Optimizing Sequencing Coverage Depth in DNA Storage: Insights From DNA Storage Data

Ruiying Cao
Center for Applied Mathematics
Tianjin University
Tianjin
Email: rayen_c@tju.edu.cn

Xin Chen
Center for Applied Mathematics
State Key Laboratory of Synthetic Biology
Tianjin University
Tianjin
Email: chen_xin@tju.edu.cn

***Abstract*—DNA storage is now being considered as a new archival storage method for its durability and high information density, but still facing some challenges like high costs and low throughput. By reducing sequencing sample size for decoding digital data, minimizing DNA coverage depth helps lower both costs and system latency. In this framework, for noiseless channels, we explore the relationship between coverage depth and the MDS code with different redundancy in log-normal distribution channels, a conclusion derived from our PCR and sequencing experimental data analysis. For noisy channels, we study the theoretical lower bounds of sequencing coverage depth required for successful data decoding with high probability, and derive several conclusions that can further guide the efficient implementation of DNA storage experiments.**

I. Introduction

With the rapid growth of global data and advancements in information technology, traditional storage media such as HDDs and SSDs may no longer meet future data storage demands [1], [2]. DNA storage, with its high density and durability, has emerged as a promising solution to address this challenge. However, current DNA storage technologies face significant obstacles, including low throughput and high costs. Reducing the sequencing sample size required to ensure a high probability of decoding all information is the main goal of coverage depth problem, which could provide valuable insights for reducing latency and associated costs.

Several studies have addressed this issue in DNA storage channels follow uniform distribution by adjusting outer error-correcting codes [3]–[7]. However, the processes of DNA synthesis and amplification exhibit a degree of randomness, which results in a non-uniform channel probability distribution. Therefore, in this study, we investigate the problem of minimizing sequencing coverage depth in a real-world channel based on Polymerase Chain Reaction (PCR) and sequencing experimental data, under the non-random access setting.

This paper is organized as follows: Section II offers a detailed description of the problem addressed in this study and then provides an overview of previous work. Section III analyzes PCR and sequencing experimental data to establish the foundation for research on minimizing sequencing coverage depth in channels where the probability distribution follows a log-normal distribution. Section IV investigates the expected value of minimum sequencing coverage depth in the noiseless channel and two lower bounds in the noisy channel, under the non-random access setting.

II. Problem Statement, Related Work

A. Problem Statement

The problem we study is built on the following DNA storage model (see Fig. 1).

Writing Process. The digital information is first converted into a binary bit stream, which is then segmented and translated into m short DNA fragments. These fragments are subsequently encoded with error-correcting codes to generate n DNA strands with redundant information, which are synthesized artificially. After synthesis, the strands undergo PCR amplification and are stored in dry powder form, unordered, in a container.

Reading Process. To read the information, a portion of the dry powder is extracted and sequenced using next-generation sequencing to obtain K reads. After the retrieval process (including clustering, sequence reconstruction), decoding, and the DNA sequence-to-bit steps, the original digital information is recovered.

Notice that due to the inherent randomness in the synthesis, amplification, and sequencing processes of DNA storage [8], it is uncertain whether all the original information can be fully recovered from the K reads obtained through sequencing. Thus, we focus on how to ensure the 100% successful decoding of the original data with high probability, using as small sequencing sample size as possible in the real channel.

We formulate this coverage depth problem under non-random access setting as a variant of the classical coupon collector's problem [9] or the urn model [10]. Specifically, we consider the scenario in which one identical ball is thrown in each round, and the probability of a ball falling into each urn varies. After throwing K rounds, we are interested in the number of urns—among n indistinguishable urns except for their labels—that contain at least a balls. To this end, we model the sample size as a function of the channel probability distribution, the MDS code [11], and the retrieval algorithm. We proceed to give a detailed explanation of three variables.

