SCIENCE CHINA
Information Sciences

• LETTER •

May 2022, Vol. 65 159304:1–159304:2
https://doi.org/10.1007/s11432-021-3394-2

Error exponent for concatenated codes in DNA data storage under substitution errors

Yuxuan SHI[1], Shuo SHAO[1], Xiaohang ZHANG[2*], Yongjian WANG[2] & Yongpeng WU[3]

[1]*School of Cyber and Engineering, Shanghai Jiao Tong University, Shanghai* 200240*, China;*
[2]*The National Computer Network Emergency Response Technical Team, Coordination Center of China, Beijing* 100029*, China;*
[3]*Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai* 200240*, China*

Received 18 August 2021/Revised 1 December 2021/Accepted 4 January 2022/Published online 13 April 2022

Citation Shi Y X, Shao S, Zhang X H, et al. Error exponent for concatenated codes in DNA data storage under substitution errors. Sci China Inf Sci, 2022, 65(5): 159304, https://doi.org/10.1007/s11432-021-3394-2

Dear editor,

The idea to store information on DNA has gained significant attention, due to the explosion of the demand for current data centers and storage techniques in recent years. DNA molecules are well known for their superior information density and lifetime [1], which are far beyond what current tapes and discs could achieve. Notice that the DNA storage problem is considered as a channel coding problem instead of a source coding one since the errors in data recovery are non-negligible in DNA-based storage. Besides the substitutions, these errors include insertions, deletions, permutation of strands, and loss of ordering information [2]. Hence it is necessary to develop error-correcting codes for reliable and efficient storage, such as indexed-based coding [3] to recover the disordered information, anchor-based coding [4] against the substitutions. Moreover, several studies concern about an arbitrarily-permuted parallel DNA channel and corresponding concatenated codes for error correction. Lenz et al. [5] provided a multi-draw channel, under which the overall capacity and achievable code rate [6] are studied. However the achievable error-correcting performance remains unclear, and it lacks numerical results with engineering competitive error-correcting families like polar and low density parity check (LDPC) codes.

Inspired by [5], we consider a specific concatenated codes framework with a pre-decoder and analyze its error-correcting performance under the DNA-based multi-draw channel. More precisely, we present the Forney's error exponent [7] and the upper bound on the achievable maximal overall rate of this code concatenation. Furthermore, the capacity achievability of this concatenated family is verified through a practical combination with polar codes.

DNA storage channel. As the channel in [5], we model the DNA multi-draw channel with substitution errors only to simplify the derivation. The model has an input strands set $\boldsymbol{\mathcal{X}}^M = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M\}$, where $\boldsymbol{X}_i \in \mathbb{F}_2^L$, $1 \leqslant i \leqslant M$ represents a binary vector of length L. From the input, a total of N strands are drawn with replacement, each uniformly at random, and received with bit flip rate p, resulting in an unordered output set $\boldsymbol{\mathcal{Y}}^N = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N\}$, where each output sequence is $\boldsymbol{Y}_j = \boldsymbol{X}_{I_j} \oplus \boldsymbol{E}_j$, where I_j denotes i.i.d. uniform random draw with $\mathbb{P}(I_j = i) = \frac{1}{M}$, and $E_j \in \mathbb{F}_2^L$ denotes independent error vector consisting of L i.i.d. Bernoulli entries with flip rate p. Furthermore, the coverage depth is denoted by $c = \frac{N}{M}$, i.e., the average number of times that each sub-block has been drawn. The exact number of times that the i-th input sequence $\boldsymbol{X}_i$ has been drawn is denoted by D_i with realization d. The D_i sampled sequences, derived from the i-th input $\boldsymbol{X}_i$, are clustered into subset $\boldsymbol{Z}_i$ according to their ordering information. After that the cluster gets retrieved to $\hat{\boldsymbol{X}}_i$ by bit-wised majority decision.

Coding scheme.

