Information Density of DNA Storage in the Short Molecule Regime

Ran Tamir
Universitat Politècnica de Catalunya
ran.tamir@upc.edu

Nir Weinberger
Technion – Israel Institute of Technology
nirwein@technion.ac.il

Albert Guillén i Fàbregas
University of Cambridge
Universitat Politècnica de Catalunya
guillen@ieee.org

***Abstract*—We complete the proof of a conjecture put forward by Shomorony and Heckel (2022) regarding the amount of reliable information bits that can be stored in a DNA-based storage system composed by short DNA molecules. While the direct part of the aforementioned conjecture was only partially proved in a recent paper, here we take a more direct approach and complete the proof for the entire range of short molecules. We analyze a random-coding scheme, where each codeword is given by an appropriate quantization of a randomly generated probability mass function from the probability simplex. By analyzing the optimal maximum-likelihood decoder, we derive an achievability bound, which matches a recently proved converse bound.**

I. Introduction

Storing information in DNA molecules has a potential for extremely high information density[1] and longevity, and can address the ever-growing demand for digital storage. Several working prototypes and system proposals [1]–[4] have sparked a surge of information-theoretic and coding-theoretic research, including coding methods [5], channel capacity and error probability analysis [6]–[12], machine-learning based systems [13], [14], secrecy [15], [16], and many more.

In this paper, we consider the commonly adopted DNA storage channel model, known as the *shuffling-sampling channel* [9]. In this channel, information is encoded as a codeword comprised of M molecules, where each molecule has length $L = \beta \log M$ symbols from the alphabet $\mathcal{A}$ (a natural choice is $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, i.e., the four DNA bases, however, other alphabets are also possible), for some length parameter $\beta > 0$. The M molecules are stored in a DNA pool, with no preservation of order. The retrieval of information is performed in two recurring steps. First, one of the M molecules is randomly chosen from the DNA pool, with a uniform distribution, and with replacement. Second, the chosen molecule is sequenced, that is, the sequence of nucleotides it is comprised from is reconstructed to obtain a read. These two steps are repeated for K times, where typically $K > M$. While the sequencing operation is noisy in practice, in this paper we address the stylized model of noiseless sequencing. Even under this assumption, the list of K output reads is still random due to the sampling operation, as some molecules may be sampled more than once, and others may not be sampled at all.

The length of the molecules, as designated by the parameter β, affects the channel capacity of this channel, as both effects – lack of molecule order and non-ideal sampling – are less severe as β increases. In case that the molecule length parameter is large enough, specifically, $\beta > \frac{1}{\log|\mathcal{A}|}$, a simple scheme is shown to achieve the channel capacity [8]: Start each molecule with a header of length $\log_{|\mathcal{A}|} M = \frac{L}{\beta \log|\mathcal{A}|}$ symbols from $\mathcal{A}$,[2] identifying the index of the molecule in $\{1, 2, \ldots, M\}$. Then, the rest of $L(1 - \frac{1}{\beta \log|\mathcal{A}|})$ symbols can be used for encoding the data. The resulting coding rate (or *information density*), i.e., total number of encoded bits divided by the total number of nucleotides used ML, is then given by $(1 - \frac{1}{\beta \log|\mathcal{A}|})$, which, as said, can be proved to be the capacity. This scheme clearly breaks if the molecule length is shorter, i.e. $0 < \beta < \frac{1}{\log|\mathcal{A}|}$, because then each molecule is too short to even just encode its index. Since the decoder can always ignore some nucleotides, it follows that the capacity is monotonically non-decreasing in β, thus the capacity equals *zero* for any $0 < \beta < \frac{1}{\log|\mathcal{A}|}$.

Nonetheless, the regime of $0 < \beta < \frac{1}{\log|\mathcal{A}|}$, which is called the *short molecule regime*, is still of interest. In this regime, the total number of different types of molecules $|\mathcal{A}|^L = |\mathcal{A}|^{\beta \log M} = M^{\beta \log|\mathcal{A}|}$ is smaller than the total number of molecules M. The pigeon hole principle then implies that each codeword must contain repeated molecules. Consequently, the information is encoded into a frequency vector, containing the relative count of each of the $M^{\beta \log|\mathcal{A}|}$ types of molecules in the DNA pool of M molecules. During reading, the sampling operation produces a noisy version of this vector, since, e.g., molecule types appearing once in the codeword may be sampled more than once, or none at all. The decoder then finds the codeword whose frequency vector is closest, in a manner to be made precise, to the frequency vector defined by the output reads.

