Southeast Asian Bulletin of Mathematics (2021) 45: 833–848

Southeast Asian
Bulletin of
Mathematics

On k-Comma Codes*

Haiyan Liu
School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming, Yunnan, China
Email: liuhy19@qq.com

K. P. Shum
Institute of Mathematics, Yunnan University, Kunming, 650091, China
Email: kpshum@ynu.edu.cn

Jing Leng†
Department of Mathematics and Physics, Mianyang Normal University, Mianyang, Sichuan, China
Email: lengjingai0813@163.com

Received 25 May 2021
Accepted 9 September 2021

Communicated by K. Denecke

Dedicated to the memory of Professor Yuqi Guo (1940–2019)

AMS Mathematics Subject Classification(2000): 20M35; 68R15; 68Q70

Abstract. Let L be a nonempty language over an alphabet A and $k \geq 0$. Then, we call L a k-comma code if L satisfies $LA^kL \cap A^+LA^+ = \emptyset$. It is obvious that the k-comma codes are a generalization of comma-free codes. Our aim in this paper is to further study the properties of the k-comma codes. As one can easily observe that the family of infix codes is not closed under composition and therefore, we concentrate on the closure properties of the k-comma codes under composition and decomposition. Consequently, we give a sufficient and necessary condition under which the concatenation of two disjoint k-comma codes X and Y is still a k-comma code under the assumption that $X \cup Y$ is an infix code. Obviously, the 1-1-comma codes are a generalization of 1-comma codes. Thus, by using the structure properties of 1-1-comma codes, we show that the 1-1-comma property is decidable for regular lan-

*This research is supported by National Natural Science Foundation of China #11861051.
†Corresponding author.

guages. Finally, we characterize the automata which accept the words in X_1 where $X_1 = \{u \in A^+ |(\forall w \in A^k) uwu \cap A^+uA^+ = \emptyset\}$.

Keywords: Decidability; Automaton; Combinatorics of words; k-Comma code.

1. Introduction

It is well-known in the literature that the codes play a crucial role in many disciplines such as information processing, data compression, cryptography, information transmission and so on [6, 7, 11, 17]. In recent years, the theory of codes has becomes an important principle in the field of theoretical computer science. In particular, the theory of codes has a wide applications in combinatorics of words. In addition, the study of codes has been rapidly developed in the area of automata, formal languages, theory of semigroup and AG groupoids. For a detailed treatment of the theory of codes, the readers are referred to the well known monogrphies of codes by S. Ling and C. Xing [19], C. E. Shannon [22], R. Hill [10], J. Berrstel, D. Perrin and C. Reutenauver [1].

Various codes with specific algebraic properties such as infix codes, comma-free codes, intercodes and k-comma codes have been initiated and studied for various purpose, for instance, the definition of k-comma codes was first initiated by the idea of genes (coding segments) are usually interrupted by noncoding segments, formerly known as junk segments. In this aspect, C. E. Shannon, R. W. Hamming, E. N. Gilbert and G. E. Sacks are big names for their outstanding contribution to establish a strong foundation of coding theory. For more information on coding theory and their applications, the readers are referred to the papers [6, 7, 9, 18, 22, 24]. It is noted that the theory of k-comma codes have been extensively studied as an independent subject in the theory of variable-length codes, the readers are referred to [2, 4, 20, 21]. On the other hand, regular languages and finite automata are also the essential concepts and tools in the study of theoretic computer science. In an early study, Y. Rabin and A. V. Scott have given some algorithms to test whether or not the language accepted by a given automaton is the empty set, or an infinite set [12]. Further, Ito et al. established the relationship between the regular languages and primitive words [13]. The theory of codes is closely related to formal languages: a code is a language. Most of the decision problems related to code properties are decidable for regular languages whereas they often become undecidable for context-free languages [17]. Decidability of general code properties is also investigated in the literature [3, 5, 8, 14, 15, 16, 21].

This paper is organized as follows. In Section 2, we define some basic notions and notations. In Section 3, we investigate the closure properties of k-comma codes with $k \geq 0$ under composition and decomposition. In Section 4, we give a sufficient and necessary condition under which the concatenation of two disjoint k-comma codes is still a k-comma code under the assumption that the union of the two codes is an infix code. The notion of 1-1-comma codes is a generalization of 1-comma codes. It is known that the family of 1-1-comma codes is equal to 2^{X_1} where $X_1 = \{u \in A^+ | (\forall w \in A^k) uwu \cap A^+uA^+ = \emptyset\}$. In Section 5, we prove the decidability of the property of being a 1-1-comma code for regular languages. In addition, we characterize the automata which accept the words in X_1.

2. Preliminaries

Let A be a finite *alphabet* containing at least two letters and A^* the free monoid generated by A. Then, we call the elements and subsets of A^* the *words* and *languages* over A respectively. Let $A^+ = A^* \backslash \{1\}$ be the free semigroup generated by A, where 1 is the identity of the monoid A^*, called the *empty word*. Then, the length of word w is denoted by $|w|$. As usual, we use $|L|$ to denote the cardinality of language L over A. For any $L_1, L_2, L \subseteq A^*$, we let

$$\begin{aligned} L_1L_2 &= \{xy|\ x \in L_1,\ y \in L_2\}, \\ L^n &= \{x_1x_2 \cdots x_n | x_i \in L,\ i = 1, 2, \cdots, n\}, \\ L_1^{-1}L_2 &= \{\ y \in A^* | (\exists x \in L_1) xy \in L_2\}, \\ L_1L_2^{-1} &= \{\ y \in A^* | (\exists x \in L_2) yx \in L_1\}. \end{aligned}$$

For any two words x, y in A^*, we call x a *prefix* (resp. *suffix*, *infix*) of y if $y = xu$ (resp. $y = ux$, $y = uxv$) for some $u, v \in A^*$. Call a prefix (resp. suffix, infix) x of y is *proper* (resp. *nontrivial*) if $x \neq y$(resp. $0 < |x| < |y|$). Call x is an internal factor if $u, v \neq 1$. It is clear to see the following three relations

$$\begin{aligned} \leq_p &= \{(x, y) \in A^* \times A^* | x \text{ is a prefix of } y\}, \\ \leq_s &= \{(x, y) \in A^* \times A^* | x \text{ is a suffix of } y\}, \\ \leq_d &= \{(x, y) \in A^* \times A^* | x \leq_p y \ \&\ x \leq_s y\}, \end{aligned}$$

on A^* are partial orders. Now we denote the set of all nontrivial prefixes (resp. nontrivial suffixes) of y by $P(y)$ (resp. $S(y)$). For a language L over A, we let

$$\begin{aligned} P(L) &= \bigcup_{y \in L} P(y), \\ S(L) &= \bigcup_{y \in L} S(y), \\ PS(L) &= P(L) \cap S(L). \end{aligned}$$

Let $y \in A^+$. Then, we call the word y *primitive* if y is not the power of another word. It is well known that every word $y \in A^+$ can be expressed uniquely as a power of a primitive word. The word y is said to be *unbordered* if $x \leq_d y$ for some $x \in A^+$ implies $x = y$. Otherwise, y is said to be *bordered*. The set of all primitive (resp. unbordered) words is denoted by Q (resp. U).

