Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems

Daniella Bar-Lev[1], **Omer Sabary**[2], **Ryan Gabrys**[1], and **Eitan Yaakobi**[2]
[1]Electrical and Computer Engineering Dept., University of California, San Diego, La Jolla, CA 92093, USA.
[2]The Henry and Marilyn Taub Faculty of Computer Science, Technion, Haifa, 3200003, Israel.

***Abstract*—Although the expenses associated with DNA sequencing have been rapidly decreasing, the current cost of sequencing information stands at roughly $120/GB, which is dramatically more expensive than reading from existing archival storage solutions today. In this work, we aim to reduce not only the cost but also the latency of DNA storage by introducing the *DNA coverage depth problem*, which aims to reduce the required number of reads to retrieve information from the storage system. Under this framework, our main goal is to understand the effect of error-correcting codes and retrieval algorithms on the required sequencing coverage depth. We establish that the expected number of reads that are required for information retrieval is minimized when the channel follows a uniform distribution. We also derive upper and lower bounds on the probability distribution of this number of required reads and provide a comprehensive upper and lower bound on its expected value. We further prove that for a noiseless channel and uniform distribution, MDS codes are optimal in terms of minimizing the expected number of reads. Additionally, we study the DNA coverage depth problem under the random-access setup, in which the user aims to retrieve just a specific information unit from the entire DNA storage system. We prove that the expected retrieval time is at least k for $[n, k]$ MDS codes as well as for other families of codes. Furthermore, we present explicit code constructions that achieve expected retrieval times below k and evaluate their performance through analytical methods and simulations. Lastly, we provide lower bounds on the maximum expected retrieval time. Our findings offer valuable insights for reducing the cost and latency of DNA storage.**

I. Introduction

The rapid expansion of digital data, predicted to reach 180 zettabytes by the end of the year, presents significant challenges for existing storage technologies, whose capacity growth falls short of demand. This gap underscores the need for alternative storage solutions, and DNA has emerged as a compelling candidate due to its unparalleled density, durability, and longevity. In DNA storage, information is encoded into synthetic DNA strands, stored in a physical medium, and retrieved through DNA sequencing. However, the high costs and low throughput of current sequencing technologies hinder the widespread adoption of DNA storage systems.

The DNA storage pipeline consists of three main components. The first is *DNA synthesis*, which produces artificial DNA strands (also known as oligos). These strands are designed to encode the user's information, typically limited to a length of 300 bases due to current synthesis technologies. The synthesis process introduces some noise, resulting in multiple imperfect copies of each strand. The second component is the *storage container*, typically a small tube that holds the synthesized strands that encode the user's information. The final component is *DNA sequencing*, which retrieves the stored information by converting the DNA strands into digital sequences over the alphabet $\{A, C, G, T\}$. These sequences are noisy copies of the synthesized strands and require decoding to reconstruct the original information. In this pipeline, the retrieval time is defined as the duration between the start of sequencing and the completion of the decoding process.

The DNA sequencing channel models the process of reading stored DNA strands and recovering their digital representation. Each stored strand is read multiple times, producing noisy copies due to errors introduced during synthesis, amplification, and sequencing. The number of reads per strand is governed by a probability distribution influenced by the sequencing technology and amplification biases. This channel accounts for several critical characteristics, including its error rates and the *channel distribution*, which defines the probability that each sequenced read corresponds to a noisy copy of a particular strand. This distribution might be uneven across the strands and can significantly impact decoding reliability. The DNA storage pipeline and the sequencing process are graphically described in Figure 1.

An important metric in DNA storage is the coverage depth, defined as the ratio of the number of sequencing reads to the number of designed strands. High coverage depth translates to higher costs and longer retrieval times, making its reduction a key objective. While various studies have explored aspects of DNA data storage, few have focused on systematically optimizing coverage depth. This gap motivates the introduction of the DNA coverage depth problem, which seeks to minimize the number of reads required for reliable data retrieval.

In this work [1], we examine the interplay between error-correcting codes, retrieval algorithms, and DNA sequencing channel characteristics, such as noise levels, error rates, and the probability distribution of reads across strands. These characteristics dictate the reliability and efficiency of data retrieval, as noise introduces decoding challenges and uneven distributions increase the required number of reads. By optimizing these components, this work aims to reduce sequencing costs, minimize retrieval time, and improve overall system efficiency.

II. Problem Definition

In general, for integers k and n, we assume that the user's data is stored in k *information strands* that are encoded into $n \geqslant k$ *encoded strands* using an (n, k) code $\mathcal{C}$. This work addresses three key problems within the context of DNA storage systems.

- **The MDS Coverage Depth Problem** assumes $\mathcal{C}$ is an $[n, k]$ MDS code, and focuses on determining the expected number of reads required for successful decoding of the user's data and establishing tight bounds on the retrieval probability.
- **The Coding Coverage Depth Problem** examines the optimal pairing of error-correcting codes with the channel

The research was funded by the European Union (ERC, DNAStorage, 865630). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work was also supported in part by NSF Grant CCF2212437.

The first two authors contributed equally to this work.

