Linear-Time Encoding/Decoding of Irreducible Words for Codes Correcting Tandem Duplications

Yeow Meng Chee, Johan Chrisnata, Han Mao Kiah, and Tuan Thanh Nguyen
School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
email: {ymchee, johanchr001, hmkiah, nguyentu001}@ntu.edu.sg

***Abstract*—Tandem duplication is the process of inserting a copy of a segment of DNA adjacent to the original position. Motivated by applications that store data in living organisms, Jain *et al.* (2017) proposed the study of codes that correct tandem duplications. All code constructions are based on *irreducible words*.**

We provide efficient encoders/decoders for codes correcting tandem duplications whose codewords are irreducible. First, we describe an (ℓ, m)-finite state encoder and show that when $m = \Theta(1/\epsilon)$ and $\ell = \Theta(1/\epsilon)$, the encoder has rate that is ϵ away from the optimal. We then use combinatorial method to reduce the space requirements for the finite state encoder.

I. Introduction

Tandem duplications or repeats is one of the two common repeats found in the human genome [1] and they are caused by slipped-strand mispairings [2]. They occur in DNA when a pattern of one or more nucleotides is repeated and the repetitions are directly adjacent to each other.

Jain *et al.* [3] first proposed the study of codes that correct errors due to tandem duplications. In the same paper, Jain *et al.* used *irreducible* words (see Section II for definition) to construct a family of codes that correct tandem duplications of lengths at most k, where $k \in \{2, 3\}$. While these codes are optimal in size for the case $k = 2$, these codes are not optimal for $k = 3$, and in fact, Chee *et al.* [4] constructed a family of codes with strictly larger size. Unfortunately, the asymptotic rate of the latter is the same as the codes in [3].

In this work, we first develop a recursive formula to find the exact number of irreducible words for arbitrary length, and hence provide a closed formula for the asymptotic rate of codes in [3]. Table I demonstrates that the rate of such codes are almost optimal for $q \geq 5$. We then look at encoding/decoding methods for irreducible words. In particular, we provide a linear-time algorithm that encodes irreducible words and the rate of such encoder is close to the asymptotic rates of irreducible words. Due to space constraints, we only summarise the results and describe the main idea of the algorithms. Details can be found in [7] and the results have been presented in ISIT 2018.

II. Notation and Terminology

Let $[n]$ denote the set $\{1, 2, \ldots, n\}$. Let $\Sigma_q = \{0, 1, \cdots q-1\}$ be an alphabet of $q \geqslant 2$ symbols. For a positive integer n, let Σ_q^n denote the set of all words of length n over Σ_q, and let Σ_q^* denote the set of all words over Σ_q with finite length. Given two words $\boldsymbol{x}, \boldsymbol{y} \in \Sigma_q^*$, we denote their concatenation by $\boldsymbol{xy}$.

We state the *tandem duplication* rules. For integers $k \leqslant n$ and $i \leqslant n - k$, we define $T_{i,k} : \Sigma_q^n \to \Sigma_q^{n+k}$ such that $T_{i,k}(\boldsymbol{x}) = \boldsymbol{uvvw}$, where $\boldsymbol{x} = \boldsymbol{uvw}$, $|\boldsymbol{u}| = i$, $|\boldsymbol{v}| = k$.

If a finite sequence of tandem duplications of length at most k is performed to obtain $\boldsymbol{y}$ from $\boldsymbol{x}$, then we say that $\boldsymbol{y}$ is a $\leqslant k$*-descendant* of $\boldsymbol{x}$, or $\boldsymbol{x}$ is a $\leqslant k$*-ancestor* of $\boldsymbol{y}$. Given a word $\boldsymbol{x}$, we define the $\leqslant k$*-descendant cone* of $\boldsymbol{x}$ is the set of all $\leqslant k$-descendants of $\boldsymbol{x}$ and denote this cone by $D^*_{\leqslant k}(\boldsymbol{x})$.

Definition 1 ($\leqslant k$-Tandem-Duplication Codes)**.** A subset $\mathcal{C} \subseteq \Sigma_q^n$ is a $\leqslant k$*-tandem-duplication code* if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ and $\boldsymbol{x} \neq \boldsymbol{y}$, we have that $D^*_{\leqslant k}(\boldsymbol{x}) \cap D^*_{\leqslant k}(\boldsymbol{y}) = \varnothing$. We say that $\mathcal{C}$ is an $(n, \leqslant k; q)$-TD code.

The *size* of $\mathcal{C}$ refers to $|\mathcal{C}|$, while the *rate* of $\mathcal{C}$ is given by $(1/n) \log_q |\mathcal{C}|$. Given an infinite family $\{\mathcal{C}_n : \mathcal{C}_n \text{ is of length } n\}_{n=1}^{\infty}$, its *asymptotic rate* is given by $\lim_{n\to\infty} (1/n) \log_q |\mathcal{C}_n|$.