Let $L \subseteq A^+$. Then, we call L a *code* over A if

$$x_1 x_2 \cdots x_m = y_1 y_2 \cdots y_n, \quad x_i, y_j \in L, \; i = 1, 2, \cdots, m, \; j = 1, 2, \cdots, n$$

implies that $m = n$ and $x_i = y_i$, $i = 1, 2, \cdots, n$. L is called a *prefix code* (resp. *suffix code*, *infix code*) if no word in L is a prefix (resp. suffix, infix) of another word in L. Also, we call L a *bifix code* if it is both a prefix code and a suffix code. Obviously, every infix code is a bifix code.

A (finite) *automaton* is a quintuple $\mathcal{A} = (S, A, \delta, s_0, T)$ such that:

S is a finite nonempty set called the set of *states*,

A is a finite alphabet called the set of *inputs*,

δ is a mapping from $S \times A$ into S called the *state transition function*,

s_0 is a distinguished element of S called the *initial state*, and T is a subset of S called the set of *final states*.

The function δ can be extended to a mapping of $S \times A^*$ into S by $\delta(s, 1) = s$ and $\delta(s, xa) = \delta(\delta(s, x), a)$ for any $s \in S$, $x \in A^*$ and $a \in A$.

For an automaton $\mathcal{A} = (S, A, \delta, s_0, T)$, we let $T(\mathcal{A}) = \{x \in A^* | \delta(s_0, x) \in T\}$. Then, $T(\mathcal{A})$ is said to be the language *accepted* by $\mathcal{A}$. Thus, we call a language $L \subseteq A^*$ *regular* if there exists an automaton $\mathcal{A}$ such that $L = T(\mathcal{A})$.

We now give the the following crucial definitions.

Definition 2.1. [23] *Let L be a nonempty language over A. Then, L is said to be a comma-free code if $L^2 \cap A^+ L A^+ = \emptyset$.*

Definition 2.2. [4] *Let L be a nonempty language over A and $k \geq 0$. Then, L is called a k-comma code if $LA^k L \cap A^+ L A^+ = \emptyset$.*

Clearly, a 0-comma code is just a comma-free code. It is well-known that each word in a k-comma code is longer than k [4].

Definition 2.3. [4] *Let L be a nonempty language over A and $k \geq 0$. Then, L is called a 1-k-comma code if every singleton of L is a k-comma code.*

3. Composition and Decomposition of k-Comma Codes

Firstly, we introduce the notion of composition of codes.

Definition 3.1. [1] *Let Y and Z be two codes over B and A respectively with $B =$ alph(Y). If there exists a bijection β: $B \to Z$, then for any word $w = b_1 b_2 \cdots b_s$, $b_i \in B$, $i = 1, 2, \cdots, s$, the correspondence*

$$w \;\mapsto\; \beta(b_1)\beta(b_2)\cdots\beta(b_s)$$

provides a free monoid homomorphism β^*: $B^* \to Z^*$. We treat β^* as β. It can be easily checked that $X = \beta(Y) \subseteq Z^+$ is a code, which is called the composition *of Y and Z by means of β and denoted by $X = Y \circ_\beta Z$ or simply $X = Y \circ Z$ when the context permits it.*

It is easy to check that the family of infix codes is not closed under composition. For example, let $B = \{c, d, e\}$, $A = \{a, b\}$,

$$Y = \{c, de\} \subseteq B^+, \quad Z = \{a^3b^2, aba^3, b^2a^2b\} \subseteq A^+$$

and $\beta : B^* \to A^*$, where $\beta(c) = a^3b^2$, $\beta(d) = aba^3$, $\beta(e) = b^2a^2b$. It is clear that $X = Y \circ_\beta Z = \{a^3b^2, aba^3b^2a^2b\}$ is not an infix code. However, the conclusion is completely different for some particular families of infix codes, for example k-comma codes, $k \geq 0$.

Firstly, we introduce a known result on the composition of comm-free codes.

Lemma 3.2. [1] *Let Y, Z be two composable codes and $X = Y \circ Z$. If Y and Z are both comma-free codes, then so is X.*

For $k \geq 0$, it is known that the family of k-comma codes and comma-free codes are incomparable, but not disjoint[4], for example, language $\{ab^{k+1}a\}$ is both a k-comma code and a comma-free code.

Theorem 3.3. *Let Y, Z be two composable codes, $X = Y \circ Z$ and $k \geq 0$. If Y and Z are both k-comma codes and comma-free codes, then X is both a k-comma code and a comma-free code.*

Proof. Assume that $X, Z \subseteq A^*$, $Y \subseteq B^*$, $\beta : B^* \to A^*$ a monomorphism and $\beta(B) = Z$, $\beta(Y) = X$. Then, by Lemma 3.2, X is common-free. Suppose that X is not a k-comma code. Then there exists $x, y, z \in X$, $u, v \in A^+$ and $w \in A^k$ such that $xwy = uzv$. Since $X \subseteq Z^*$, we can let

$$\begin{aligned} x &= z_1^{(1)} z_2^{(1)} \cdots z_m^{(1)}, \\ y &= z_1^{(2)} z_2^{(2)} \cdots z_n^{(2)}, \\ z &= z_1^{(3)} z_2^{(3)} \cdots z_l^{(3)}. \end{aligned}$$

where $z_1^{(1)}, z_2^{(1)}, \cdots, z_m^{(1)}, z_1^{(2)}, z_2^{(2)}, \cdots, z_n^{(2)}, z_1^{(3)}, z_2^{(3)}, \cdots, z_l^{(3)} \in Z$, $m, n, l \geq 1$, that is,

$$xwy = z_1^{(1)} z_2^{(1)} \cdots z_m^{(1)} w z_1^{(2)} z_2^{(2)} \cdots z_n^{(2)} = uzv.$$

Since X is an infix code, z is neither an infix of x nor y. According to the lengths of u and x, we divide the proof into the following two cases:

Case 1. If $|u| < |x|$.

