Discrete Applied Mathematics 163 (2014) 334–342

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Average number of occurrences of repetitions in a necklace

Kazuhiko Kusano *, Ayumi Shinohara

Graduate School of Information Sciences, Tohoku University, Aramaki aza Aoba 6-3-09, Aoba-ku, Sendai-shi 980-8579, Japan

ARTICLE INFO

Article history:
Received 12 December 2011
Received in revised form 22 February 2013
Accepted 25 May 2013
Available online 22 June 2013

Keywords:
Repetition
Run
Combinatorics on words

ABSTRACT

In this paper, we consider the average number of occurrences of primitively rooted repetitions in a necklace. First, we define *circular repetitions* for a string and show the average number of them. Using these results, we can obtain the average number of squares, cubes, runs and cubic runs and the average sum of exponents of runs and cubic runs in a necklace, exactly.

1. Introduction

Repetitions are fundamental properties of strings. They can be applied to string processing or data compression. Especially, we are interested in a run (as known as maximal repetition), which is non-extendable repetition. Kolpakov and Kucherov [10,11] showed that the maximal number $\rho(n)$ of runs in a string of length n is $\rho(n) \leq cn$ for some constant c. Although the exact value of $\rho(n)$ is still unknown, it is conjectured [10] that $\rho(n) < n$. The current best upper bound is $\rho(n) < 1.029n$ due to Crochemore, Ilie, and Tinta [5,4]. On the other hand, there are several approaches [8,16] to show some lower bounds of $\rho(n)$ by constructing run-rich strings. The current best lower bound is $0.945n < \rho(n)$ by Matsubara et al. [15] and Simpson [18]. Concerning with the maximal sum $\sigma(n)$ of exponents of runs, Kolpakov and Kucherov [10,11] proved that $\sigma(n)$ is also linear, and conjectured that $\sigma(n) < 2n$. The conjecture was recently disproved by Crochemore et al. [6], who showed the current best bounds $2.035n < \sigma(n) < 4.1n$. Note that, Crochemore and Ilie [2] claimed that the upper bound could be improved to $2.9n$ employing computer experiments. Other results are summarized in [3].

A square is a repetition of exponent 2, and a cube is a repetition of exponent 3. We consider squares and cubes which are primitively rooted, and count these occurrences instead of distinct repetitions. Counting squares in this way, it is known that the maximal number of squares is $O(n \log n)$ [1].

Although the maximal number $\rho(n)$ of runs in a string of length n is unknown, the average number $R_s(n, \sigma)$ is exactly estimated by Puglisi and Simpson [17] as follows:

$$R_s(n, \sigma) = \sum_{p=1}^{\frac{n}{2}} \sigma^{-2p-1} ((n-2p+1)\sigma - (n-2p)) \sum_{d|p} \mu\left(\frac{p}{d}\right) \sigma^d,$$

where σ is the alphabet size and $\mu(n)$ is the Möbius function. The average number $S_s(n, \sigma)$ of squares and the average sum $E_s(n, \sigma)$ of exponents of runs in a string are presented by Kusano et al. [12]:

$$S_s(n, \sigma) = \sum_{p=1}^{\frac{n}{2}} \sigma^{-2p} (n-2p+1) \sum_{d|p} \mu\left(\frac{p}{d}\right) \sigma^d,$$

* Corresponding author. Tel.: +81 22 795 4535; fax: +81 22 795 4536.
E-mail addresses: kusano@shino.ecei.tohoku.ac.jp (K. Kusano), ayumi@ecei.tohoku.ac.jp (A. Shinohara).

0166-218X/$ – see front matter
http://dx.doi.org/10.1016/j.dam.2013.05.019

$$E_s(n,\sigma) = \sum_{p=1}^{\frac{n}{2}} \sigma^{-2p-1}\left(2(n-2p+1)\sigma - \left(2-\frac{1}{p}\right)(n-2p)\right)\sum_{d|p}\mu\left(\frac{p}{d}\right)\sigma^d.$$

In this paper, we focus on the average number of repetitions in a necklace, that is a word with its ends joined. In [15,18], in order to construct run-rich strings, authors considered repeated strings or necklaces. We define *circular repetitions* for strings, and show the average number of them, exactly.

In Section 2, we give some definitions and basic facts. Section 3 shows the average number of circular repetitions and the average sum of exponents of circular runs in a string. In Section 4, we derive the average numbers of squares, cubes, runs and cubic runs and the average sum of exponents of runs and cubic runs in a necklace.

A preliminary version of this article appeared in [13].

2. Preliminary

Let $\Sigma_\sigma = \{a_1, a_2, \ldots, a_\sigma\}$ be an alphabet of size σ. We often abbreviate Σ_σ by Σ, and denote the letters in Σ by $\mathtt{a}$, $\mathtt{b}$ and $\mathtt{c}$ for illustration purposes. We denote by Σ^n the set of all strings of length n over Σ, and $|w|$ denotes the length of a string w. We denote by $w[i]$ the ith letter of w, and $w[i\ldots j]$ is a substring $w[i]w[i+1]\cdots w[j]$ of w.

A *necklace* is a word which can be obtained by joining the ends of a string. We denote by $\langle w\rangle$ the necklace obtained from w. The length of $\langle w\rangle$ is $|w|$ and denoted by $|\langle w\rangle|$. Necklaces are equivalent under rotation, but not reversion. For example, $\langle\mathtt{abc}\rangle = \langle\mathtt{bca}\rangle = \langle\mathtt{cab}\rangle$, while $\langle\mathtt{abc}\rangle \neq \langle\mathtt{acb}\rangle$. Let NL_σ^n be the set of all necklaces of length n over Σ_σ. We say that d is a *divisor* of n and write $d|n$ if n/d is a positive integer. It is known that the number of necklaces is as follows.