Definition 2. A word is $\leqslant k$*-irreducible* if it cannot be deduplicated into shorter words with deduplications of length at most k. We use $\mathrm{Irr}_{\leqslant k}(n, q)$ to denote the set of all $\leqslant k$-irreducible words of length n over Σ_q.

Construction 1 (Jain *et al.* [3])**.** *For $k \in \{1, 2, 3\}$ and $n \geqslant k$. An $(n, \leqslant k; q)$-TD-code $\mathcal{C}(n, \leqslant k; q)$ is given by*

$$\mathcal{C}(n, \leqslant k; q) \triangleq \bigcup_{i=1}^{n} \{\xi_{n-i}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{Irr}_{\leqslant k}(i, q)\}.$$

Here, $\xi_i(\boldsymbol{x}) = \boldsymbol{x}z^i$, where z is the last symbol of $\boldsymbol{x}$.

Let $I_{\leqslant k}(n, q) \triangleq |\mathrm{Irr}_{\leqslant k}(n, q)|$. Then the size of $\mathcal{C}(n, \leqslant k; q)$ is given by $\sum_{i=1}^{n} I_{\leqslant k}(i, q)$. Let $\mathrm{rate}_{\leqslant k}(n, q)$ and $\mathrm{rate}_{\leqslant k}(q)$ denote the rate and asymptotic rate of $\mathcal{C}(n, \leqslant k; q)$, respectively. In other words, $\mathrm{rate}_{\leqslant k}(n, q) \triangleq (1/n) \log_q |\mathcal{C}(n, \leqslant k; q)|$ and $\mathrm{rate}_{\leqslant k}(q) \triangleq \lim_{n\to\infty} \mathrm{rate}_{\leqslant k}(n, q)$. Jain *et al.* observed that $\bigcup_{n=1}^{\infty} \mathrm{Irr}_{\leqslant k}(n, q)$ is a regular language and hence,

$$\mathrm{rate}_{\leqslant k}(q) = \lim_{n\to\infty} \frac{\log_q I_{\leqslant k}(n, q)}{n}. \quad (1)$$

Furthermore, using Perron-Frobenius theory (see [6]), Jain *et al.* computed $\mathrm{rate}_{\leqslant 3}(3)$ to be approximately 0.347934. In view of (1), we look at encoding of the words in $\mathrm{Irr}_{\leqslant k}(n, q)$ instead and the extension of our encoding methods to $\mathcal{C}(n, \leqslant k; q)$ is straightforward.

In this work, we focus on the case $k \in \{2, 3\}$ as the results for $k = 1$ is well known. Specifically, the size of $\mathrm{Irr}_{\leqslant 1}(n, q)$ is given by $q(q-1)^{n-1}$ and linear-time encoding methods can be obtained via differential coding (see for example, [6]).

III. Enumerating Irreducible Words

Proposition 1. *We have that $I_{\leqslant 2}(2, q) = q(q-1)$, $I_{\leqslant 2}(3, q) = q(q-1)^2$, and*

$$I_{\leqslant 2}(n, q) = (q-2) I_{\leqslant 2}(n-1, q) + (q-2) I_{\leqslant 2}(n-2, q) \quad (2)$$

for $n \geqslant 4$. Therefore, the asymptotic rate is $\mathrm{rate}_{\leqslant 2}(q) = \log_q \lambda_2$, where $\lambda_2 = (q - 2 + \sqrt{q^2 - 4})/2$.

Proposition 2. *We have that* $I_{\leqslant 3}(3,q) = q(q-1)^2$*,* $I_{\leqslant 3}(4,q) = q^2(q-1)(q-2)$*,* $I_{\leqslant 3}(5,q) = q(q-1)(q-2)(q^2-q-1)$ *and*

$$I_{\leqslant 3}(n,q) = (q-2)I_{\leqslant 3}(n-1,q) + (q-3)I_{\leqslant 3}(n-2,q) \\ + (q-2)I_{\leqslant 3}(n-3,q) \qquad (3)$$

for $n \geqslant 6$*. Therefore,* $\mathrm{rate}_{\leqslant 3}(q) = \log_q \lambda_3$*, where* λ_3 *is the largest real root of equation* $x^3-(q-2)x^2-(q-3)x-(q-2) = 0$*.*