When $|y| \leq |v| < |wy|$, since Z is a k-comma code, every word in Z is longer than k. The fact that Z is an infix code implies that $z_l^{(3)}$ is a proper infix of $z_m^{(1)} w$, it contradicts to Z being a k-comma code. When $|z_2^{(2)} \cdots z_n^{(2)}| \leq |v| < |y|$, if $|v| = |z_2^{(2)} \cdots z_n^{(2)}|$, since Z is an infix code, $z_{l-1}^{(3)}$ is a proper infix of $z_m^{(1)} w$, which contradicts to Z being a k-comma code. If $|z_2^{(2)} \cdots z_n^{(2)}| < |v| < |y|$, the fact that Z is an infix code also yields that $z_l^{(3)}$ is an internal factor of $z_m^{(1)} w z_1^{(2)}$, a contradiction. When $|v| < |z_2^{(2)} \cdots z_n^{(2)}|$, there exists a positive integer p and $z' \in A^*$, where $2 \leq p \leq n$, $z' <_s z_p^{(2)}$ such that $v = z' z_{p+1}^{(2)} \cdots z_n^{(2)}$. If $z' = 1$, notice that $|u| < |x|$, so $p < l$. Since Z is an infix code, we have $z_{p-i}^{(2)} = z_{l-i}^{(3)}$, $i = 0, 1, \cdots, p-1$, and $z_{l-p}^{(2)}$ is a proper infix of $z_m^{(1)} w$, a contradiction. If $z' \neq 1$, Z is an infix code also implies that $z_l^{(3)}$ is an internal factor of $z_{p-1}^{(2)} z_p^{(2)}$, which contradicts to Z being comma-free.

Case 2. If $|x| \leq |u| < |xw|$, since Z is an infix code and every word in Z is longer than k, we have $|xw| < |uz_1^{(3)}| \leq |xwz_1^{(2)}|$, but in this case, $z_1^{(3)}$ is a proper infix of $wz_1^{(2)}$, which contradicts to Z being a k-comma code.

Therefore, we complete the proof. ■

Now we discuss the decomposition of k-comma codes. Firstly, we consider the special case $k = 0$.

Theorem 3.4. *Let X, Y, Z be codes with $X = Y \circ Z$. If X is a comma-free code, then so is Y.*

Proof. Assume that $X, Z \subseteq A^*$, $Y \subseteq B^*$, $\beta : B^* \to A^*$ be a monomorphism with $\beta(B) = Z$ and $\beta(Y) = X$. Suppose that Y is not a comma-free code. Then, there exist $x, y, z \in Y$ and $u, v \in B^+$ such that $xy = uzv$. It implies that

$$\beta(xy) = \beta(x)\beta(y) = \beta(uzv) = \beta(u)\beta(z)\beta(v).$$

Notice $\beta(u), \beta(v) \in Z^+$ and $\beta(x), \beta(y), \beta(z) \in X$. Hence X is not a comma-free code, a contradiction. ■

As a consequence of the above theorem, we give the following remarks:

Remark 3.5. For three codes X, Y, Z with $X = Y \circ Z$, if X is a comma-free code, then Z may not be a comma-free code. For example, let $A = \{a, b\}$, $B = \{c, d\}$,

$$Y = \{cd\} \subseteq B^+, \quad Z = \{a^2b, aba\} \subseteq A^+$$

and $\beta : B^* \to A^*$, where $\beta(c) = a^2b$, $\beta(d) = aba$. It is easy to check that $X = \{a^2baba\}$, X is a comma-free code, but Z not.

Remark 3.6.

(i) For X, Y, Z be codes over A with $X = Y \circ Z$, $k \geq 1$, if X is a k-comma code, then both Y and Z may not be k-comma codes. For example, let $A = \{a, b\}$, $B = \{c\}$,

$$Y = \{c^2\} \subseteq B^+, \quad Z = \{ab^ka\} \subseteq A^+$$

and $\beta : B^* \to A^*$, where $\beta(c) = ab^k a$. It is clear that neither Y nor Z is a k-comma code, but $X = \{(ab^k a)^2\}$ is a k-comma code.

(ii) For X, Y, Z be codes over A with $X = Y \circ Z$, $k \geq 1$, if X is a k-comma code, then Y and Z may not be a comma-free code. For example, let $A = \{a, b\}$, $B = \{c\}$,

$$Y = \{c^2\} \subseteq B^+, \quad Z = \{ab^k ab^k\} \subseteq A^+$$

and $\beta : B^* \to A^*$, where $\beta(c) = ab^k ab^k$. It is easy to check that $X = \{(ab^k ab^k)^2\}$ is a k-comma code, neither Y nor Z is a comma-free code.

4. Results Related to Concatenation

In this section, we concentrate on the concatenation of k-comma codes. Let L_1, L_2 be two k-comma codes over A. Then we first give a characterization theorem $L_1 L_2$ to be a k-comma code under the assumption that $\{L_1, L_2\}$ is a partition of an infix code L.

For $\emptyset \neq L_1, L_2 \subseteq A^+$, we denote

$$\begin{aligned}
\bar{P}(L_1) &= \left(\bigcup_{x \in P(L_1)} L_1 x^{-1} \right) \setminus \{1\}, \\
\bar{S}(L_1) &= \left(\bigcup_{x \in S(L_1)} x^{-1} L_1 \right) \setminus \{1\}, \\
\bar{P}(L_2, L_1) &= \left(\bigcup_{x \in P(L_1)} L_2 x^{-1} \right) \setminus \{1\}, \\
\bar{S}(L_2, L_1) &= \left(\bigcup_{x \in S(L_2)} x^{-1} L_1 \right) \setminus \{1\}.
\end{aligned}$$

In the following theorems, we use $lg(L)$ to denote the minimum length of the words in some language L over A. Then, we consider the product of infix codes and obtain a characterization theorem of k-comma codes.