• Encoding. The original data sequences come from a data archive, each of which is a q-ary vector with length K. A piece of message $\boldsymbol{h} \in \mathbb{F}_q^K$ is first encoded via an $[M, K, d_{\text{out}}]$ maximum distance separable (MDS) code $\mathcal{C}_{\text{out}}$ over $\mathbb{F}_{q^z}$, whose code rate $R = \frac{K}{M}$. Secondly, the binary unique representation of these sequences are encoded via an $[L(1-\beta), k, d_{\text{in}}]$ inner code $\mathcal{C}_{\text{in}}$ over $\mathbb{F}_2$. Thirdly, passing the index encoder increases the code length to L and output the strands set $\boldsymbol{\mathcal{X}}^M$, resulting in the inner code rate $r = \frac{k}{L}$.

• Clustering. Under the assumption that labels are transmitted error-free, the receiver is able to construct i-th exact cluster without errors according to the labels $I_j = i$. The expectation of the cluster size equals to coverage depth, i.e., $\frac{1}{M}\sum_{i=1}^{M} D_i = c$.

• Decoding. After clustering, we obtain M clusters and the distribution of their sizes, $\boldsymbol{d}^M$. We use a minimum distance decoder which conducts bit-wise majority decision on i-th cluster if the size satisfies $d = |\boldsymbol{Z}_i| \geqslant \theta$, and discards otherwise, where θ is artificially controlled to exclude the symbols from bad sub-channels. These discarded clusters are treated as erasures and corrected by the outer MDS code. Then the estimated output $\{\hat{\boldsymbol{X}}_i\}_1^M$ is decoded by indexing,

* Corresponding author (email: zhangxiaohang@cert.org.cn)

inner and outer decoders successively, finally recovered to the estimated message $\hat{\boldsymbol{h}}$.

Error exponent. By combining the above inner and outer codes, we obtain a specific DNA error-correcting code $[ML, Kk, \geqslant d_{\text{out}}d_{\text{in}}]$ with overall code rate $\mathcal{R} = Rr = \frac{Kk}{ML}$, which we denote as $\mathcal{C}_{\text{cont}}$. With the aforementioned coding scheme, we give the statement of error exponent of $\mathcal{C}_{\text{cont}}$.

Theorem 1. There exists $\mathcal{C}_{\text{cont}}$ with overall rate $\mathcal{R}$, inner rate r, rate loss β due to index coding, coverage depth c and a pre-decoding parameter θ, whose probability of decoding error $P_{\text{err}}(\mathcal{C}_{\text{cont}})$ is upper bounded by

$$-\frac{\log_q P_{\text{err}}(\mathcal{C}_{\text{cont}})}{LM} \geqslant E_D(p^\star, \mathcal{R}) - o(1), \tag{1}$$

where $p^\star = \sum_{i=\theta/2}^{\theta} B_{\theta,p}(i)$ and

$$E_D(p^\star,\mathcal{R}) = \max_{r\leqslant C} \frac{1}{2} E_G(p^\star, r(1-\beta)) \left(\sum_{d\geqslant\theta} p_c(d) - \frac{\mathcal{R}}{r(1-\beta)} \right),$$

$$C = \sum_{d=0}^{\infty} p_c(d) C_d - \beta \left(1 - \mathrm{e}^{-c}\right),$$

$$C_d = 1 + \sum_{i=0}^{d} B_{d,p}(i) \log \left(\frac{B_{d,p}(i)}{B_{d,p}(i) + B_{d,p}(1-i)} \right).$$

$E_G(p, r)$ is Gallager error exponent of bit flip rate p and code rate r. C is the upper bound of the capacity of the DNA multi-drawing channel, and the capacity of sub-channel only concerns about the drawing times d, hence we denote it by C_d. Meanwhile $p_c(d) = \mathrm{e}^{-c}c^d/d!$ and $B_{\theta,p}(i) = \binom{\theta}{i} p^i (1-p)^{\theta-i}$ refer to the probability mass function of Poisson and binomial distribution, respectively.

Remark 1. Notably $E_D(p^\star, \mathcal{R})$ describes the performance of the pre-decoding threshold θ on the error exponent. As follows we obtain the corollary on the optimal code rate choices intuitively by satisfying $E_D(p^\star, \mathcal{R}) \geqslant 0$. If we use some optimal pre-decoder process, i.e., the threshold of cluster size at the decoder is set to $\theta = \min\{d \in \mathbb{N} : C_d \geqslant r\}$, the corollary coincides with the Theorem 1 in [6] without consecutive strands.

