{Z \geq \theta \mathbb{E}[Z]\} \geq (1-\theta)^2 \frac{(\mathbb{E}[Z])^2}{\mathbb{E}[Z^2]}.$$

We use the Paley-Zygmund inequality to prove a lower bound on the right tail of the sum of i.i.d. random variables that occur in the proofs of our lower bound results.

Lemma 1.3: Let $\ell \geq 2$ be a fixed positive integer. Let $\{Y_i\}_{i=1}^n$ be i.i.d. random variables that are uniform on $\{1, 2, \ldots, \ell\}$. For λ satisfying $0 < \lambda \leq \ell n/50$ we have that

$$\mathbb{P}\left\{\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \geq \lambda\right\} \geq \left(\exp\left(\frac{12.5\lambda^2}{\ell^2 n}\right) - 1\right)^2 \exp\left(-\frac{400\lambda^2}{\ell^2 n}\right). \quad (37)$$

In particular, when $\ell\sqrt{n}/3 \leq \lambda \leq \ell n/50$, we have that

$$\mathbb{P}\left\{\sum_{i=1}^n (Y_i - \mathbb{E}[Y_i]) \geq \lambda\right\} \geq \exp\left(-\frac{400\lambda^2}{\ell^2 n}\right). \quad (38)$$

We note that this bound is sharp up to a universal multiplicative constant in the exponent, as witnessed by Hoeffding's inequality (Theorem 1.1).

Proof: To abbreviate notation, let $Z_i := Y_i - \mathbb{E}[Y_i]$ for $i \in [n]$, and let $S_n := \sum_{i=1}^n Z_i$. For every $t > 0$ we have that

$$\mathbb{P}\{S_n \geq \lambda\} = \mathbb{P}\{\exp(tS_n) \geq \exp(t\lambda)\} = \mathbb{P}\left\{\exp(tS_n) \geq \frac{\exp(t\lambda)}{\mathbb{E}[\exp(tS_n)]} \mathbb{E}[\exp(tS_n)]\right\}. \quad (39)$$

We now apply the Paley-Zygmund inequality (Lemma 1.2) with $Z := \exp(tS_n)$ and

$$\theta := \frac{\exp(t\lambda)}{\mathbb{E}[\exp(tS_n)]}.$$

In order to do so, we must have $\theta \in [0,1]$, so λ must satisfy

$$\lambda \le \frac{1}{t}\log \mathbb{E}[\exp(tS_n)]. \tag{40}$$

For $t > 0$ and λ satisfying (40) we have by (39) and Lemma 1.2 that

$$\begin{aligned}\mathbb{P}\{S_n \ge \lambda\} &\ge \left(1 - \frac{\exp(t\lambda)}{\mathbb{E}[\exp(tS_n)]}\right)^2 \frac{(\mathbb{E}[\exp(tS_n)])^2}{\mathbb{E}[\exp(2\ tS_n)]} \\ &= \frac{(\mathbb{E}[\exp(tS_n)] - \exp(t\lambda))^2}{\mathbb{E}[\exp(2\ tS_n)]}. \end{aligned}\tag{41}$$

In the following we analyze this expression. By an explicit calculation we have that

$$\mathbb{E}[\exp(tY_1)] = \frac{e^t\left(e^{t\ell}-1\right)}{\ell\left(e^t-1\right)}.$$

Since $\mathbb{E}[Y_1] = (\ell+1)/2$, we thus have that

$$\begin{aligned}\mathbb{E}[\exp(tZ_1)] &= \mathbb{E}[\exp(tY_1)]\, e^{-t(\ell+1)/2} \\ &= \frac{e^{t/2}\left(e^{t\ell/2} - e^{-t\ell/2}\right)}{\ell\left(e^t-1\right)}.\end{aligned}$$