As mentioned, the channel capacity of the shuffling-sampling channel is zero in the short molecule regime $\beta <$

The research of N. Weinberger was partially supported by the Israel Science Foundation (ISF), grant no. 1782/22. The research of R. Tamir and A. Guillén i Fàbregas was supported in part by the European Research Council under Grants 101142747 and 101158232, and in part by the Spanish Government under Grants PID2020-116683GB-C22 and PID2021-128373OB-I00.

[1]In information theory, the expression "information density" is commonly referred to the random variable whose expectation is the mutual information. In the current context, this expression should be understood as the amount of information bits per gram of DNA.

[2]For simplicity, we ignore here integer constraints.

$\frac{1}{\log|\mathcal{A}|}$. This implies that the total number of reliably stored bits scales at most *sub-linearly* with the total number of nucleotides ML. However, just a few grams of DNA contain a large amount of nucleotides ML, and so for a given M and L, the potential total number of reliably stored bits may be still very large. This motivated an analysis of this regime in [9, Sec. 7.3], leading to a conjecture on the maximal log-cardinality of a reliable codebook as a function of ML. Specifically, [9, Conjecture 4] states that for $\beta \in (0, \frac{1}{\log|\mathcal{A}|})$ this log-cardinality is asymptotically

$$\frac{1-\beta\log|\mathcal{A}|}{2} \cdot M^{\beta\log|\mathcal{A}|}\log M. \tag{1}$$

Evidently, this total number of bits (given the logarithm is in the binary base) is $o(ML)$, but still increases with M. It should be mentioned that in [9, Sec. 7.3], a Poisson sampling model was considered. In this model, the total number of output reads is not fixed in advance, and the random number of times that each type of molecule appears at the output reads is distributed as a Poisson random variable, whose parameter is given by the number of times that this type of molecule appears in the input codeword. This may be compared to the original sampling model, whose sampling operation can be described by a multinomial random variable. The conjecture leading to (1) is then based on relating the frequency-based channel to a power-constrained Poisson channel, for which the asymptotic scaling of its capacity, as a function of the input power, is known [17]. However, (1) remained a conjecture since the Poisson channel obtain from the reduction is non-standard: Its power constraint increases with the blocklength (amounts to $M^{\beta\log|\mathcal{A}|}$), and its inputs are restricted to the *integers*.

In [18] a rigorous approach to this conjecture was made, based on the original multinomial sampling model, rather than the Poisson sampling model. A converse result, based on a Poissonization of the multinomial (e.g., [19, Ch. 5]) shows that the log-cardinality cannot be better than (1), up to an $o(\frac{1}{\log M})$ additive term. On the other hand, an achievability result shows that (1) can be achieved, however, under the additional condition that $\beta \in (\frac{1}{2\log|\mathcal{A}|}, \frac{1}{\log|\mathcal{A}|})$, that is, the molecule is not *very* short. The proof of achievability in [18] is based on Feinstein's maximal coding bound [20] [21, Thm. 20.7], and is rather intricate. It is based on several steps, aiming to rigorously address the reduction of the multinomial channel to a Poisson channel, the integer constraints on its input, and other constraints which stem from these reductions. This resulted the additional condition, which left open the rigorous establishment of the conjecture whenever $\beta \in (0, \frac{1}{2\log|\mathcal{A}|})$.

Our main contribution in this paper is to complete the picture, and rigorously establish [9, Conjecture 4] in the entire short-molecule regime $\beta \in (0, \frac{1}{\log|\mathcal{A}|})$. We achieve this by a direct, and rather different, proof method. We conduct a random coding analysis, in which codewords are drawn by first randomly choosing a point in the probability simplex based on a Dirichlet distribution, and then rounded to integer count vectors. Directly analyzing the average error probability of this ensemble, leads to an achievable bound on the log-cardinality, which matches (1).

Works directly related to this paper, are as follows. In [22], which was also motivated by the short-molecule regime with Poisson sampling, the capacity of Poisson channels with integer (lattice) inputs was considered. In [23], we considered the short-molecule regime, but without a constraint on the input being an integer, and obtained random-coding error bounds on the error probability. This setup is motivated by the fact that the actual sequencing costs are for *different* molecules, since once a molecule is synthesized, the costs of duplicating it are relatively low. Thus, any arbitrary molecule frequency vector can be accurately approximated.