Theorem 4.1. *Let L be an infix code over A with $lg(L) > k$, $\{L_1, L_2\}$ a partition of L. Then $L_1 L_2$ is a k-comma code if and only if the following three conditions hold:*

(i) *For every $x \in \bar{S}(L_1)$, $y \in PS(L_2)$ with $xy \in L_2$, if there exists $z \in \bar{S}(L_2)$ such that $yz \in L_2$, then $|z| > k$ and $z \notin A^k P(L_1)$.*

(ii) *For every $x \in PS(L_1)$, $y \in \bar{P}(L_2)$ with $xy \in L_1$, if there exists $z \in \bar{P}(L_1)$ such that $zx \in L_1$, then $|z| > k$ and $z \notin S(L_2) A^k$.*

(iii) *For every $x \in \bar{S}(L_2, L_1)$, $y \in \bar{P}(L_2, L_1)$, one has $|xy| \neq k$.*

Proof. We first notice that L_1L_2 is not a k-comma code if and only if there exist $u_1, u_2, u_3 \in L_1,\ v_1, v_2, v_3 \in L_2,\ w \in A^k,\ r, s \in A^+$ such that $u_1v_1wu_2v_2 = ru_3v_3s$ if and only if one of the following three cases holds:

Case 1. $u_3 = z_1x$ for some $z_1 \in S(u_1)$, $v_3 = yz$, $v_1 = xy$, $w = zz_2$ for some $z_2 \in A^{<k}$ or $z = wz_2$ for some $z_2 \in P(u_2)$, see Figure 1.

Case 2. $v_3 = yz_2$ for some $z_2 \in P(v_2)$, $u_2 = xy$, $u_3 = zx$, $w = z_1z$ for some $z_1 \in A^{<k}$ or $z = z_1w$ for some $z_1 \in S(v_1)$, see Figure 2.

Case 3. $u_3 = z_1x$, $v_3 = yz_2$, $w = xy$ for some $z_1 \in S(v_1)$, $z_2 \in P(u_2)$, see Figure 3.

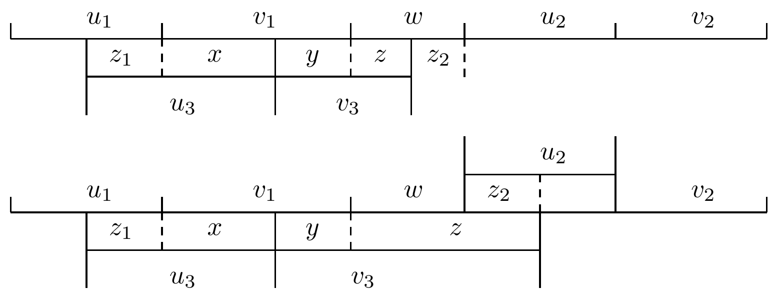

Figure 1: **Case 1**

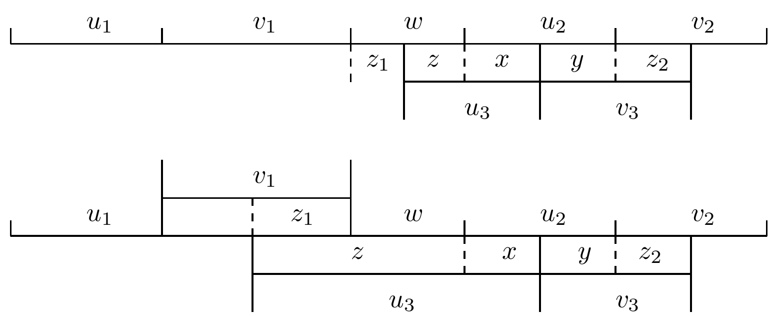

Figure 2: **Case 2**

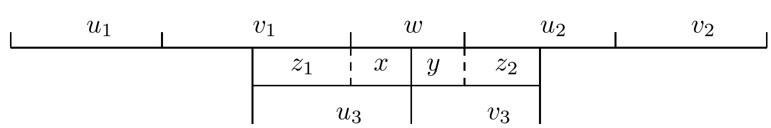

Figure 3: **Case 3**

These results contradict to the condition 1, 2 and 3 respectively. Therefore, the conditions 1, 2 and 3 hold if and only if L_1L_2 is a k-comma code. ■

By the proof of Theorem 4.1, it is easy to obtain an easier way to construct a k-comma code of the form L_1L_2 as follows:

Corollary 4.2. *Let L be an infix code over A with $lg(L) > k$, $\{L_1, L_2\}$ a partition of L. If L_1 and L_2 satisfy the following two conditions:*

(i) $\bar{P}(L_2) \cap \bar{S}(L_1) = \emptyset$,

(ii) *one of the following three statements is true:*

  (a) $P(L_1) \cap S(L_2) = \emptyset$,

  (b) $\bar{P}(L_2, L_1) = \emptyset$,

  (c) $\bar{S}(L_2, L_1) = \emptyset$,

then L_1L_2 is a k-comma code.

Proof. If L_1L_2 is not a k-comma code, by the proof of Theorem 4.1, then there are three cases. In case 1 and case 2, $x \in \bar{P}(L_2) \cap \bar{S}(L_1)$ and $y \in \bar{P}(L_2) \cap \bar{S}(L_1)$, respectively, a contradiction; in case 3, $\{z_1, z_2\} \subseteq P(L_1) \cap S(L_2)$, $x \in \bar{P}(L_2, L_1)$ and $y \in \bar{S}(L_2, L_1)$, also a contradiction. ■

5. Automata and Regular Languages

In this section, we first prove that the property of being a 1-1-comma code is decidable for regular languages which is based on the following two lemmas (Lemmas 5.7 and 5.8, respectively). In proving these two lemmas, we adopt the techniques which are similar to those arguments used for the proof of the well-known Pumping Lemma.

Then, we give a characterization of the automata which accept the words in X_1 (see Lemma 5.3). For the sake of convenience, we first state the following useful lemmas.

Lemma 5.1. [23] *If $ux = xy$, $u, x, y \in A^*$, $u \neq 1$, then $u = (\alpha\beta)^i$, $x = (\alpha\beta)^j\alpha$, $y = (\beta\alpha)^i$ for some $\alpha\beta \in Q$, $i, j \geq 1$.*

By Lemma 5.1, we are able to obtain that for any bordered primitive word y, y can be written to be $(\alpha\beta)^i\alpha$, $\alpha\beta \in Q$. Furthermore, we define the following three equalities:

$$\begin{aligned} NR^{>1} &= \{u \in A^+ | u = f^i, f \in Q,\ |f| > 1,\ i > 1\}, \\ Q_B^{>1} &= \{u \in Q | u \in (xy)^+x,\ xy \in Q,\ x, y \neq 1,\ |y| > 1\}, \\ Q_B^{=1} &= \{u \in Q | u \in (xy)^+x,\ xy \in Q,\ x, y \neq 1,\ |y| = 1\}. \end{aligned}$$

Lemma 5.2. [23] *Let $u \in A^+$. If u is bordered, then there exist $v \in A^+$ and $z \in A^*$ such that $u = vzv$.*

Lemma 5.3. [4] *Let L be a language over A and $k \geq 0$. Then L is a 1-k-comma code if and only if $L \in 2^{X_k} \backslash \emptyset$, where*

$$X_k = \left\{u \in A^+ | (\ \forall w \in A^k) uwu \cap A^+uA^+ = \emptyset\right\}.$$

Lemma 5.4. [4] $X_1 = (U \setminus A) \cap Q_B^{\geq 1} \cup NR^{>1}$.