1) **Channel probability distribution**. The inherent randomness in processes such as synthesis and PCR amplification leads to varying copy numbers of each designed

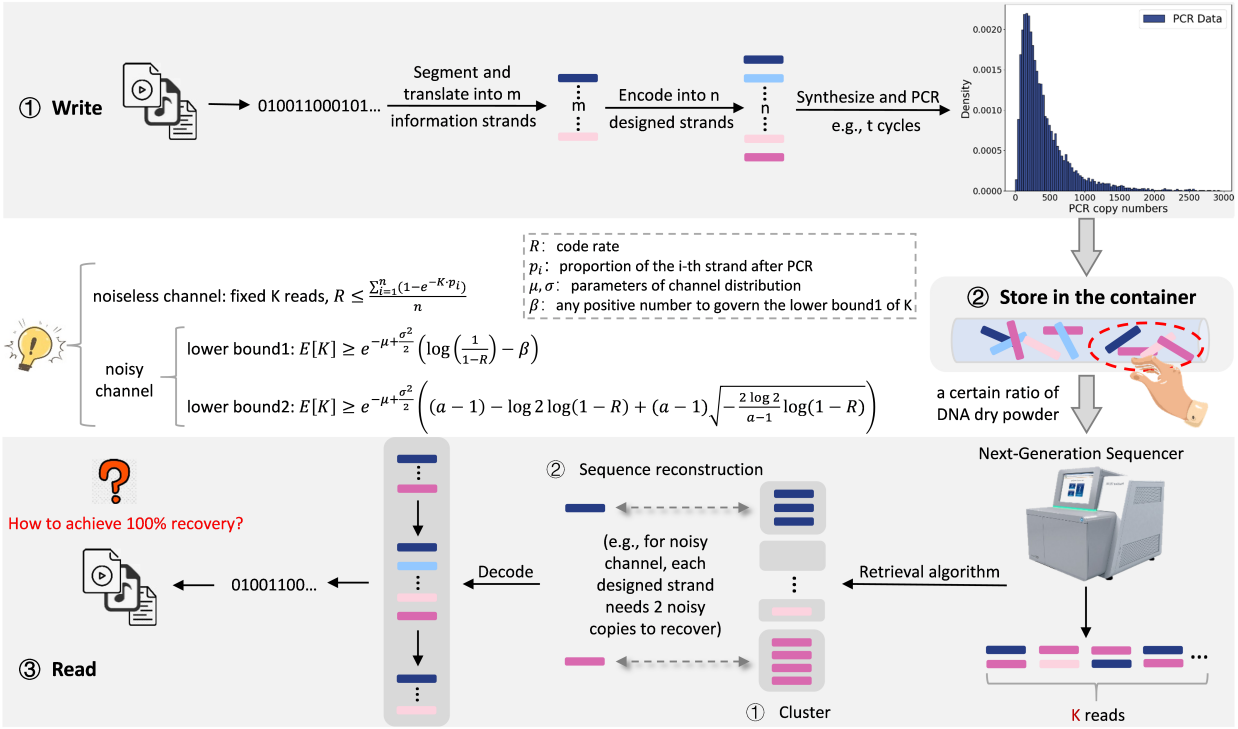


Fig. 1. Framework of this paper.

strand in the sequencing pool, resulting in unequal proportions of each strand in the overall population. We assume that the probability of each designed strand observed in the sequencing reads follows a probability distribution $\boldsymbol{p}_t = \left(p_1^{(t)}, p_2^{(t)}, \ldots, p_n^{(t)}\right)$, where $p_i^{(t)}$ denotes the probability of sampling a read corresponding to the i-th strand after t cycles of PCR amplification. Notice that for simplicity, we consider the distribution $\boldsymbol{p}_t$ solely as a function of the DNA storage channel, without accounting for potential influences from strand design [3]. Accordingly, we refer to $\boldsymbol{p}_t$ as the channel probability distribution in this paper.

2) **The MDS code**. When an $[n, m]$ MDS code is used to encode the information strands, successful retrieval of any m out of the n encoded strands is sufficient to fully recover the original information. We denote the code rate of the $[n, m]$ MDS code by $R = \frac{m}{n}$.
3) **The retrieval algorithm**. The probability of successful retrieval of each designed strand depends on the number of its noisy copies in the sequencing pool, the use of inner codes during encoding, and the channel error rate [3]. In this paper, we model the retrieval algorithm by introducing a positive integer parameter $a \ge 1$, assuming that each designed strand can be successfully retrieved if at least a reads corresponding to it are available.