A full version of the paper containing full proofs and all omitted details can be found in [24].

II. Notation Conventions, Settings and Problem Formulation

A. Notation Conventions

For a positive integer n, we use the notation $[n]$ to denote the set $\{1, 2, \ldots, n\}$. The probability of an event $\mathcal{A}$ will be denoted by $\mathbb{P}[\mathcal{A}]$. The expectation of a random variable X will be denoted by $\mathbb{E}[X]$. The indicator function of an event $\mathcal{A}$ will be denoted by $\mathbb{1}[\mathcal{A}]$. The cardinality of a finite set $\mathcal{A}$ will be denoted by $|\mathcal{A}|$. The *floor* function of a real number x, denoted by $\lfloor x \rfloor$, is defined as $\lfloor x \rfloor := \max\{y \in \mathbb{Z} : \ y \leq x\}$. The $(n-1)$-dimensional probability simplex, denoted by $\mathcal{P}_n$, is defined as $\mathcal{P}_n := \{(x_1, \ldots, x_n) \in [0,1]^n : \ \sum_{i=1}^n x_i = 1\}$.

B. Settings and Problem Formulation

Let $\mathcal{C}_M$ be a codebook for data storage in a system that relies on short molecules. Each codeword in $\mathcal{C}_M$ is composed by at most M molecules[3]. More specifically, for any $m \in \{1, 2, \ldots, |\mathcal{C}_M|\}$, the codeword $\boldsymbol{x}(m)$ is given by a set of sequences of the form $(\boldsymbol{x}_1^L(m), \boldsymbol{x}_2^L(m), \ldots, \boldsymbol{x}_{J(m)}^L(m))$, where $J(m) \leq M$ and for every $i \in [J(m)]$, $\boldsymbol{x}_i^L \in \mathcal{A}^L$. In the short-molecule regime, we assume that $L = \beta \log M$ for some $\beta \in (0, \frac{1}{\log|\mathcal{A}|})$, and then $|\mathcal{A}^L| = M^{\beta\log|\mathcal{A}|}$.

We assume that the message m is drawn with a uniform distribution from the set $\{1, 2, \ldots, |\mathcal{C}_M|\}$ and that all the molecules that make up the codeword $\boldsymbol{x}(m)$ are grouped with no preservation of order. When the message is restored, we assume that exactly K sequences $\boldsymbol{y} := (\boldsymbol{y}_1^L, \boldsymbol{y}_2^L, \ldots, \boldsymbol{y}_K^L)$ are independently sampled (with replacement) from the DNA pool. We assume that the *coverage depth* $\xi := \frac{K}{M}$ is fixed.

Based on the sampled sequences, the decoder estimates the message as $\hat{m}(\boldsymbol{y})$. The probability of error is given by

$$\varepsilon_M = \mathbb{P}\left[\hat{m}(\boldsymbol{Y}) \neq m\right], \tag{2}$$

which is taken with respect to the randomness of the message selection, the (possibly) random codebook generation, and the sampling process.

Our main objective in this work is to resolve the direct part of [9, Conjecture 4], which states that for $\mathcal{A} = \{0, 1\}$,

[3]Note that distinct codewords may be of different sizes. However, we assume a uniform upper bound on their sizes because the cost of the input is related to the number of molecules synthesized.

there exists a sequence of codes $\{\mathcal{C}_M\}_{M\geq 1}$ with $\xi = 1$ and a vanishing error probability, for which

$$\limsup_{M\to\infty} \frac{\log|\mathcal{C}_M|}{M^\beta \log M} = \frac{1-\beta}{2}. \tag{3}$$

Although [9, Conjecture 4] was postulated for the special cases $\mathcal{A} = \{0,1\}$ and $\xi = 1$, in this work we address the general case of an arbitrary $\mathcal{A}$ (with $|\mathcal{A}|< \infty$) and $\xi > 0$.

III. Main Result and Discussion

In our model, each codeword is composed of M short DNA molecules, and the system designer is allowed to choose how many copies to take from each possible string in $\mathcal{A}^L$. In other words, each codeword is equivalent to an empirical probability mass function (PMF) over $M^{\beta\log|\mathcal{A}|}$ entries; hence, generating a codebook means choosing many empirical PMFs. While such a codebook may be composed deterministically, it turns out that the random coding methodology commonly used when studying ordinary channel coding scenarios may be adapted quite easily to the case at hand: Instead of drawing vectors from $\mathcal{X}^n$, one can draw PMFs from the probability simplex and quantize the given realizations to attain empirical PMFs.