Lemma 5.5. [4] $\left(Q_B^{=1} \cup \{a^i | a \in A,\ i \geq 1\}\right) \cap X_1 = \emptyset$.

Lemma 5.6. [4] $A^* = \left(Q_B^{=1} \cup \{a^i | a \in A,\ i \geq 1\}\right) \cup X_1$.

Now, we state the following two crucial lemmas.

Lemma 5.7. *Let $\mathcal{A} = (S, A, \delta, s_0, T)$ be a finite automaton with $|S| = n \geq 2$. Then $T(\mathcal{A}) \cap Q_B^{=1} \neq \emptyset$ if and only if there exists $y \in T(\mathcal{A}) \cap Q_B^{=1}$ with $|y| \leq n^{n+2} + n^{n+1}$.*

Proof. Here, we only prove the sufficiency part as the necessity part is obvious. Now we may assume that $T(\mathcal{A}) \cap Q_B^{=1} \neq \emptyset$. Take $y \in T(\mathcal{A}) \cap Q_B^{=1}$ such that

$$|y| = min\{|x| | x \in T(\mathcal{A}) \cap Q_B^{=1}\}.$$

Suppose that $|y| > n^{n+2} + n^{n+1}$. Let $y = (\alpha\beta)^i\alpha$, $\alpha\beta \in Q,\ i \geq 1,\ |\beta| = 1$. Then, we distinguish the following two cases.

Case 1. If $i \leq n$, $|y| > n^{n+2} + n^{n+1}$ implies that $|\alpha| \geq n^{i+1}$.

If $|\alpha| \geq n^{i+1}$, then α can be decomposed into $\alpha = \alpha_1\alpha_2\alpha_3,\ |\alpha_2| \geq 1$ with

$$\delta(s_0, (\alpha\beta)^j\alpha_1) = \delta(s_0, (\alpha\beta)^j\alpha_1\alpha_2),\ j = 0, 1, \cdots, i.$$

Let $\overline{\alpha} = \alpha_1\alpha_3$. Then, we can easily see that $\delta(s_0, (\overline{\alpha}\beta)^i\overline{\alpha}) = \delta(s_0, (\alpha\beta)^i\alpha)$. Thus, we have $(\overline{\alpha}\beta)^i\overline{\alpha} \in T(\mathcal{A})$. Notice that $(\overline{\alpha}\beta)^i\overline{\alpha} \in Q_B^{=1}$ and $|(\overline{\alpha}\beta)^i\overline{\alpha}| < |y|$, which contradicts to the definition of y.

Case 2. If $i > n$, there exist positive integers s, t, $1 \leq s < t \leq i$ such that

$$\delta(s_0, (\alpha\beta)^s) = \delta(s_0, (\alpha\beta)^t).$$

Thus

$$\delta(s_0, (\alpha\beta)^s(\alpha\beta)^{i-t}\alpha) = \delta(s_0, (\alpha\beta)^t(\alpha\beta)^{i-t}\alpha).$$

Hence $\delta(s_0, (\alpha\beta)^i\alpha) = \delta(s_0, (\alpha\beta)^{i-t+s}\alpha)$. Therefore $(\alpha\beta)^{i-t+s}\alpha \in T(\mathcal{A})$. It is clear that $(\alpha\beta)^{i-t+s}\alpha \in Q_B^{=1}$, which implies that $(\alpha\beta)^{i-t+s}\alpha \in T(\mathcal{A}) \cap Q_B^{=1}$. While $|(\alpha\beta)^{i-t+s}\alpha| < |y|$, it also contradicts to the definition of y. ∎

Lemma 5.8. *Let $\mathcal{A} = (S, A, \delta, s_0, T)$ be a finite automaton with $|S| = n \geq 2$. Then*

$$T(\mathcal{A}) \cap \{a^i | a \in A,\ i \geq 1\} \neq \emptyset$$

if and only if there is a letter a in A, $k < max\Big\{n, |A|\Big\}$ such that $a^k \in T(\mathcal{A})$.

Proof. The sufficient part is straightforward and we hence omit the proof. Now, we proceed to show the converse part.

Since A is finite, without loss of generality, we may assume that $T(\mathcal{A}) \cap \{a^i | a \in A,\ i \geq 2\} \neq \emptyset$. Take $y \in T(\mathcal{A}) \cap \{a^i | a \in A,\ i \geq 2\}$ with

$$|y| = min\Big\{|x|\Big|x \in \Big(T(\mathcal{A}) \cap \{a^i | a \in A,\ i \geq 2\}\Big)\Big\}.$$

If $y = a^i$, $a \in A$, $i \geq max\Big\{n, |A|\Big\}$, then $i \geq n$. The fact that $|Q| = n$ implies that there exist positive integers j, k, l such that

$$a^i = a^j a^k a^l \text{ and } \delta(s_0, a^j a^k a^l) = \delta(s_0, a^j (a^k)^* a^l).$$

Thus $a^{j+l} \in T(\mathcal{A})$. But $j + l < i$, this yields a contradiction. ∎

Therefore, for regular languages, we have the following theorem.