Based on the assumptions above, we define a new notation $K \triangleq K_a^{\boldsymbol{p}_t}(n, m)$, which means the sample size K to decode the complete information under the condition that after t cycles of PCR, retrieving at least a reads of each of m out of n designed strands. For a uniformly distributed channel, we will omit $\boldsymbol{p}_t$, and denote it as $K_a(n, m)$. The main problems we investigate in this paper are defined below.

Problem 1. (Simulation of DNA storage channel distribution) Given values of $n \ge m \ge 1$, $t \ge 1$, the synthesis amount of each designed strand $c_1, c_2, \ldots, c_n \overset{i.i.d}{\sim} p(c)$, and the amplification efficiency of each designed strand $r_1, r_2, \ldots, r_n$, we focus on the following questions:

1) The probability distribution of $\nu_i^{(t)}$, i.e., copy numbers of the i-th designed strand after t cycles of PCR amplification.
2) The probability distribution of $p_i^{(t)} = \frac{\nu_i^{(t)}}{\sum_{j=1}^{n}\nu_j^{(t)}}$, i.e., the proportion of the i-th designed strand in the population after t cycles of PCR amplification.

Problem 2. (MDS coverage depth problem in the real channel) Given values of $n \ge m \ge 1$, $a \ge 1$, we focus on the following questions:

1) The expectation value $\mathbb{E}[K_a(n, m)]$.
2) Given $t \ge 1$, the expectation value $\mathbb{E}[K_1^{\boldsymbol{p}_t}(n, m)]$.
3) Given $t \ge 1$, the lower bounds of $K_a^{\boldsymbol{p}_t}(n, m)$.

B. Previous Work

By directly mapping the coupon collector's problem, urn and dixie cup problem to the sequencing coverage depth problem, we obtain from [12]–[15] that,

$$\mathbb{E}[K_1(n = m, m)] = m \log m + \gamma m + \mathcal{O}(1),$$

where $\gamma \sim 0.577$ is the Euler–Mascheroni constant.

$$\mathbb{E}[K_1(n, m)] = n(H_n - H_{n-m}),$$

where H_n is the n-th harmonic number.

$$\mathbb{E}[K_a(n=m,m)] = m\log m + m(a-1)\log\log m + mC_a + o(m),$$

where C_a is a constant corresponding to a.

$$\mathbb{E}[K_a^{\boldsymbol{p}}(n,m)] = \sum_{q=0}^{m-1}\int_0^{\infty}[v^q]Q(v)dr, \tag{1}$$

where $\boldsymbol{p}$ represents any general discrete distribution, $[v^q]Q(v)$ represents the coefficients of the q-th terms of the polynomial $Q(v)$, and $Q(v) = \prod_{i=1}^{n}(e_{t-1}(p_i r)+v(e^{p_i r}-e_{t-1}(p_i r)))e^{-r}$, $e_t(x) = \sum_{i=0}^{t}\frac{x^i}{i!}$.

Based on the models above, previous work has been carried out mainly in channels following uniform distribution. For example, [3] proved that for any $\epsilon > 0$,

$$\log\left(\frac{1}{1-R}\right) + f_c(n,R) \le \mathbb{E}\left[\frac{K_a(n,m)}{n}\right] \le K^*,$$

where $K^* = \left(\log\left(\frac{1}{1-R}\right) + a\log\log n + 2\log(a+1)\right)\cdot(1+2\epsilon)$, and $f_c(n,R) = \mathcal{O}(\frac{1}{n^2})$.

Equation (1) does not provide a closed-form expression and is not straightforward to compute. Thus, we will derive the expectation value of $K_a^{\boldsymbol{p}_t}(n,m)$ in a noiseless channel, as well as its lower bounds in a noisy channel building upon the MDS coverage depth problem in [3] in Section IV.

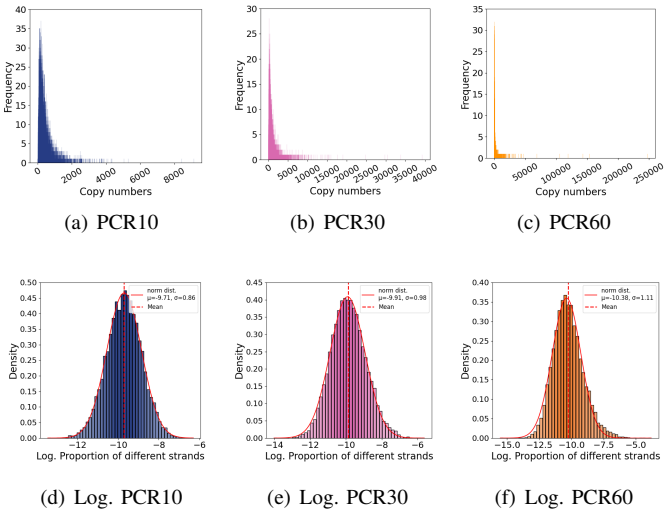
III. Simulation of DNA Storage Channel Probability Distribution