Corollary 1. Consider the setting of Theorem 1, there exists a kind of concatenated codes with overall rate $\mathcal{R}$, inner rate r, rate loss of indexing β, and coverage depth c, whose maximum achievable code rate is given by

$$\mathcal{R} \leqslant \sum_{d\geqslant\theta} p_c(d) r(1-\beta). \tag{2}$$

Numerical results. To verify the capacity-achievable capability of the concatenated codes in such a DNA multi-drawing channel, we compare the channel capacity from [5] with the theoretical upper bound of code rate given in Corollary 1. Furthermore, the numerical results of the maximum achievable rate of the RS-polar codes are also presented. In [8], this family is provably capacity-achievable in arbitrarily-permuted parallel channels.

We fix $p = 0.1$, $\beta = 0.03$, optimal θ in Remark 1, and an RS-polar family with adjustable inner rates which combines 2^7-length RS and 2^6-length interleaved polar codes. In Figure 1 the overall rates/capacity are plotted over the inner code rates under different coverage depth c. The dashed line stands for the theoretical achievable overall rate $\mathcal{R}$ and the circle line represents the rate of simulation results $\mathcal{R}_{\text{RS-polar}}$. The channel capacity is illustrated by the dash-dotted lines. In both cases of c, the achievable rates of RS-polar codes are close to (2). Besides, the RS-Polar codes present a faster descent than the theoretic bound on code rates, due to the finite code length and non-negligible errors in sub-blocks. Given larger c, the overall rate increases obviously, e.g., with $c = 10$, $\mathcal{R}_{\text{RS-polar}}$ approaches 0.8. Furthermore, the achievable code rate of code concatenation appears a large gap in comparison to the channel capacity when c is small, while the gap converges to 0 with increasing coverage depth. We conclude the upper bound on maximal achievable overall rates can approach the channel capacity with large enough c.

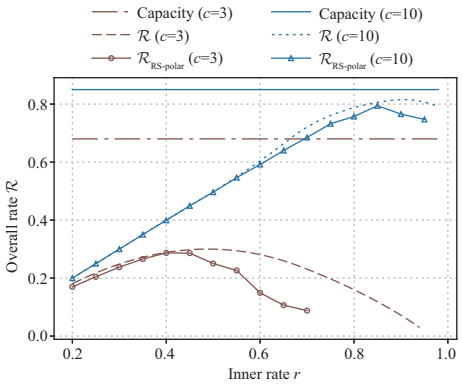

Figure 1 (Color online) Comparison between capacity, theoretical, and specific achievable code rates.

Conclusion. Under this multi-drawing channel model, we present the upper bound on the error probability and the maximal achievable overall rate of this concatenated scheme. Moreover, given large enough coverage depth, the achievable overall rate of $\mathcal{C}_{\text{cont}}$ is tight enough in comparison with the channel capacity.

Acknowledgements This work was supported in part by National Key R&D Program of China (Grant No. 2018YFB1801102) and National Natural Science Foundation of China (Grant Nos. 62122052, 62071289).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Yazdi S M H T, Kiah H M, Garcia-Ruiz E, et al. DNA-based storage: trends and methods. IEEE Trans Mol Biol Multi-Scale Commun, 2015, 1: 230–248
2 Lenz A, Siegel P H, Wachter-Zeh A, et al. Coding over sets for DNA storage. IEEE Trans Inform Theory, 2020, 66: 2331–2351
3 Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science, 2012, 337: 1628–1628
4 Lenz A, Siegel P H, Wachter-Zeh A, et al. Anchor-based correction of substitutions in indexed sets. In: Proceedings of IEEE International Symposium on Information Theory (ISIT), 2019. 757–761
5 Lenz A, Siegel P H, Wachter-Zeh A, et al. An upper bound on the capacity of the DNA storage channel. In: Proceedings of IEEE Information Theory Workshop (ITW), 2019
6 Lenz A, Welter L, Puchinger S. Achievable rates of concatenated codes in DNA storage under substitution errors. In: Proceedings of International Symposium on Information Theory and Its Applications, 2020. 269–273
7 Forney G. Concatenated Codes. Cambridge: MIT Press, 1966
8 Hof E, Sason I, Shamai S, et al. Capacity-achieving polar codes for arbitrarily permuted parallel channels. IEEE Trans Inform Theory, 2013, 59: 1505–1516