We now describe more specifically the encoding-decoding algorithm. Let us denote $n := M^{\beta\log|\mathcal{A}|}$ and let $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n$ be an ordered set of all strings in $\mathcal{A}^L$. Each codeword in $\mathcal{C}_M$ is generated in the following procedure. For the message m, a random PMF $\boldsymbol{P}_m = (P_m(1),\ldots,P_m(n))$ is drawn from the $(n-1)$-dimensional simplex $\mathcal{P}_n$ according to the Dirichlet distribution with vector parameters $\boldsymbol{\alpha} := (1,\ldots,1)$, which is equivalent to the uniform measure over $\mathcal{P}_n$.

The m-th codeword is made up of $\lfloor MP_m(\ell)\rfloor$ copies of the string $\boldsymbol{a}_\ell$, where $\ell \in [n]$. The m-th codeword is also represented by the empirical probability vector $\hat{\boldsymbol{P}}_m := (\hat{P}_m(1),\ldots,\hat{P}_m(n))$, where for any $\ell\in[n]$,

$$\hat{P}_m(\ell) = \frac{\lfloor MP_m(\ell)\rfloor}{\sum_{k=1}^n \lfloor MP_m(k)\rfloor}. \tag{4}$$

Given $\boldsymbol{y} := (\boldsymbol{y}_1^L, \boldsymbol{y}_2^L,\ldots,\boldsymbol{y}_K^L)$, the decoder first calculates the frequency vector $\hat{\boldsymbol{Q}} := (\hat{Q}(1),\ldots,\hat{Q}(n))$, where for any $\ell\in[n]$,

$$\hat{Q}(\ell) := \frac{1}{K}\sum_{i=1}^K \mathbb{1}[\boldsymbol{y}_i^L = \boldsymbol{a}_\ell]. \tag{5}$$

One can show (as in [23, Subsection II.B]) that the maximum likelihood decoder is equivalent to a decoder that estimates the message as the one whose codeword minimizes the Kullback–Leibler divergence with $\hat{\boldsymbol{Q}}$. To this end, the decoder estimates the transmitted message according to

$$\hat{m}(\boldsymbol{y}) = \operatorname*{arg\,min}_{m\in[|\mathcal{C}_M|]} D(\hat{\boldsymbol{Q}}\|\hat{\boldsymbol{P}}_m). \tag{6}$$

Theorem 1: Consider an error-free shuffling-sampling channel with $\beta \in (0, \frac{1}{\log|\mathcal{A}|})$ and a coverage depth $\xi > 0$. There exists a sequence of codes $\{\mathcal{C}_M\}_{M\geq 1}$ with vanishing error probabilities ($\varepsilon_M \to 0$), such that

$$\lim_{M\to\infty} \frac{\log|\mathcal{C}_M|}{M^{\beta\log|\mathcal{A}|}\log M} = \frac{1-\beta\log|\mathcal{A}|}{2}. \tag{7}$$

It is interesting to note that although we have considered a general coverage depth $\xi > 0$, the asymptotic log-cardinality of the largest storage codebook is independent of ξ. While the optimal information density is independent of ξ, the error probability converges faster for larger values of ξ, as was also observed in the recent studies [7], [11], [12].

The idea of generating PMF codewords using the Dirichlet distribution has been borrowed from [23], but with one major modification: while the channel model in [23] was assumed to be with infinite input-resolution, this can no longer be assumed in the current work, because each codeword is given by the assignment of a specific number of DNA molecules into all possible molecule types. Hence, the channel in our model is restricted to a finite input resolution. To satisfy this restriction, we generate a quantized version of each PMF codeword as in (4). As a result of this quantization operation, the number of DNA molecules composing each codeword is not fixed and can be easily shown to be in the range $\{M-n,\ldots,M\}$. Since $n = M^{\beta\log|\mathcal{A}|} \ll M$, all the codewords have roughly the same size. While other quantization techniques may be implemented to yield a codebook with a fixed number of DNA molecules in each codeword, we prefer to stick to the specific quantization technique because of a technical reason. At the beginning of the proof of Theorem 1, when handling the pairwise error probability $\mathbb{P}[D(\hat{\boldsymbol{Q}}\|\hat{\boldsymbol{P}}) \leq D(\hat{\boldsymbol{Q}}\|\hat{\boldsymbol{p}})]$, where $\hat{\boldsymbol{p}}$ is the true codeword, $\hat{\boldsymbol{P}}$ is a competing codeword, and $\hat{\boldsymbol{Q}}$ is the empirical distribution of the vector of samples, we upper-bound this probability using Chernoff's inequality, and then, in the next step, we need to handle an expectation of the form