Theorem 5.9. *Let L be a regular language. It is decidable whether or not L is a* 1-1-*comma code.*

Proof. For a regular language L, by Lemma 5.3, 5.5, 5.6, 5.7 and 5.8, in order to decide whether L is a 1-1-comma codes, we need only to check whether the words of length less than $min\{n^{n+2} + n^{n+1}, |A|\}$ contain the words in $Q_B^{=1}$ or the power of a letter. ∎

Now, we are going to characterize the automata which accepts the words in X_1 (see Lemma 5.3). For this purpose, we need the following three lemmas:

Lemma 5.10. *Let $\mathcal{A} = (S, A, \delta, s_0, T)$ be a finite automaton with $|S| = n \geq 2$. Then,*

$$T(\mathcal{A}) \cap NR^{>1} \neq \emptyset$$

if and only if there exists $y \in T(\mathcal{A}) \cap NR^{>1}$ with

$$|y| \leq (n-1)(n+2)^{n-1} - n.$$

Proof. The sufficient part is trivial and we hence omit its proof. To prove the necessity part, we assume that $T(\mathcal{A}) \cap NR^{>1} \neq \emptyset$. Take $y \in T(\mathcal{A}) \cap NR^{>1}$ with

$$|y| = min\{|x||x \in T(\mathcal{A}) \cap NR^{>1}\}.$$

Suppose that $|y|$ exceeds the bound claimed in the lemma. Let $y = f^i,\ f \in Q,\ i > 1$. Then there are the following two cases:

Case 1. If $i \geq n$, then there exist two integers $p, q,\ 1 \leq p < q \leq i$ such that $\delta(s_0, f^p) = \delta(s_0, f^q)$, which yields that

$$\delta(s_0, f^{p+i-q}) = \delta(s_0, f^{q+i-q}) = \delta(s_0, f^i).$$

Hence, $f^{p+i-q} \in T(\mathcal{A})$, clearly, $f^{p+i-q} \in NR^{>1}$. Thus, we deduce

$$f^{p+i-q} \in T(\mathcal{A}) \cap NR^{>1}.$$

Notice that $p+i-q < i$. This fact implies that $|f^{p+i-q}| < |y|$, which contradicts to the minimality of y.

Case 2. If $i < n$, $|Q| = n$ and $|y| \geq (n-1)(n+2)^{n-1} - n + 1$ imply that $|f| \geq (n+2)^i$, then f has a decomposition

$$f = f_1 f_2 f_3,\ |f_1 f_3| > 1,\ |f_2| \geq 1,$$

such that

$$\delta(s_0, f^j f_1) = \delta(s_0, f^j f_1 f_2),\ j = 0, 1, \cdots, i-1. \tag{1}$$

Let $\bar{f} = f_1 f_3$. Then, for (1), we can obtain that $\delta(s_0, \bar{f}^i) = \delta(s_0, f^i)$, which implies that $\bar{f}^i \in T(\mathcal{A})$. Hence

$$\bar{f}^i \in T(\mathcal{A}) \cap NR^{>1}.$$

It is clear that $|\bar{f}| < |f|$, a contradiction. ■

Lemma 5.11. *Let $\mathcal{A} = (S, A, \delta, s_0, T)$ be a finite automaton with $|S| = n \geq 2$. Then, $T(\mathcal{A}) \cap Q_B^{>1} \neq \emptyset$ if and only if there exists $y \in T(\mathcal{A}) \cap Q_B^{>1}$ with*

$$|y| \leq n^{n+2} + n^{n+1} + n(n+2)^n - n - 1.$$

Proof. The proof of sufficient part is immediate. We now proceed to prove the necessity part. Assume that $T(\mathcal{A}) \cap Q_B^{>1} \neq \emptyset$. Take $y \in T(\mathcal{A}) \cap Q_B^{>1}$ with

$$|y| = min\{|x| | x \in T(\mathcal{A}) \cap Q_B^{>1}\}.$$

Suppose that $|y| > n^{n+2} + n^{n+1} + n(n+2)^n - n - 1$. Let $y = (\alpha\beta)^i\alpha,\ \alpha\beta \in Q,\ i \geq 1,\ |\beta| > 1$. Then, we divide our proof into the following two cases.

Case 1. If $i \leq n$, notice that $|y| > n^{n+2} + n^{n+1} + n(n+2)^n - n - 1$, then we have $|\alpha| \geq n^{i+1}$ or $|\beta| \geq (n+2)^i$.

If $|\alpha| \geq n^{i+1}$, then α has a decomposition $\alpha = \alpha_1\alpha_2\alpha_3$ with $|\alpha_2| \geq 1$ such that

$$\delta(s_0, (\alpha\beta)^j \alpha_1) = \delta(s_0, (\alpha\beta)^j \alpha_1\alpha_2),\ j = 0, 1, \cdots, i.$$

Let $\overline{\alpha} = \alpha_1\alpha_3$. Then, we have $\delta(s_0, (\overline{\alpha}\beta)^i\overline{\alpha}) = \delta(s_0, (\alpha\beta)^i\alpha)$. Hence, $(\overline{\alpha}\beta)^i\overline{\alpha} \in T(\mathcal{A})$. It is clear that $(\overline{\alpha}\beta)^i\overline{\alpha} \in Q_B^{\geq 1}$ and $|(\overline{\alpha}\beta)^i\overline{\alpha}| < |y|$, contradicting the minimality of y.

If $|\beta| \geq (n+2)^i$, then β has a decomposition

$$\beta = \beta_1\beta_2\beta_3,\ |\beta_2| > 1,\ |\beta_1\beta_3| > 1$$

and for all $j = 1, 2, \cdots, i-1$, we have

$$\begin{aligned} \delta(s_0, \alpha\beta_1) &= \delta(s_0, \alpha\beta_1\beta_2), \\ \delta(s_0, (\alpha\beta)^j\beta_1) &= \delta(s_0, (\alpha\beta)^j\beta_1\beta_2). \end{aligned} \tag{2}$$

Denote $\beta_1\beta_3 = \overline{\beta}$. By (2), we have $\delta(s_0, (\alpha\overline{\beta})^i\alpha) = \delta(s_0, (\alpha\beta)^i\alpha)$, thus $(\alpha\overline{\beta})^i\alpha \in T(\mathcal{A})$. Observe that $(\alpha\overline{\beta})^i\alpha \in Q_B^{\geq 1}$ and $|(\alpha\overline{\beta})^i| < |y|$, which contradicts to the choice of y.