By multiplying and dividing by t, we slightly rewrite this expression as

$$\mathbb{E}[\exp(tZ_1)] = \frac{te^{t/2}}{e^t-1}\cdot\frac{e^{t\ell/2}-e^{-t\ell/2}}{t\ell} \tag{42}$$

and bound the two factors separately. For the first factor we have for all $t > 0$ that

$$e^{-t^2/20} \le \frac{te^{t/2}}{e^t-1} \le 1. \tag{43}$$

For the second factor, note that

$$e^{x^2/25} \le \frac{e^{x/2}-e^{-x/2}}{x} \le e^{x^2/20},$$

where the first inequality holds for all $x \in (0,1]$ and the second inequality holds for all $x > 0$. Thus for all $t \in (0, 1/\ell]$ we have that

$$e^{t^2\ell^2/25} \le \frac{e^{t\ell/2}-e^{-t\ell/2}}{t\ell} \le e^{t^2\ell^2/20}. \tag{44}$$

Putting together the bounds in (43) and (44), and plugging them into (42) we obtain the upper bound

$$\mathbb{E}[\exp(tZ_1)] \le e^{t^2\ell^2/20},$$

which holds for all $t > 0$. Similarly, we obtain the lower bound

$$\begin{aligned}\mathbb{E}[\exp(tZ_1)] &\ge \exp\left(t^2\ell^2/25 - t^2/20\right) \\ &= \exp\left(t^2\frac{4\ell^2-5}{100}\right) \ge e^{t^2\ell^2/40},\end{aligned}$$

which holds for all $t \in (0, 1/\ell]$; here in the last inequality we used that $4\ell^2 - 5 \ge 2.5\ell^2$, which holds for all $\ell \ge 2$. In summary, we have shown that for all $t \in (0, 1/\ell]$ we have that

$$e^{t^2\ell^2/40} \le \mathbb{E}[\exp(tZ_1)] \le e^{t^2\ell^2/20}. \tag{45}$$

Since $\{Z_i\}_{i=1}^n$ are i.i.d. we have that $\mathbb{E}[\exp(tS_n)] = (\mathbb{E}[\exp(tZ_1)])^n$. Therefore by (45) we have for all $t \in (0, 1/\ell]$ that

$$e^{t^2\ell^2 n/40} \le \mathbb{E}[\exp(tS_n)] \le e^{t^2\ell^2 n/20}. \tag{46}$$

Plugging these inequalities back into (41) we obtain the lower bound

$$\mathbb{P}\{S_n \ge \lambda\} \ge \frac{\left(\exp\left(t^2\ell^2 n/40\right) - \exp(t\lambda)\right)^2}{\exp(t^2\ell^2 n/5)}. \tag{47}$$

The inequality in (47) holds whenever $t \in (0, 1/\ell]$ and the inequality $t\lambda \le t^2\ell^2 n/40$ holds. We now choose t to be

$$t := \frac{50\lambda}{\ell^2 n},$$

a choice which requires $\lambda > 0$. This choice of t satisfies the inequality $t\lambda \le t^2\ell^2 n/40$. Furthermore, we then have $t \le 1/\ell$ if and only if $\lambda \le \ell n/50$. Plugging in this choice of t into (47) we obtain that

$$\mathbb{P}\{S_n \ge \lambda\} \ge \frac{\left(\exp\left(\frac{62.5\lambda^2}{\ell^2 n}\right) - \exp\left(\frac{50\lambda^2}{\ell^2 n}\right)\right)^2}{\exp\left(\frac{500\lambda^2}{\ell^2 n}\right)},$$

from which (37) follows directly. Finally, note that when $\lambda \ge \ell\sqrt{n}/3$, then $\exp\left(12.5\lambda^2/\left(\ell^2 n\right)\right) - 1 \ge e - 1 > 1$, and (38) follows. □

APPENDIX B
PROOF OF LEMMA 4.1

In this section we prove Lemma 4.1 about the empirical quantiles of the distribution of the random variable $\mathsf{cost}_R(S)$, where S is a random strand.