$$\mathbb{E}\left[\prod_{i=1}^n \hat{P}(i)^{\theta\hat{Q}(i)}\right], \tag{8}$$

where $\theta > 0$ is a parameter. While the expectation in (8) is not easy to solve, one can tightly upper-bound it by a constant times the expectation

$$\mathbb{E}\left[\prod_{i=1}^n P(i)^{\theta\hat{Q}(i)}\right], \tag{9}$$

where $\boldsymbol{P} = (P(1),\ldots,P(n))$ is the original, unquantized PMF drawn from the Dirichlet distribution. Unlike products of independent random variables, where their expectations can be calculated by pulling the multiplication operation outside, when dealing with products of dependent random variables, things are usually more complicated. Although the various components of the vector $\boldsymbol{P}$ are statistically dependent, it turns out that the product moment in (9) can be precisely evaluated when $\boldsymbol{P}$ follows a Dirichlet distribution. The following result concerning product moments of the Dirichlet distribution can be found, e.g., in [25, p. 274].

Proposition 2: Let $(\alpha_1,\ldots,\alpha_n)$ and $(\beta_1,\ldots,\beta_n)$ be positive vectors and let $(X_1,\ldots,X_n) \sim \mathrm{Dir}(\alpha_1,\ldots,\alpha_n)$. Then, it holds that

$$\mathbb{E}\left[\prod_{i=1}^n X_i^{\beta_i}\right] = \frac{\Gamma\left(\sum_{i=1}^n \alpha_i\right)}{\Gamma\left(\sum_{i=1}^n(\alpha_i+\beta_i)\right)} \cdot \prod_{i=1}^n \frac{\Gamma\left(\alpha_i+\beta_i\right)}{\Gamma\left(\alpha_i\right)}. \tag{10}$$

As can be easily seen, the expectation in (9) can be evaluated exactly using the result of Proposition 2, and as a consequence, we now see that, at least from a technical point of view, the specific PMF quantization technique as defined in (4) carries a major advantage.

The work in [18] considered a more general frequency-based channel, derived tight lower and upper bounds on its capacity, and then specialized these results to the DNA-based storage channel with short molecules. The achievability bound in [18], when implemented to the DNA storage channel, yields that

$$\frac{\log|\mathcal{C}|}{M^{\beta\log|\mathcal{A}|}L} \geq \frac{1-\beta\log|\mathcal{A}|}{2\beta} - \frac{2.773}{2\beta\log M} + o\left(\frac{1}{\log M}\right), \quad (11)$$

which agrees with Theorem 1 when $M \to \infty$ for an arbitrary $\xi > 0$. It is important to mention here that the bound in (11) holds as long as $\beta \in (\frac{1}{2\log|\mathcal{A}|}, \frac{1}{\log|\mathcal{A}|})$, where the result in Theorem 1 holds for any $\beta \in (0, \frac{1}{\log|\mathcal{A}|})$. Hence, as opposed to the achievability result in [18], the result in the current paper holds also for very short molecules.

Furthermore, the converse bound in [18], when implemented to the DNA storage channel, provides that

$$\frac{\log|\mathcal{C}|}{M^{\beta\log|\mathcal{A}|}L} \leq \frac{1-\beta\log|\mathcal{A}|}{2\beta} + o\left(\frac{1}{\log M}\right), \quad (12)$$

which holds for any $\beta \in (0, \frac{1}{\log|\mathcal{A}|})$, and thus, when combined with Theorem 1, provides a complete characterization for the scaling of the log-cardinality of the largest codebook in a DNA-based storage system with short molecules.