Lemma 1 ([9]). $\left|NL_\sigma^n\right| = \frac{1}{n}\sum_{d|n}\phi\left(\frac{n}{d}\right)\sigma^d$.

Here, $\phi(n)$ is the Euler's phi function, which is defined to be the number of integers less or equal to n which are co-prime to n and can be written as $\phi(n) = \sum_{d|n}\mu\left(\frac{n}{d}\right)d$.

For a string w of length n and a positive integer $p \le n$, we say that p is a *period* of w if and only if $w[i] = w[i+p]$ holds for any $1 \le i \le n-p$. We denote by *period* (w) the set of all periods of w. For instance, *period* $(\mathtt{abaaba}) = \{3, 5, 6\}$.

Lemma 2 ([7]). *If* $p, q \in$ *period* (w) *and* $|w| \ge p + q - \gcd(p,q)$, *then* $\gcd(p,q) \in$ *period* (w).

A string w is *primitive* if w cannot be written as $w = u^k$ by any string u and any integer $k \ge 2$. We denote by $Prim_\sigma^n$ the set of all primitive strings in Σ_σ^n. It is known that the number of primitive strings can be represented using the Möbius function.

Lemma 3 ([14]). $\left|Prim_\sigma^n\right| = \sum_{d|n}\mu\left(\frac{n}{d}\right)\sigma^d$.

A substring $w[i\ldots j]$ of w is called a *repetition* in w if the smallest period p of $w[i\ldots j]$ satisfies $p \le (j-i+1)/2$. We denote the repetition by a triplet $\langle i, j, p\rangle$. The *root* of the repetition is the string $w[i\ldots i+p-1]$, and the *exponent* of the repetition is the value $(j-i+1)/p$. By definition, the exponent is at least 2, and the root is always primitive.

An *integer repetition* is a repetition whose exponent is an integer. Especially, a *square* (resp. *cube*) is an integer repetition whose exponent is exactly 2 (resp. 3). A repetition $\langle i, j, p\rangle$ in w is called a *run* if it satisfies the following two conditions: (1) either $i = 1$ or $w[i-1] \neq w[i+p-1]$, and (2) either $j = n$ or $w[j+1] \neq w[j-p+1]$. That is, *run* is a maximal repetition which is extendable neither to the left nor to the right. A *cubic run* is a run whose exponent is at least 3. For example, squares in a string $w = \mathtt{aaabab}$ are $\langle 1,2,1\rangle$, $\langle 2,3,1\rangle$ and $\langle 3,6,2\rangle$, and runs in w are $\langle 1,3,1\rangle$ and $\langle 3,6,2\rangle$. The repetition $\langle 1,3,1\rangle$ is a cube and a cubic run in w.

Concerning with repetitions in a necklace, ambiguity arises. For instance, a necklace $\langle\mathtt{abc}\rangle$ contains a repetition $\ldots\mathtt{abcabcabc}\ldots$, but which part is the maximal one? In order to avoid this ambiguity, we regard the circular repetitions in an infinite string w^ω as repetitions in a necklace $\langle w\rangle$, that are defined as follows.

For a string w of length n, we denote by w^ω the infinite string obtained by repeating w both to the left and right, and we define $w^\omega[i] = w[i\%n]$, where $i\%n$ is the positive integer z satisfying $1 \le z \le n$ and $z \equiv i \pmod n$ for any (possibly negative) integer i.

In this paper, we define *circular squares* in a string w as squares in w^ω which start at between 1 and $|w|$. We also define *circular cubes*, *circular runs* and *circular cubic runs* in the same way. Collectively, we call these repetitions *circular repetitions*. For example, circular squares in $w = \mathtt{aabab}$ are $\langle 1,2,1\rangle$, $\langle 2,5,2\rangle$, $\langle 3,6,2\rangle$, $\langle 4,9,3\rangle$, $\langle 1,10,5\rangle$, $\langle 2,11,5\rangle$, $\langle 3,12,5\rangle$, $\langle 4,13,5\rangle$ and $\langle 5,14,5\rangle$. Moreover, circular runs in w are $\langle 1,2,1\rangle$, $\langle 2,6,2\rangle$ and $\langle 4,9,3\rangle$ (see Fig. 1). For a string w and an integer $k \ge 2$, we denote by *cirep* (w,k) the number of circular integer repetitions of exponent k in w. Moreover, we denote by *crun* (w,k) (resp. *cexp* (w,k)) the number (resp. the sum of exponents) of circular runs of exponent at least k in w. For instance, for $w = \mathtt{aabab}$, *cirep* $(w,2) = 4$, *crun* $(w,2) = 3$, and *cexp* $(w,2) = 2/1 + 5/2 + 6/3 = 13/2$. Note that, the length of circular repetitions in a string of length n can be longer than n. For example, a string $\mathtt{abaab}$ of length 5 contains a circular square $\langle 1,6,3\rangle$ of length 6.

For a necklace $\langle w\rangle$, we define the number *irep* $(\langle w\rangle, k)$ of integer repetitions of exponent k, the number *run* $(\langle w\rangle, k)$ of runs, and the sum *exp* $(\langle w\rangle, k)$ of exponents of runs as follows:

$$irep(\langle w\rangle, k) = cirep(w,k), \qquad run(\langle w\rangle, k) = crun(w,k), \qquad exp(\langle w\rangle, k) = cexp(w,k).$$