According to the definition of the channel probability distribution in Section II, it is closely related to the proportion of each designed strand in the population after PCR amplification. In this section, by analyzing PCR and sequencing experimental data, we find that the real channel probability distribution follows a log-normal distribution. Based on this, we propose a theoretical model for the simulation of the real channel probability distribution, which serves as the foundation for the subsequent research on minimizing sequencing coverage depth.

A. Analysis of Channel Probability Distribution Based on PCR and Sequencing Experimental Data

We first analyze the real channel probability distribution based on the PCR and sequencing experimental data. We synthesized 11,520 oligos, each 150 nucleotides long, using inkjet printing technology at Twist Bioscience (i.e., $n = 11520$), and PCR amplification is performed using these oligos as template strands. After 10 cycles of amplification, sequencing 4,970,786 reads to generate Dataset PCR10; after 30 cycles, sequencing 11,001,029 reads to generate Dataset PCR30; and after 60 cycles, sequencing 11,180,177 reads to generate Dataset PCR60. The relationship between different PCR cycles and the copy numbers of different oligos at corresponding cycle is shown in Fig. 2(a), 2(b) and 2(c).

Our primary interest lies in the relationship between the number of PCR cycles and the proportion of each strand in the population. Given that the sequencing data exhibit a clear skewed distribution, we assume that the PCR data follow a log-normal distribution [16]. Then we perform a logarithmic transformation on the normalized data before fitting a normal distribution, and Fig. 2(d), 2(e) and 2(f) show the results of a good fit on all three datasets, which means after PCR amplification, the channel actually follows a log-normal distribution.



(a) PCR10 (b) PCR30 (c) PCR60

(d) Log. PCR10 (e) Log. PCR30 (f) Log. PCR60

Fig. 2. Visualization of different cycles of PCR data.

Assuming X is a random variable that follows a normal distribution with parameters μ and σ^2 as its expectation and variance after logarithmic transformation, we denote it as $X \sim \mathcal{LN}(\mu,\sigma^2)$. After fitting the PCR data to a log-normal distribution, the sample distributions of the proportion of the i-th strand in the population after different PCR cycles are presented in Table I.

Since all three datasets are samples from the overall population generated by PCR amplification for 10, 30, and 60 cycles, we will next perform maximum likelihood estimation (MLE) to estimate the parameters of the population. Let $x_1, x_2, \ldots, x_n$ be a simple random sample from the population $X \sim \mathcal{LN}(\mu,\sigma^2)$, the MLE for μ and σ^2 are denoted as $\hat{\mu}_{MLE}$ and $\hat{\sigma}^2_{MLE}$ respectively. Then, from [17], we have

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i), \tag{2}$$

and

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(\ln(x_i) - \hat{\mu}_{MLE})^2. \tag{3}$$

Based on the calculations from (2) and (3), the population distributions of the proportions of different strands after 10, 30, and 60 PCR cycles are presented in Table I.

B. Real Channel Probability Distribution Model

In the previous subsection, we analyzed PCR experimental data to obtain the real channel distributions after 10, 30, and 60 cycles of PCR amplification. However, in practical

TABLE I
THE DISTRIBUTION OF DIFFERENT CYCLES OF PCR DATA.

Cycles	Sample distribution	Population distribution
PCR10	$\mathcal{LN}(-9.71, 0.86^2)$	$\mathcal{LN}(-9.72, 0.74^2)$
PCR30	$\mathcal{LN}(-9.91, 0.98^2)$	$\mathcal{LN}(-9.86, 0.96^2)$
PCR60	$\mathcal{LN}(-10.38, 1.11^2)$	$\mathcal{LN}(-10.25, 1.38^2)$

experiments, additional sequencing of PCR data to obtain the parameter information of the probability distribution is typically not performed. This necessitates the simulation of the channel probability distribution under the condition that only the synthetic amount c_i and amplification efficiency r_i of each designed strand are known, where PCR amplification efficiency is the rate at which the target DNA fragment is amplified during each cycle in the exponential phase. Theoretically, perfect replication would result in 100% efficiency, doubling the DNA amount per cycle. However, in practice, efficiency is typically below 100% and ranges from 80% to 110% due to various factors [22].