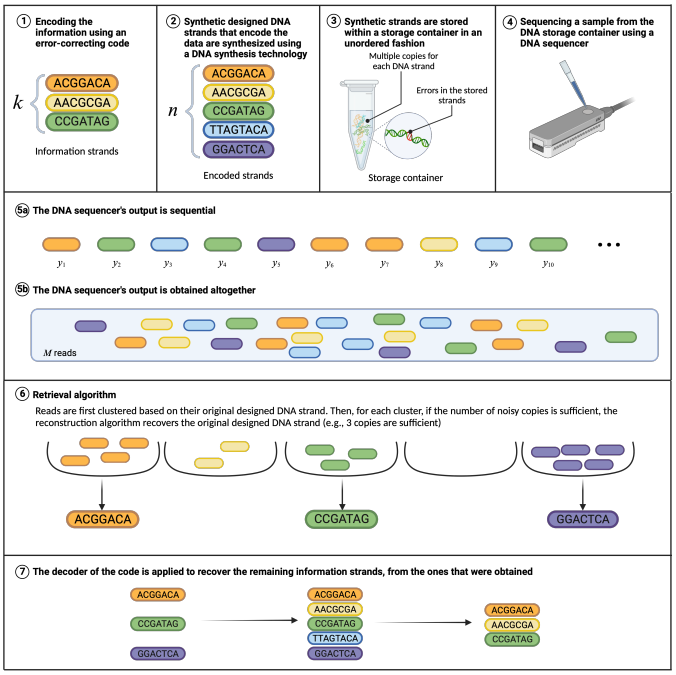

Figure 1. A description of the DNA storage pipeline. Created with BioRender.

to minimize the number of required reads, showing that MDS codes are optimal under certain conditions.

- **The Singleton Coverage Depth Problem** investigates the expected retrieval time for a specific information strand (out of the k strands) in random-access setups.

Together, these problems form a comprehensive framework for optimizing coverage depth and retrieval efficiency in DNA-based storage systems.

III. Related Work

The coverage depth problem draws parallels to classical probabilistic problems like *the coupon collector's problem*, *the dixie cups problem*, and their generalizations [5], [3], [6]. These problems study the number of samples required to collect all (or subset of) items in a set, each at least $t \geqslant 1$ times. This number corresponds to the minimum reads needed to retrieve the information stored in the DNA strands. Extensions of this framework, such as the MDS coverage depth problem, explore how error-correcting codes can further optimize this number.

In the more practical aspect, coverage depth for DNA-based data storage was also studied in the experimental setup, as was done in [4]. In their work, they introduced a Luby transform-based coding scheme, demonstrating the feasibility of encoding digital information into synthetic DNA strands. By diluting synthesized strands, they analyzed the effect of reduced coverage depth, showing successful decoding with an optimized coding strategy despite challenging sequencing conditions. Chandak et al. [2] extended these ideas by formalizing the concepts of writing cost (synthesized bases per information bit) and reading cost (sequenced bases per information bit). Their work highlighted the trade-offs between these metrics, particularly for noiseless channels modeled as erasure channels. For noisy channels, they also proposed LDPC-based coding schemes that improve reading costs through redundancy optimization, validated via simulations.

This work introduces a unified framework to minimize coverage depth under both, full-data and random-access retrieval scenarios, offering novel insights into coding strategies tailored for DNA-based storage.

IV. Main Results

This paper introduces a new framework for analyzing the DNA coverage depth problem and provides significant theoretical and practical insights.

The MDS Coverage Depth Problem

- We establish that the expected number of reads required for retrieval is minimized when the read distribution of the sequencing channel is uniform, meaning that the probability of obtaining a read of a specific strand is equal across all of them. This contrasts with other channel models (see, e.g., [7]).
- Upper and lower bounds on the probability distribution of the required number of reads are derived, offering tight estimation of the expected value of this number.
- Practical bounds are provided for planning read requirements to ensure successful decoding in real-world DNA storage systems.

The Coding Coverage Depth Problem

- For noiseless channels with uniform read distribution, we prove that MDS codes are optimal for minimizing coverage depth.
- We show that for fixed k increasing the number of encoded strands (n) reduces the expected coverage depth.

The Singleton Coverage Depth Problem

- When $n = k$, retrieval without coding is optimal and minimizes the retrieval time.
- We introduce the concepts of *retrieval sets* and *minimal retrieval sets*, which represent the subsets of encoded strands required for successful decoding of an information strand. Using these concepts, we analyze retrieval times and show that standard coding strategies, like simple parity codes and MDS codes, maintain k as the expected time to retrieve an information strands.
- We propose two novel coding schemes that achieve faster retrieval times than traditional approaches (smaller than k). These schemes are validated both analytically and through simulations, showcasing their practical advantages.
- Finally, we establish lower bounds on the retrieval time of a specific strand in the singleton coverage depth problem.

Our contribution in this work has extended the understanding of DNA-based storage systems. Our results can be used for designing cost-effective and efficient DNA storage solutions.

References

[1] D. Bar-Lev, O. Sabary, R. Gabrys, and E. Yaakobi, "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems," *IEEE Tran. on Inf. Theory*, vol. 71, no. 1, pp. 192–218, 2025.

[2] S. Chandak, K. Tatwawadi, B. Lau, J. Mardia, M. Kubit, J. Neu, P. Griffin, M. Wootters, T. Weissman, H. Ji, "Improved read/write cost tradeoff in DNA-based data storage using LDPC codes," *Annual Allerton Conference on Communication, Control, and Computing*, 2019.

[3] P. Erdős, and A. Rényi, "On a classical problem of probability theory," *Magyar Tud. Akad. Mat. Kutató Int.* vol. 6, pp. 215–220, 1961.

[4] Y. Erlich, and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," *Science*, vol. 335, no. 6328, pp. 950–954, 2017.

[5] W. Feller, "An introduction to probability theory and its applications," *Wiley*, vol. 1, 2nd edition, 1967.

[6] P. Flajolet, D. Gardy, and L. Thimonier, "Birthday paradox, coupon collectors, caching algorithms and self-organizing search," *Discrete Applied Mathematics*, vol. 39, no. 3, pp. 207–229, 1992.

[7] M. Gandelman and Y. Cassuto, "Treeplication: An Erasure Code for Distributed Full Recovery Under the Random Multiset Channel," in IEEE Transactions on Information Theory, vol. 67, no. 6, pp. 3542-3556, 2021.