Proof of Lemma 4.1: By the Dvoretzky–Kiefer–Wolfowitz inequality ([55], [56]), the empirical cumulative distribution function for $\mathsf{cost}_R(S)$ is very close to the cumulative distribution function for $\mathsf{cost}_R(S) \le t$ with high probability over the random choice of the set $\mathcal{S}_{D,M}$. Specifically, we have

$$\mathbb{P}\left\{\left|\frac{|\{S \in \mathcal{S} : \mathsf{cost}_R(S) \le t\}|}{|\mathcal{S}|} - \mathbb{P}\{\mathsf{cost}_R(S) \le t\}\right| \le \varepsilon \right. \\ \left. \text{for all } t \in \mathbb{R}\right\} \ge 1 - 2e^{-2M\varepsilon^2}.$$

Let

$$\mathcal{E} := \left\{\left|\frac{|\{S \in \mathcal{S} : \mathsf{cost}_R(S) \le t\}|}{|\mathcal{S}|} - \mathbb{P}\{\mathsf{cost}_R(S) \le t\}\right| \le \varepsilon \right. \\ \left. \text{for all } t \in \mathbb{R}\right\}.$$

We show that if $\mathcal{E}$ occurs then for all $q \in (\varepsilon, 1-\varepsilon)$, we have

$$Q_{q-\varepsilon,R}(D) \le \widetilde{Q}_{q,R}(\mathcal{S}_{D,M}) \le Q_{q+\varepsilon,R}(D),$$

and, therefore, (6) holds. Inequality (7) follows from (6) by a union bound over all R in $\mathcal{R}$.

We first show that $\widetilde{Q}_{q,R}(\mathcal{S}_{D,M}) \geq Q_{q-\varepsilon,R}(D)$ for all $q \in (\varepsilon, 1)$. Consider an arbitrary $q \in (\varepsilon, 1)$ and let $t^* = Q_{q-\varepsilon,R}(D) - 1$. Since $t^* < Q_{q-\varepsilon,R}(D)$, we have $\mathbb{P}\{\mathsf{cost}_R(S) \leq t^*\} < q - \varepsilon$ and

$$\frac{|\{S \in \mathcal{S} : \mathsf{cost}_R(S) \leq t^*\}|}{|\mathcal{S}|} \leq \mathbb{P}\{\mathsf{cost}_R(S) \leq t^*\} + \varepsilon < q.$$

Hence, $\widetilde{Q}_{q,R}(\mathcal{S}_{D,M}) > t^*$, and since $\widetilde{Q}_{q,R}(\mathcal{S}_{D,M})$ is an integer, $\widetilde{Q}_{q,R}(\mathcal{S}_{D,M}) \geq Q_{q-\varepsilon,R}(D)$.

We now show that $\widetilde{Q}_{q,R}(\mathcal{S}_{D,M}) \leq Q_{q+\varepsilon,R}(D)$ for all $q \in (0, 1-\varepsilon)$. Consider an arbitrary $q \in (0, 1-\varepsilon)$ and let $t^{**} = Q_{q+\varepsilon,R}(D)$. By the definition of $Q_{q,R}(D)$, we have $\mathbb{P}\{\mathsf{cost}_R(S) \leq t^{**}\} \geq q + \varepsilon$ and

$$\frac{|\{S \in \mathcal{S} : \mathsf{cost}_R(S) \leq t^{**}\}|}{|\mathcal{S}|} \geq \mathbb{P}\{\mathsf{cost}_R(S) \leq t^{**}\} - \varepsilon \geq q.$$

Hence, $\widetilde{Q}_{q,R}(\mathcal{S}_{D,M}) \leq t^{**} = Q_{q+\varepsilon,R}(D)$. □

ACKNOWLEDGMENT

The authors would like to thank Karin Strauss for helpful discussions about this work.

REFERENCES

[1] M. S. Neiman, "On the molecular memory systems and the directed mutations," *Radiotekhnika*, vol. 1, no. 6, pp. 1–8, 1965.

[2] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, p. 1628, 2012.

[3] N. Goldman *et al.*, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, pp. 77–80, Jan. 2013.