Therefore, for given c_i and r_i, where $i \in [1, n]$, we model the channel probability distribution as a function of PCR cycles. First, we show in [21] that after t cycles of PCR amplification, the expected copy number of the i-th strand is $\mathbb{E}\left[\nu_i^{(t)}\right] = c_i(1 + r_i)^t$. Thus, we can derive the proportion of the i-th strand in the population after t cycles of PCR amplification as $p_i^{(t)} = \frac{c_i(1+r_i)^t}{\sum_{j=1}^{n} c_j(1+r_j)^t}$, then the expectation and variance of $p_i^{(t)}$ can be calculated, denoted by $\mathbb{E}\left[p_i^{(t)}\right]$ and $\mathrm{Var}\left[p_i^{(t)}\right]$. From the conclusion we obtain in the previous subsection, we have $p_i^{(t)} \sim \mathcal{LN}\left(\mu^{(t)}, {\sigma^{(t)}}^2\right)$. According to [17], we can calculate that,

$$\mu^{(t)} = \log\left(\mathbb{E}\left[p_i^{(t)}\right]\right) - \frac{1}{2}\log\left(1 + \frac{\mathrm{Var}\left[p_i^{(t)}\right]}{\mathbb{E}^2\left[p_i^{(t)}\right]}\right), \quad (4)$$

and

$$\sigma^{(t)} = \sqrt{\ln\left(1 + \frac{\mathrm{Var}\left[p_i^{(t)}\right]}{\mathbb{E}^2\left[p_i^{(t)}\right]}\right)}. \quad (5)$$

Remark. For the simulation of theorems in the following section, we use the population distribution of PCR amplification data from those three different cycles.

IV. MDS COVERAGE DEPTH PROBLEM IN THE LOG-NORMAL DISTRIBUTION CHANNEL

Under the condition of non-random access, the main goal of our study is to find the minimum sequencing sample size K that ensures the successful decoding of all digital data in the real channel that follows the log-normal distribution after PCR. For given n designed strands, we denote $\alpha \triangleq \frac{K}{n}$ as the sequencing coverage depth.

Due to space limitations, the proofs of all lemmas and theorems presented in this section can be found in [21].

A. MDS Coverage Depth Problem in the Noiseless Channel

In practical experiments, in addition to focusing on the expected sequencing coverage depth required to decode all the digital information (i.e., $\mathbb{E}[\alpha]$), we are also concerned with the probability of successfully decoding all data in a single sequencing run (i.e., $\mathrm{Var}[\alpha]$). Theorem 1 and Theorem 2 respectively describe the probability distribution of the number of designed strands that can be successfully retrieved when sequencing with a fixed sample size of K in practical channels and uniform channels. Let N denote the number of successfully retrieved strands from K fixed sequencing reads.

Remark. In a noiseless channel, $a = 1$.

Theorem 1. For any channel probability distribution $\boldsymbol{p}_t = (p_1^{(t)}, p_2^{(t)}, \ldots, p_n^{(t)})$ and any $K \geq n \geq m \geq 1$, it holds that

$$N \sim \mathcal{N}\Bigg(\sum_{i=1}^{n}\left(1 - e^{-K\cdot p_i^{(t)}}\right), \sum_{i=1}^{n} e^{-K\cdot p_i^{(t)}}\left(1 - e^{-K\cdot p_i^{(t)}}\right) - K\left(\sum_{i=1}^{n} p_i^{(t)} e^{-K\cdot p_i^{(t)}}\right)^2\Bigg),$$

where $X \sim \mathcal{N}(\mu, \sigma^2)$ represents r.v. X follows a normal distribution with mean μ and variance σ^2, and this assumption is carried forward in Theorem 2.