The proof of the achievability bound in [18, Sec. IV] is based on Feinstein's maximal coding bound [20], which bounds the maximal error probability of the optimal codebook of a given cardinality via the cumulative distribution function of the information density of the channel (i.e., the information spectrum). Specifically, the authors of [18] use the extended version stated in [21, Theorem 20.7], which also takes into account input constraints. In contrast, Theorem 1 is proved via a direct route; the pairwise error probability of the optimal maximum likelihood decoder is upper-bounded using Chernoff's bound and the resulting product moment admits a closed-form expression due to Proposition 2 and the fact that the codewords are drawn from the Dirichlet distribution. It turns out that the proof of Theorem 1 is considerably shorter and somewhat easier to follow than the proof in [18, Sec. IV].

IV. Future Work

In this work, we have considered the information density of the DNA storage channel with short molecules. While in the current work we only dealt with the randomness that stems from the random (multinomial) sampling of the DNA molecules, it may be interesting to consider a generalized model, where also the sequencing process is noisy. E.g., one may assume that each of the M molecules is read thru a discrete memoryless channel. A more complicated settings may also consider insertions and deletions while sequencing.

References

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

[2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *nature*, vol. 494, no. 7435, pp. 77–80, 2013.

[3] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. S., "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, pp. 2552–2555, 2015.

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

[5] O. Sabary, H. M. Kiah, P. H. Siegel, and E. Yaakobi, "Survey for a decade of coding for DNA storage," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 10, no. 2, pp. 253–271, 2024.

[6] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2331–2351, 2019.

[7] N. Weinberger, "Error probability bounds for coded-index DNA storage channels," *IEEE Transactions on Information Theory*, vol. 68, no. 11, pp. 7005–7022, 2022.

[8] I. Shomorony and R. Heckel, "DNA-based storage: Models and fundamental limits," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3675–3689, 2021.

[9] ——, "Information-theoretic foundations of DNA data storage," *Foundations and Trends® in Communications and Information Theory*, vol. 19, no. 1, pp. 1–106, 2022.

[10] N. Weinberger and N. Merhav, "The DNA storage channel: Capacity and error probability bounds," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 5657–5700, 2022.

[11] Y. H. Ling and J. Scarlett, "Exact error exponents of concatenated codes for dna storage," *IEEE Transactions on Information Theory*, 2025.

[12] Y. H. Ling, N. Weinberger, and J. Scarlett, "Error exponents for DNA storage codes with a variable number of reads," *arXiv preprint arXiv:2504.17337*, 2025.

[13] A. Kobovich and N. Weinberger, "Input optimization in the composite dna storage channel," *IEEE Journal on Selected Areas in Information Theory*, 2025.

[14] D. Bar-Lev, I. Orr, O. Sabary, T. Etzion, and E. Yaakobi, "Scalable and robust DNA-based storage via coding theory and deep learning," *Nature Machine Intelligence*, pp. 1–11, 2025.

[15] P. K. Vippathalla and N. Kashyap, "The secure storage capacity of a DNA wiretap channel model," *IEEE Transactions on Information Theory*, 2023.

[16] W. Zhang and Z. Wang, "Secret sharing for DNA probability vectors," in *ICC 2024-IEEE International Conference on Communications*. IEEE, 2024, pp. 4578–4583.

[17] A. Lapidoth and S. M. Moser, "On the capacity of the discrete-time Poisson channel," *IEEE Transactions on Information Theory*, vol. 55, no. 1, pp. 303–322, 2008.

[18] Y. Gerzon, I. Shomorony, and N. Weinberger, "Capacity of frequency-based channels: Encoding information in molecular concentrations," *IEEE Transactions on Information Theory*, 2025.

[19] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, 2017.

[20] A. Feinstein, "A new basic theorem of information theory," 1954.

[21] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge University Press, 2023+. [Online]. Available: https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf

[22] F. Bello, A. Martin, T. Rischewski, and G. Seroussi, "The lattice-input discrete-time Poisson channel," in *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024, pp. 3624–3629.

[23] R. Tamir and N. Weinberger, "Achievable rates and error probability bounds of frequency-based channels of unlimited input resolution," *arXiv preprint arXiv:2504.18364*, 2025.

[24] R. Tamir, N. Weinberger, and A. Guillén i Fàbregas, "DNA storage in the short molecule regime," *arXiv preprint arXiv:2511.14284*, 2025.

[25] N. Balakrishnan and V. B. Nevzorov, *A primer on statistical distributions*. John Wiley & Sons, 2004.