[4] S. H. T. Yazdi, H. M. Kiah, E. Garcia-Ruiz, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based storage: Trends and methods," *IEEE Trans. Mol. Biol. Multi-Scale Commun.*, vol. 1, no. 3, pp. 230–248, Sep. 2015.

[5] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," in *Proc. 21st Int. Conf. Archit. Support Program. Lang. Operating Syst.*, Mar. 2016, pp. 637–649.

[6] S. L. Shipman, J. Nivala, J. D. Macklis, and G. M. Church, "CRISPR–cas encoding of a digital movie into the genomes of a population of living bacteria," *Nature*, vol. 547, no. 7663, pp. 345–349, 2017.

[7] S. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Sci. Rep.*, vol. 7, p. 5011, Jul. 2017.

[8] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.

[9] L. Organick *et al.*, "Random access in large-scale DNA data storage," *Nature Biotechnol.*, vol. 36, no. 3, pp. 242–248, 2018.

[10] L. Ceze, J. Nivala, and K. Strauss, "Molecular digital data storage using DNA," *Nature Rev. Genet.*, vol. 20, no. 8, pp. 456–466, Aug. 2019.

[11] L. C. Meiser *et al.*, "Reading and writing digital data in DNA," *Nature Protocols*, vol. 15, no. 1, pp. 86–101, 2020.

[12] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis: Technologies and applications," *Nature Methods*, vol. 11, no. 5, pp. 499–507, May 2014.

[13] H. Lee *et al.*, "Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage," *Nature Commun.*, vol. 11, no. 1, pp. 1–9, Dec. 2020.

[14] S. Chandak *et al.*, "Overcoming high nanopore basecaller error rates for DNA storage via basecaller-decoder integration and convolutional codes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8822–8826.

[15] M. L. Bochman, K. Paeschke, and V. A. Zakian, "DNA secondary structures: Stability and function of G-quadruplex structures," *Nature Rev. Genet.*, vol. 13, no. 11, pp. 770–780, Nov. 2012.

[16] C. Rashtchian *et al.*, "Clustering billions of reads for DNA data storage," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3360–3371.

[17] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proc. 15th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2004, pp. 910–918.

[18] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results," in *Proc. 19th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2008, pp. 389–398. [Online]. Available: http://dl.acm.org/citation.cfm?id=1347082.1347125

[19] K. Viswanathan and R. Swaminathan, "Improved string reconstruction over insertion-deletion channels," in *Proc. 19th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2008, pp. 399–408.

[20] Y. Peres and A. Zhai, "Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice," in *Proc. IEEE 58th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2017, pp. 228–239.

[21] N. Holden, R. Pemantle, and Y. Peres, "Subpolynomial trace reconstruction for random strings and arbitrary deletion probability," in *Proc. Conf. Learn. Theory (COLT)*, 2018, pp. 1799–1840.

[22] S. Hannenhalli, E. Hubbell, R. Lipshutz, and P. A. Pevzner, "Combinatorial algorithms for design of DNA arrays," in *Chip Technology* (Advances in Biochemical Engineering/Biotechnology), vol. 77. Berlin, Germany: Springer, 2002, pp. 1–19.

[23] A. B. Kahng, I. Măndoiu, P. A. Pevzner, S. Reda, and A. Zelikovsky, "Border length minimization in DNA array design," in *Proc. Int. Workshop Algorithms Bioinformat.* Berlin, Germany: Springer, 2002, pp. 435–448.

[24] S. Rahmann, "The shortest common supersequence problem in a microarray production setting," *Bioinformatics*, vol. 19, pp. 156–161, Sep. 2003.

[25] A. B. Kahng, I. I. Măndoiu, P. A. Pevzner, S. Reda, and A. Z. Zelikovsky, "Scalable heuristics for design of DNA probe arrays," *J. Comput. Biol.*, vol. 11, nos. 2–3, pp. 429–447, Mar. 2004.

[26] K. Ning and H. W. Leong, "The distribution and deposition algorithm for multiple oligo nucleotide arrays," *Genome Informat.*, vol. 17, no. 2, pp. 89–99, 2006.