Case 2. If $i > n$, then there exist two positive integers s, t, $1 \leq s < t \leq i$ such that

$$\delta(s_0, (\alpha\beta)^s) = \delta(s_0, (\alpha\beta)^t),$$

which implies that

$$\delta(s_0, (\alpha\beta)^s(\alpha\beta)^{i-t}\alpha) = \delta(s_0, (\alpha\beta)^t(\alpha\beta)^{i-t}\alpha).$$

The above two equalities yield that $\delta(s_0, (\alpha\beta)^{i-t+s}\alpha) = \delta(s_0, (\alpha\beta)^i\alpha)$. Hence, $(\alpha\beta)^{i-t+s}\alpha \in T(\mathcal{A})$. Notice that

$$(\alpha\beta)^{i-t+s}\alpha \in Q_B^{\geq 1}.$$

Hence, we have $(\alpha\beta)^{i-t+s}\alpha \in T(\mathcal{A}) \cap Q_B^{\geq 1}$. While $|(\alpha\beta)^{i-t+s}\alpha| < |y|$, a contradiction. ∎

Lemma 5.12. *Let $\mathcal{A} = (S, A, \delta, s_0, T)$ be a finite automaton with $|S| = n \geq 2$. Then,*

$$T(\mathcal{A}) \cap (U \setminus A) \neq \emptyset$$

if and only if there exists $y \in T(\mathcal{A}) \cap (U \setminus A)$ with $|y| < 4n$.

Proof. If there exists $y \in T(\mathcal{A}) \cap (U \setminus A)$ with $|y| < 4n$, the assertion is obviously correct.

Conversely, assume $T(\mathcal{A}) \cap (U \setminus A) \neq \emptyset$. Let $y \in T(\mathcal{A}) \cap (U \setminus A)$, and assume that $|y|$ is minimal with the property. Suppose that $|y| \geq 4n$. Then, y has a decomposition:

$$y = uvwx,\ |u| = |x| \geq n,\ n = |v| \leq |w| \leq n+1.$$

$|v| = n$ and $|Q| = n$ imply that v has a decomposition $v = v_1v_2v_3$ with $v_2 \in A^+$ such that

$$\delta(s_0, uv_1) = \delta(s_0, uv_1v_2).$$

Hence, we have $\delta(s_0, uv_1v_3wx) = \delta(s_0, uv_1v_2v_3wx) = \delta(s_0, y) \in T(\mathcal{A})$.

By the choice of y, we have $uv_1v_3wx (= y') \notin U \setminus A$. Since $y' \notin A$, y' is bordered. By Lemma 5.2, there exist $z \in A^+$ and $s \in A^*$ such that $y' = zsz$. Since $|v_1v_3| < |v| \le |w|$, $|z| < |wx|$. If $|z| \le |x| (= |u|)$, then $z \le_d y$, which contradicts to $y \in U \setminus A$. Then $|x| < |z| < |wx|$. Thus there exist $v', w' \in A^+$ such that $z = uv' = w'x,\ w' <_s w$. Of course, y also has another decomposition

$$y = u\bar{v}\bar{w}x,\ |u| = |x| \ge n,\ n \le |\bar{w}| \le |\bar{v}| \le n+1.$$

Similar to the above discussion, there exists $\bar{y}$ such that

$$\bar{y} = u\bar{v}\bar{w}_1\bar{w}_3x = \bar{z}t\bar{z} \in T(\mathcal{A}) \setminus (U \cap A)$$

with $|u| < |\bar{z}| < |u\bar{v}|$. Hence we have $\bar{z} = u\bar{v'} = \bar{w'}x,\ \bar{v'} <_p \bar{v}$. According to the lengths of z and $\bar{z}$, there are the following three cases:

Case 1. $|z| = |\bar{z}|$.

Since $z = uv'$ and $\bar{z} = u\bar{v'}$, $|v'| = |\bar{v'}|$. $w' <_s w$ implies that $|w'| \le n$. Notice that $z = uv' = w'x$, thus $v' <_s x$. Similarly, $|\bar{v'}| \le n$. Since $\bar{z} = u\bar{v'} = \bar{w'}x$ and $\bar{v'} <_s x$, $v' = \bar{v'}$, and hence $z = \bar{z}$. Thus $z = \bar{z} <_p u\bar{v} <_p y$ and $\bar{z} = z <_s wx <_s y$, which yields $z <_d y$, which contradicts to the fact that $y \in U \setminus A$.

Case 2. $|z| > |\bar{z}|$, as shown in Figure 4 below.

Since $z = uv'$ and $\bar{z} = u\bar{v'}$, $|v'| > |\bar{v'}|$. Notice that $|v'| = |w'| \le n \le |x|$, hence $|\bar{v'}| = |\bar{w'}| \le n \le |x|$. While $z = uv' = w'x$ and $\bar{z} = u\bar{v'} = \bar{w'}x$, then $v', \bar{v'} <_s x$. Thus there exists $r \in A^+$ such that $v' = r\bar{v'}$. Since uv' and $u\bar{v'}$

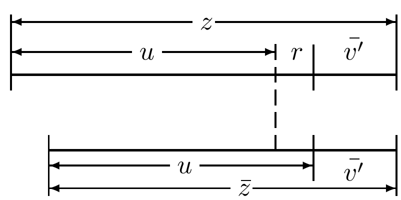

Figure 4: $|z| > |\bar{z}|$

have common suffix x, there exists $\bar{u}$ such that $ur = \bar{u}u$. $|r| < |x|$ implies that $u = r_2r^i,\ i \ge 1,\ r_1r_2 = r$. Hence

$$z = uv' = ur\bar{v'} = r_2r^ir\bar{v'} = (r_2r_1)(r_2r^i)\bar{v'} = (r_2r_1)\bar{z}.$$

Thus, $\bar{z} <_s z$. Since $\bar{z} <_p u\bar{v} <_p y$ and $z <_s wx <_s y$, $\bar{z} <_s y$ and $\bar{z} <_p y$, which yields that $\bar{z} <_d y$. It contradicts to $y \in U \setminus A$.

Case 3. $|z| < |\bar{z}|$, as shown in Figure 5.

Since $z = w'x$ and $\bar{z} = u\bar{v'} = \bar{w'}x$, $|\bar{w'}| > |w'|$. Notice that $z = uv' = w'x$ and $\bar{z} = u\bar{v'} = \bar{w'}x$ with $|w'| \le n \le |u|$ and $|\bar{w'}| \le n \le |u|$, we have that w' and $\bar{w'}$ have a common initial segment u. Then there exists $t \in A^+$ such that $\bar{w'} = w't$.