We compute the values of $\mathrm{rate}_{\leqslant k}(q)$ for $k \in \{2,3\}$ in Table I. Let $T(n,q)$ be the largest size of an $(n, \leqslant 3; q)$-TD code and define $\tau(q) \triangleq (1/n)\limsup_{n\to\infty} \log_q T(n,q)$. From [3], [4], we have that that $\mathrm{rate}_{\leqslant 3}(q) \leqslant \tau(q) \leqslant \mathrm{rate}_{\leqslant 2}(q)$. Therefore, Table I demonstrates that $\mathcal{C}(n, \leqslant 3; q)$ is *almost* optimal for $q \geqslant 5$.

q	3	4	5	6	7	8
$\mathrm{rate}_{\leqslant 2}(q)$	0.4380	0.7249	0.8280	0.8788	0.9081	0.9269
$\mathrm{rate}_{\leqslant 3}(q)$	0.3479	0.7054	0.8208	0.8753	0.9062	0.9258

TABLE I: The asymptotic information rates for k-irreducible words for $k \in \{2,3\}$

IV. Finite State Encoder

For integers $\ell < m$, an (ℓ,m)-*finite state encoder* is triple $(\mathcal{S},\mathcal{E},\mathcal{L})$, where $\mathcal{S}$ is a set of *states*, $\mathcal{E} \subset \mathcal{S}\times\mathcal{S}$ is a set of *directed edges*, and $\mathcal{L} : \mathcal{E} \to \Sigma_q^\ell \times \Sigma_q^m$ is an *edge labeling*.

To encode irreducible words, we choose $m \geqslant 2k-1$, and set

$$\mathcal{S} \triangleq \mathrm{Irr}_{\leqslant k}(m,q) \text{ and } \mathcal{E} \triangleq \{(\boldsymbol{x},\boldsymbol{x}') : \boldsymbol{x}\boldsymbol{x}' \in \mathrm{Irr}_{\leqslant k}(2m,q)\}.$$

For $\boldsymbol{x} \in \mathcal{S}$, we define the *neighbours* of $\boldsymbol{x}$ to be $N(\boldsymbol{x}) \triangleq \{\boldsymbol{x}' : (\boldsymbol{x},\boldsymbol{x}') \in \mathcal{E}\}$. We also consider the quantity $\Delta_{\leqslant k}(m,q) \triangleq \min\{|N(\boldsymbol{x})| : \boldsymbol{x} \in \mathcal{S}\}$ and choose ℓ such that

$$\Delta_{\leqslant k}(m,q) \geqslant q^\ell. \qquad (4)$$

We now define the edge labelling $\mathcal{L}$ using this choice of ℓ. For $\boldsymbol{x} \in \mathcal{S}$, since $|N(\boldsymbol{x})| \geqslant q^\ell$, we may use the set Σ^ℓ to index the first q^ℓ words in $N(\boldsymbol{x})$. Hence, for $\boldsymbol{x}' \in S$, if $\boldsymbol{x}'$ is one of the first q^ℓ words, we let $\boldsymbol{y}_{\boldsymbol{x}'} \in \Sigma^\ell$ denote the index. Otherwise, we simply set $\boldsymbol{y}_{\boldsymbol{x}'} = -$. Therefore, for $(\boldsymbol{x},\boldsymbol{x}') \in \mathcal{E}$, we set $\mathcal{L}(\boldsymbol{x},\boldsymbol{x}') = (\boldsymbol{y}_{\boldsymbol{x}'},\boldsymbol{x}')$. Finally, we call this triple an (ℓ,m)- *finite state encoder for irreducible words*.

A. Encoding

Let s be a positive integer and set $n = s\ell$. Suppose the message $\boldsymbol{y} = \boldsymbol{y}_1\boldsymbol{y}_2\ldots\boldsymbol{y}_s \in \Sigma^{s\ell}$.

To encode $\boldsymbol{y}$ using an (ℓ,m)-finite state encoder for irreducible words, we do the following:

(I) Set $\boldsymbol{x}_0$ to the first word in $\mathcal{S} = \mathrm{Irr}_{\leqslant k}(m,q)$.

(II) For $i \in [s]$, set $\boldsymbol{x}_i$ to be the unique word such that $\mathcal{L}(\boldsymbol{x}_{i-1},\boldsymbol{x}_i) = (\boldsymbol{y}_i,\boldsymbol{x}_i)$.

(III) The encoded irreducible word is $\boldsymbol{x} = \boldsymbol{x}_1\boldsymbol{x}_2\ldots\boldsymbol{x}_s$.

Since the encoded word has length sm, the (ℓ,m)-finite state encoder for irreducible words has rate ℓ/m. Pick $\epsilon > 0$. We find suitable values for ℓ and m so that the encoding rate satisfies

$$\ell/m \geqslant \mathrm{rate}_{\leqslant k}(q) - \epsilon. \qquad (5)$$

Recall that ℓ and m are required to satisfy (4). Hence, we determine $\Delta_{\leqslant k}(m,q)$.