[27] S. Rahmann, "Subsequence combinatorics and applications to microarray production, DNA sequencing and chaining algorithms," in *Proc. Annu. Symp. Combinat. Pattern Matching*. Berlin, Germany: Springer, 2006, pp. 153–164.

[28] A. Kumar, M. Cho, and D. Z. Pan, "DNA microarray placement for improved performance and reliability," in *Proc. Int. Symp. VLSI Design, Automat. Test*, Apr. 2010, pp. 275–278.

[29] D. Trinca and S. Rajasekaran, "Fast local-search-based parallel algorithms for DNA probe placement on small oligonucleotide arrays," *Adv. Model. Optim.*, vol. 12, no. 1, pp. 45–55, 2010.

[30] K. Ning and H. W. Leong, "The multiple sequence sets: Problem and heuristic algorithms," *J. Combinat. Optim.*, vol. 22, no. 4, pp. 778–796, Nov. 2011.

[31] S. Srinivasan, V. Kamakoti, and A. Bhattacharya, "A review of algorithms for border length minimization problem," *IETE Tech. Rev.*, vol. 31, no. 5, pp. 369–382, Sep. 2014.

[32] E. Hubbell and P. A. Pevzner, "Fidelity probes for DNA arrays," in *Proc. 7th Int. Conf. Intell. Syst. Mol. Biol.*, 1999, pp. 113–117.

[33] C. J. Colbourn, A. C. H. Ling, and M. Tompa, "Construction of optimal quality control for oligo arrays," *Bioinformatics*, vol. 18, no. 4, pp. 529–535, Apr. 2002.

[34] R. Sengupta and M. Tompa, "Quality control in manufacturing oligo arrays: A combinatorial design approach," *J. Comput. Biol.*, vol. 9, no. 1, pp. 1–22, Jan. 2002.

[35] O. Milenkovic, "Error and quality control coding for DNA microarrays," in *Proc. Inaugural Workshop Center Inf. Theory Appl.*, San Diego, CA, USA, 2006, pp. 1–19.

[36] E.-J. Räihä and E. Ukkonen, "The shortest common supersequence problem over binary alphabet is NP-complete," *Theor. Comput. Sci.*, vol. 16, no. 2, pp. 187–198, 1981.

[37] T. Jiang and M. Li, "On the approximation of shortest common supersequences and longest common subsequences," *SIAM J. Comput.*, vol. 24, no. 5, pp. 1122–1139, 1995.

[38] A. Lenz, Y. Liu, C. Rashtchian, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding for efficient DNA synthesis," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2885–2890.

[39] L. Anavy, I. Vaknin, O. Atar, R. Amit, and Z. Yakhini, "Data storage in DNA with fewer synthesis cycles using composite DNA letters," *Nature Biotechnol.*, vol. 37, no. 10, pp. 1229–1236, Oct. 2019.

[40] H. H. Lee, R. Kalhor, N. Goela, J. Bolot, and G. M. Church, "Terminator-free template-independent enzymatic DNA synthesis for digital information storage," *Nature Commun.*, vol. 10, no. 1, pp. 1–12, Dec. 2019.