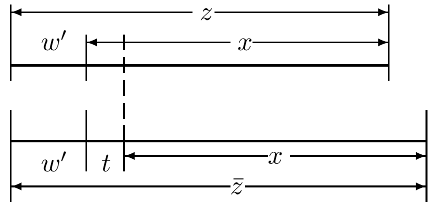

Figure 5: $|z| < |\bar{z}|$

Since z and $\bar{z}$ have common prefix u, there exists $\bar{x}$ such that $x\bar{x} = tx$. $|t| < |u|$ implies that $x = t^j t_1,\ j \geq 1,\ t_1 t_2 = t$. Then $\overline{z} = w'tx = w'tt^j t_1 = w't^j t_1(t_2 t_1) = w'x(t_2 t_1) = z(t_2 t_1)$. Hence, we obtain $z <_p \bar{z}$. Observe that $\bar{z} <_p u\bar{v} <_p y$ and $z <_s wx <_s y$. Thus, we derive that $z <_s y$ and $z <_p y$. Therefore, we obtain $z <_d y$, this contradicts to y is unbordered. ■

We now proceed to establish a theorem for finite automata.

Theorem 5.13. *Let $\mathcal{A} = (S, A, \delta, s_0, T)$ be a finite automaton with $|S| = n \geq 2$. Then $T(\mathcal{A}) \cap X_1 \neq \emptyset$ if and only if there exists $y \in T(\mathcal{A}) \cap X_1$ with*

$$|y| \leq n^{n+2} + n^{n+1} + n(n+2)^n - n - 1.$$

Proof. The proof of this theorem is a direct consequence of Lemmas 5.4, 5.10, 5.11 and 5.12 in this section. ■

As an application of the automata theory, an automaton is said to be *coaccessible* if for any $s \in Q$, $\delta(s, x) \in T$ for some $x \in A^*$, this means that start from any state, then we arrive a final state. For a coaccessible automaton, we now can get a theorem below which is completely different from Theorem 5.13.

Theorem 5.14. *Let $\mathcal{A} = \{S, A, \delta, s_0, T\}$ be a coaccessible automaton with $|A| \geq 2$, $|S| = n$, and $k \geq 0$. Then, the following two statements hold:*

(i) $|T(\mathcal{A}) \cap X_k| = \infty$,

(ii) *there is $y \in T(\mathcal{A}) \cap X_k$ with $2n + k + 1 \leq |y| \leq n + k + 2$.*

Proof. Take $a, b \in A$ with $a \neq b$. Since $\mathcal{A}$ is coaccessible, we have

$$(\forall i \geq 1)(\exists x_i \in A^*)|x_i| \leq k \ \&\ \delta(s_0, a^i b^{k+1} a x_i) \in T(\mathcal{A}).$$

If $i \neq j$, then $a^i b^{k+1} a x_i \neq a^j b^{k+1} a x_j$; If $i \geq n$, then, it is clear that $a^i b^k a x_i \in X_k$. Thus $\left\{a^i b^{k+1} a x_i | i \geq n\right\} \subseteq T(\mathcal{A}) \cap X_k$. Hence, we have $a^i b^{k+1} a x_1 \in T(\mathcal{A}) \cap X_k$ and $n + k + 2 \leq |a^{2k+1} b x_1| \leq 2n + k + 1$. Thus, our proof is completed. ■

References

[1] J. Berstel, D. Perrin, C. Reutenauer, *Codes and Automata*, Cambridge University Press, Cambridge, 2010.

[2] C.H. Cao, H.Y. Liu, D. Yang, Characterizations of k-comma codes and k-comma intercodes, *Acta Inf.* **53** (2016) 23–33.

[3] D.J. Cho, Y. Han, S.K. Ko, Decidability of involution hypercodes, *Theoret. Comput. Sci.* **550** (2014) 90–99.

[4] B. Cui, L. Kari, S. Seki, k-comma codes and their generalizations, *Fund. Inf.* **107** (2011) 1–18.

[5] H. Fernau, K. Reinhardt, L. Staiger, Decidability of code properties, *RAIRO-Inf. Theor. Appl.* **41** (3) (2007) 243–259.

[6] E.N. Gilbert, A comparison of signalling alphabets, *Bell System Technical Journal* **31** (3) (1952) 504–522.

[7] R.W. Hamming, Error detecting and error correcting codes, *Bell System Technical Journal* **20** (2) (1950) 147–160.

[8] Y.S. Han, K. Salomaa, D. Wood, Intercode regular languages, *Fund. Inf.* **76** (2007) 113–128.

[9] B. Hayes, The invention of the genetic code, *American Scientist* **86** (1998) 8–14.

[10] R. Hill, *A First Course in Coding Theory*, Oxford University press, New York, 1986.

[11] H. Hon and S. Han, A new construction and an efficient decoding method for Rabin like Codes, *IEEE Transaction* **66** (2) (2018) 521–538.

[12] J.E. Hopcroft and T.D. Ullman, *Formal Languages and Their Relation to Automata*, Addison-Wesley, Massachusetts, 1976.

[13] M. Ito, M. Katsura, H.J. Shyr, S.S. Yu, Automata accepting primitive words, *Semigroup Forum* **37** (1988) 45–52.

[14] M. Ito, H. Jürgensen, H.J. Shyr, G. Thierrin, n-Prefix-suffix languages, *Intern. J. Comput. Math.* **30** (1989) 37–56.

[15] H. Jürgensen, K. Salomaa, S. Yu, Decidability of the intercode property, *Elektronische Informationsverarbeitung und Kybernetik* **29** (6) (1993) 375–380.

[16] H. Jürgensen, K. Salomaa, S. Yu, Transducers and the decidability of independence in free monoids, *Theor. Comput. Sci.* **134** (1994) 107–117.

[17] H. Jürgensen, S. Konstantinidis, Codes, In: *Word, Language, Grammar*, Ed. by G. Rozenberg and A. Salomaa, Handbook of Formal Languages, Vol. **1**, Springer-Verlag, 1997.

[18] B. Lewin, *Genes IX*, Jones and Bartlett Publishers, 2007.

[19] S. Ling and C. Xing, *Coding Theory: A First Course*, Cambridge University Press, 2004.

[20] H.Y. Liu and Y.Q. Guo, A note on 1-(k, m)-comma codes, *J. Shandong University (Natural Sci.)* **54** (6) (2019) 11–15.

[21] H.Y. Liu, K. P. Shum, Y.Q. Guo, Some results on (k, m)-comma codes, *B. Iran. Math. Soc.* **46** (4) (2020) 1143–1162.

[22] C.E. Shannon, A mathematical theory of communication, I, II, *The Bell System Technical Journal* **27** (3) (1948) 379–423, 623–656.

[23] H.J. Shyr, *Free Monoids and Languages*, **3rd** Ed., Hon Min Book Company, Taichung, 2001.

[24] J. Watson, *Genes, Girls and Gamow: After the Double Helix*, Oxford University Press, Oxford, 2001.