On the basis of Theorem 1, let $\boldsymbol{p}_t = \left(p_1^{(t)}, p_2^{(t)}, \ldots, p_n^{(t)}\right) = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$, the following result is straightforward.

Theorem 2. For a uniform distribution channel, and any $K \geq n \geq m \geq 1$, it holds that

$$N \sim \mathcal{N}\left(n\left(1 - e^{-\alpha}\right), n\left(e^{-\alpha} - e^{-2\alpha} - \alpha e^{-2\alpha}\right)\right).$$

In summary, for n designed strands, when sequencing fixed K reads, the expected ratio of successfully decoding designed strands is $\frac{\mathbb{E}[N]}{n}$. That is, if m information strands are encoded with the $\left[\frac{n\cdot m}{\mathbb{E}[N]}, m\right]$ MDS code, then theoretically speaking, sequencing K reads in a noiseless channel recovers complete information.

Applying the conclusions of Theorem 1 and Theorem 2, Fig. 3 shows a comparison of the minimum sequencing coverage depth between real channels and a uniform distribution channel. Fig. 4 visualizes the variance derivative function to intuitively show the probability of successfully decoding all digital data in a single sequencing experiment, where $f(K) \triangleq e^{-K\mathbb{E}\left[p_i^{(t)}\right]} - e^{-2K\mathbb{E}\left[p_i^{(t)}\right]} - \frac{n}{K}\left(K\mathbb{E}\left[p_i^{(t)}\right]\right)^2 e^{-2K\mathbb{E}\left[p_i^{(t)}\right]}$.

We conclude as follows:

1) When encoding m information strands with identical redundancy, the minimum sequencing coverage depth required to decode all data in a uniform distribution channel is considerably smaller than in a real channel.
2) The lower the code rate, the smaller the minimum sequencing coverage depth in any channels.
3) In a log-normal distribution channel, the minimum sequencing coverage depth increases with the number of PCR cycles. This is attributed to varying amplification efficiencies among the designed strands, leading to the

over-amplification and sequencing of a small subset of strands as the cycle count rises.

4) The variance reaches its maximum between $\alpha = 1$ and $\alpha = 2$ for PCR cycles between 10 and 60, and tends to minimize when $\alpha > 7$ and $\alpha > 8$ respectively, which means that when the expected coverage depth is between 1 and 2, although sequencing αn strands is sufficient to decode all data, the probability of successfully decoding all data in a single experiment is minimized. We recommend increasing the sequencing ratio or the code rate to reduce the variance.

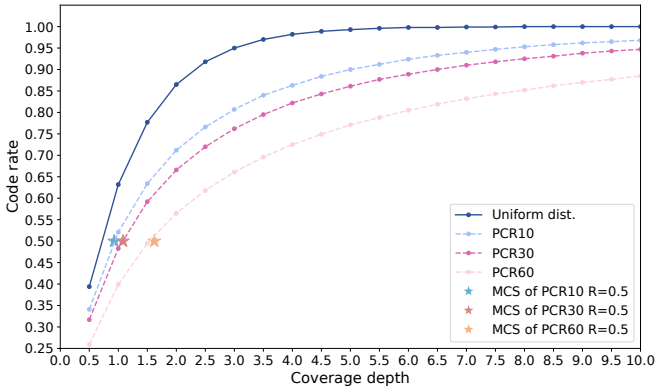


Fig. 3. Expected coverage depth required for successful decoding under different code rates. The *Uniform dist.* curve represents the relationship between coverage depth and code rate in a uniform distribution channel. The *PCR10*, *PCR30*, and *PCR60* curves correspond to the empirical channels derived from Dataset PCR10, Dataset PCR30, and Dataset PCR60, respectively, showing how coverage depth varies with code rate. Three points denote Monte Carlo simulations performed under the respective channel with code rate $R = 0.5$.