[41] S. Jain, F. Farnoud, M. Schwartz, and J. Bruck, "Coding for optimized writing rate in DNA storage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 711–716.
[42] A. Flaxman, A. W. Harrow, and G. B. Sorkin, "Strings with maximally many distinct subsequences and substrings," *Electron. J. Combinatorics*, vol. 11, no. 1, p. R8, Jan. 2004.
[43] D. S. Hirschberg and M. Regnier, "Tight bounds on the number of string subsequences," *J. Discrete Algorithms*, vol. 1, no. 1, pp. 123–132, 2000.
[44] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys. Doklady*, vol. 28, no. 8, pp. 707–710, 1966.
[45] R. Pemantle and M. C. Wilson, "Asymptotics of multivariate sequences II: Multiple points of the singular variety," *Combinatorics, Probab. Comput.*, vol. 13, nos. 4–5, pp. 735–761, Jul. 2004.
[46] R. Pemantle and M. C. Wilson, "Twenty combinatorial examples of asymptotics derived from multivariate generating functions," *SIAM Rev.*, vol. 50, no. 2, pp. 199–272, Jan. 2008.
[47] B. Bukh and C. Cox, "Periodic words, common subsequences and frogs," *Ann. Appl. Probab.*, vol. 32, no. 2, pp. 1295–1332, Apr. 2022.
[48] V. Chvátal and D. Sankoff, "Longest common subsequences of two random sequences," *J. Appl. Probab.*, vol. 12, no. 2, pp. 306–315, Jun. 1975.
[49] V. Dančík and M. Paterson, "Upper bounds for the expected length of a longest common subsequence of two binary sequences," *Random Struct. Algorithms*, vol. 6, no. 4, pp. 449–458, Jul. 1995.
[50] C. Houdré and H. Matzinger, "Closeness to the diagonal for longest common subsequences in random words," *Electron. Commun. Probab.*, vol. 21, pp. 1–19, Jan. 2016.
[51] M. Kiwi, M. Loebl, and J. Matoušek, "Expected length of the longest common subsequence for large alphabets," *Adv. Math.*, vol. 197, no. 2, pp. 480–498, Nov. 2005.
[52] G. S. Lueker, "Improved bounds on the average length of longest common subsequences," *J. ACM*, vol. 56, no. 3, pp. 1–38, May 2009.
[53] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, 2001.
[54] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford Univ. Press, 2013.
[55] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann. Math. Statist.*, vol. 27, no. 3, pp. 642–669, Sep. 1956.
[56] P. Massart, "The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality," *Ann. Probab.*, vol. 18, no. 3, pp. 1269–1283, Jul. 1990.

Konstantin Makarychev is a Professor of Computer Science at Northwestern University. He is interested in designing efficient algorithms for computationally hard problems. Before joining Northwestern, he was a Researcher at Microsoft and IBM Research Labs. He graduated from Princeton University with the Ph.D. degree in computer science in 2007.

Miklós Z. Rácz is an Assistant Professor at Princeton University in the ORFE Department, as well as an Associated Faculty Member at the Center for Statistics and Machine Learning (CSML). Before coming to Princeton, he received the Ph.D. degree in statistics from UC Berkeley and was then a Postdoctoral Researcher in the Theory Group at Microsoft Research, Redmond. His research interests lie broadly at the interface of probability, statistics, computer science, and information theory. His research and teaching has been recognized by Princeton's Howard B. Wentz, Jr. Junior Faculty Award, a Princeton SEAS Innovation Award, and an Excellence in Teaching Award.

Cyrus Rashtchian is currently a Senior Research Scientist at Google. Prior to this, he was a Postdoctoral Researcher in the Computer Science & Engineering Department at the University of California, San Diego. He received the Ph.D. degree in computer science & engineering in 2018, advised by Paul Beame, from the University of Washington, Seattle, WA, USA, and the B.S. degree in computer science from the University of Illinois, Urbana–Champaign.

He has broad research interests in the foundations of data science, including robust and explainable machine learning, DNA data storage, statistical reconstruction, clustering, and distributed algorithms. He has published in top machine learning and theoretical computer science conferences, including SODA, COLT, ITCS, ICML, NeurIPS, and AISTATS, and journals such as *Nature Biotechnology* and the *Annals of Applied Probability*.

Sergey Yekhanin received the Specialist Diploma degree from Moscow State University in 2002, and the Ph.D. degree from MIT in 2007. In 2007–2008 he was a member of the School of Mathematics at the Institute for Advanced Study at Princeton. In 2008, he joined Microsoft Research where he is currently the Manager of Redmond Lab's Algorithms Group. His research interests are in differential privacy, coding for distributed data storage, and coding for DNA data storage. He was a recipient of the ACM Doctoral Dissertation Award (2007) and the IEEE Communications Society and Information Theory Society Joint Paper Award (2014). He has been an invited speaker at the 2014 International Congress of Mathematicians.