B. Approaching the Asymptotic Information Rate

Proposition 3. *We have that* $\Delta_{\leqslant 2}(3,q) = q(q-2)^2$*,* $\Delta_{\leqslant 2}(4,q) = (q-2)^2(q^2-q-1)$*, and for* $m \geqslant 5$*,*

$$\Delta_{\leqslant 2}(m,q) = (q-2)\Delta_{\leqslant 2}(m-1,q)+(q-2)\Delta_{\leqslant 2}(m-2,q). \quad (6)$$

Proposition 4. *We have that*

$$\Delta_{\leqslant 3}(5,q) = (q-2)(q^2-2q-1)^2,$$
$$\Delta_{\leqslant 3}(6,q) = (q-1)(q^5-6q^4+9q^3+4q^2-8q-9),$$
$$\Delta_{\leqslant 3}(7,q) = (q-2)(q^6-6q^4+9q^3+4q^2-8q-10q+3),$$

and for $m \geqslant 8$*,*

$$\Delta_{\leqslant 3}(m,q) = (q-2)\Delta_{\leqslant 3}(m-1,q) + (q-3)\Delta_{\leqslant 3}(m-2,q) \\ + (q-2)\Delta_{\leqslant 3}(m-3,q). \qquad (7)$$

Set κ_2 such that $\Delta_{\leqslant 2}(m,q) \geqslant \kappa_2\lambda_2^m$ for $m \in \{3,4\}$. Similarly, set κ_3 so that $\Delta_{\leqslant 3}(m,q) \geqslant \kappa_3\lambda_3^m$ for $m \in \{5,6,7\}$.

Theorem 1. *Let* $k \in \{2,3\}$*. Set* $c_k = \mathrm{rate}_{\leqslant k}(q) = \log_q \lambda_k$*. For* $\epsilon > 0$*, if we choose* m *and* ℓ *such that*

$$\ell = \left\lceil \frac{(c_k-\epsilon)(c_k-\log_q \kappa_k)}{\epsilon} \right\rceil, m = \left\lceil \frac{\ell - \log_q \kappa_k}{c_k} \right\rceil,$$

then the (ℓ,m)*-finite state encoder has rate at least* $\mathrm{rate}_{\leqslant k}(q) - \epsilon$*.*

Therefore, to achieve encoding rates at least $\mathrm{rate}_{\leqslant k}(q) - \epsilon$, we only require $\ell = \Theta(1/\epsilon)$ and $m = \Theta(1/\epsilon)$. If we naively use a lookup table to represent $(\mathcal{S},\mathcal{E},\mathcal{L})$, we require $q^{\Theta(1/\epsilon)}$ space. Furthermore, using binary search, the (ℓ,m)-finite state encoder for irreducible words encodes in $O(n/\epsilon)$ time. In fact, we can use combinatorial insights from (2) and (3) to reduce the space requirement to $O(1/\epsilon^2)$ (refer to our preprint [7]).

V. Further Work

We combine our finite state encoder and Knuth's balancing method [8] to obtain GC-balanced codes that correct tandem duplications. Our recent results also include constructions of codes for duplication length $k \geq 4$.

References

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh et al., "Initial sequencing and analysis of the human genome", *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[2] N. I. Mundy and A. J. Helbig, "Origin and evolution of tandem repeats in the mitochondrial DNA control region of shrikes (lanius spp.)," *Journal of Molecular Evolution*, vol. 59, no. 2, pp. 250–257, 2004.

[3] S. Jain, F. Farnoud, M. Schwartz and J. Bruck, "Duplication-Correcting Codes for Data Storage in the DNA of Living Organisms," *IEEE Trans. Inform. Theory*, vol. 63, no. 8, pp. 4996–5010, 2017.

[4] Y. M. Chee, J. Chrisnata, H. M. Kiah, and T. T. Nguyen, "Deciding the confusability of words under tandem repeats," *preprint arXiv:1707.03956*, 2017.

[5] S. Jain, F. Farnoud, M. Schwartz and J. Bruck, "Noise and Uncertainty in String-Duplication Systems," in *Proc. 2017 IEEE Intl. Symp. Inform. Theory, Aachen, Germany*, Jun. 2017, pp. 3120–3124.

[6] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An Introduction to Coding for Constrained System", Oct 2001.

[7] Y. M. Chee, J. Chrisnata, H. M. Kiah, and T. T. Nguyen, "Efficient encoding/decoding of irreducible words for codes correcting tandem duplications", *preprint arXiv:1801.02310*, 2018.

[8] D. E. Knuth, "Efficient Balanced Codes", *IEEE Trans. Inform. Theory*, vol. IT-32, no. 1, pp. 51-53, Jan. 1986.