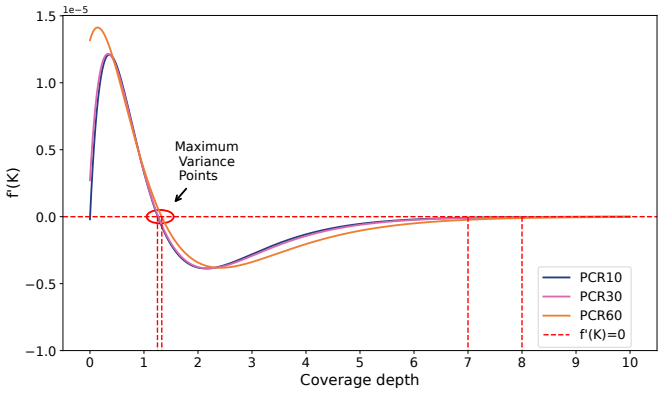


Fig. 4. This plot illustrates the trend of the probability that a single experiment fails to decode all information when sequencing is performed at the expected sample size. This probability is quantified by the variance of each designed strand under that sequencing sample size, denoted as $f(K)$, thus the trend of this probability is the derivative $f'(K)$, which indicates that the variance has only one peak, corresponding to the maximum variance point as annotated in the figure.

B. MDS Coverage Depth Problem in the Noisy Channel

Theorem 3 and Theorem 4 of this section provide two lower bounds on $K_a^{\boldsymbol{p}_t}(n,m)$ in noisy channels obeying a log-normal distribution. According to Section III, if given the synthetic amount and the amplification efficiency of each strand, denoted as $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)$, $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$ respectively, we can deduce the parameters $\mu^{(t)}$ and $\sigma^{(t)}$ of the real log-normal distribution channel after t cycles of PCR.

Remark. In a noisy channel, $a > 1$.

In [21], we show that using $\frac{1}{\mathbb{E}\left[p_i^{(t)}\right]}$ is sufficient for Theorems 3 and 4. To achieve a tighter lower bound, we can replace it with $\mathbb{E}\left[\frac{1}{p_i^{(t)}}\right]$, where $\mathbb{E}\left[p_i^{(t)}\right] = e^{\mu^{(t)}+\frac{\sigma^{(t)2}}{2}}$, $\mathbb{E}\left[\frac{1}{p_i^{(t)}}\right] = e^{-\mu^{(t)}+\frac{\sigma^{(t)2}}{2}}$ [19]. This replacement is justified by Jensen's inequality, since $\frac{1}{x}$ is convex, implying $\mathbb{E}\left[\frac{1}{x}\right] > \frac{1}{\mathbb{E}[x]}$.

Theorem 3. For given $\boldsymbol{c}$, $\boldsymbol{r}$, for any $t \geq 1$, $\beta > 1$, $a > 1$, let $R = \frac{m}{n}$, it holds that $P\left[K_a^{\boldsymbol{p}_t}(n,m) \leq K_1\left(\mu^{(t)}, \sigma^{(t)}, a, R\right)\right] \leq e^{-\beta}\left(1 + \frac{m}{n-m}\right)$, where

$$K_1(\mu^{(t)}, \sigma^{(t)}, a, R) \triangleq e^{-\mu^{(t)}+\frac{\sigma^{(t)2}}{2}}\left(\log\left(\frac{1}{1-R}\right) - \beta\right).$$

Let N_K represent the number of urns that contain less than a balls after K rounds. When using an $[n,m]$ MDS code, recovering all data means successfully decoding m out of n strands, implying that at most $n-m$ strands cannot be successfully decoded. According to Claim 2 in [3], Lemma 1 follows straightforwardly.

Lemma 1. For $K > (a-1)\cdot\mathbb{E}\left[\frac{1}{p_i^{(t)}}\right]$, we have $\mathbb{E}[N_K] \leq n-m$, if $\frac{K\cdot\mathbb{E}\left[p_i^{(t)}\right]}{a-1}e^{-\frac{K\cdot\mathbb{E}\left[p_i^{(t)}\right]}{a-1}} \leq \frac{1}{e}(1-R)^{\frac{\log 2}{a-1}}$.

Based on Lemma 1, the second tighter lower bound is as follows.

Theorem 4. For given $\boldsymbol{c}$, $\boldsymbol{r}$, for any $a > 1$, $t \geq 1$, let $R = \frac{m}{n}$, we have $\mathbb{E}[N_K] \leq n-m$, if $K_a^{\boldsymbol{p}_t}(n,m) > K_2\left(\mu^{(t)}, \sigma^{(t)}, a, R\right)$, where

$$K_2\left(\mu^{(t)}, \sigma^{(t)}, a, R\right) \triangleq e^{-\mu^{(t)}+\frac{\sigma^{(t)2}}{2}}\Bigg((a-1) - \log 2\log(1-R) + (a-1)\sqrt{-\frac{2\log 2}{a-1}\log(1-R)}\Bigg).$$

V. Conclusion

In this paper, by working with PCR and sequencing data, we analyze and simulate the probability distribution of the real channel, and mainly investigate the MDS coverage depth problem based on experimental data, under the non-random access setting. That is, we prove the expected coverage depth and its theoretical lower bounds in real noiseless and noisy channels respectively, with the lower bounds shown in Table II in [21], and first propose the problem of decoding all data successfully in a single sequencing experiment under the expected sample size.

Acknowledgment

This work was supported by the National Key R&D Program of China under Grant No. 2020YFA0712100.

REFERENCES

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

[2] J. Rydning, "Worldwide IDC global datasphere forecast, 2022–2026: Enterprise organizations driving most of the data growth," *International Data Corporation (IDC)*, 2022.

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

[4] H. Abraham, R. Gabrys, and E. Yaakobi, "Covering all bases: The next inning in DNA sequencing efficiency," *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 464-469, 2024.

[5] A. Gruica, D. Bar-Lev, A. Ravagnani, E.Yaakobi, "A combinatorial perspective on random access efficiency for DNA storage," *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 675-680, 2024.

[6] I. Preuss, B. Galili, Z. Yakhini, L. Anavy, "Sequencing coverage analysis for combinatorial DNA-based storage systems," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 10, no. 2, pp. 297 - 316, 2024.

[7] A. Gruica, M. Montanucci, F. Zullo, "The Geometry of Codes for Random Access in DNA Storage," *arXiv preprint arXiv: 2411.08924*, 2024.

[8] L. Organick, S. D. Ang, Y. J. Chen, et al, "Erratum: Random access in large-scale DNA data storage," *Nature Biotechnology*, vol. 36, no.7, pp. 660, 2018.

[9] P. Neal, "The Generalised Coupon Collector Problem," *Journal of Applied Probability*, vol. 45, no. 3, pp. 621-629, 2008.

[10] P. C. Consul, "A Simple Urn Model Dependent upon Predetermined Strategy," *Sankhyā: The Indian Journal of Statistics, Series B*, vol. 36, no. 4, pp. 391-399, 1974.

[11] R. Singleton, "Maximum distance q-nary codes," *IEEE Transactions on Information Theory*, vol. 10, no. 2, pp. 116-118, 1964.

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

[13] 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.

[14] D. J. Newman, "The Double Dixie Cup Problem," *The American Mathematical Monthly*, vol. 67, no. 1, pp. 58-61, 1960.

[15] P. Erdős, and A. Rex nyi, "On a classical problem of probability theory," *Magyar Tud. Akad. Mat. Kutató Int*. vol. 6, no. 1-2, pp. 215–220, 1961.

[16] V. Svensson, K. N. Natarajan, L. Ly, R. J Miragaia, C. Labalette, I. C Macaulay, A. Cvejic, S. A Teichmann, "Power analysis of single-cell RNA-sequencing experiments," *Nature methods*, vol. 14, no. 4, pp. 381-387, 2017.

[17] Ginos, B. Faith, "Parameter estimation for the lognormal distribution," Brigham Young University, 2009.

[18] V. P. Chistyakov, "On the calculation of the power of the test of empty boxes," *Theory of Probability & Its Applications*, vol. 9, no. 4, pp. 648-653, 1964.

[19] J. Aitchison, J. A. C. Brown, "The Lognormal Distribution," *Cambridge: Cambridge University Press*, pp. 8, 1963.

[20] A. N. Philippou, C. Georghiou, G. N. Philippou, "A generalized geometric distribution and some of its properties," *Statistics & Probability Letters*, vol. 1, no. 4, pp. 171–175, 1983.

[21] R. Cao, X. Chen, "Optimizing Sequencing Coverage Depth in DNA Storage: Insights From DNA Storage Data," *arXiv preprint arXiv: 2501.06801*, 2025.

[22] J. M. Ruijter, C. Ramakers, W. M. H. Hoogaars, Y. Karlen, O. Bakker, M. J. B. van den Hoff, A. F. M. Moorman, "Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data," *Nucleic Acids Research*, vol. 37, no. 6, pp. e45, 2009.