Average Number of Runs and Squares in Necklace

Kazuhiko Kusano* and Ayumi Shinohara

Graduate School of Information Science, Tohoku University,
Aramaki aza Aoba 6-6-05, Aoba-ku, Sendai-shi 980-8579, Japan
{kusano@shino., ayumi@}ecei.tohoku.ac.jp

Abstract. A repetition is an important property of a string. In this paper we consider the average number of occurrences of primitively rooted repetitions in necklace. First, we define *circular square* and *circular run* for a string and show the average number of them. Using these results, we obtain the average number of squares, the average number of runs and the average sum of exponents of runs in a necklace, exactly.

Keywords: repetition, run, combinatorics on words

1 Introduction

A repetition is a fundamental property of a string. It can be applied to string processing or data compression. We are interested in run (as known as maximal repetition), which is non-extendable repetition. Kolpakov and Kucherov showed that the maximal number $\rho(n)$ of runs in a string of length n is $\rho(n) \leq cn$ for some constant c [6]. The exact value of $\rho(n)$ is still unknown and it is conjectured that $\rho(n) < 1$. The current best upper bound is $\rho(n) < 1.029n$ [3,4]. On the other hand, there are approaches to show the lower bounds of $\rho(n)$ constructing run-rich strings. The best lower bound is $\rho(n) > 0.945$ [9,11]. A repetition count of a run is called an exponent. It is proved that the maximal sum of exponents is also linear and the current best upper bound is $2.9n$ [2]. It is conjectured that the sum is less than $2n$ [7].

A square is a substring of the form u^2. We consider the primitively rooted square and count occurrences of squares instead of distinct squares. Counting squares in this way, it is known that the maximal number of squares is $O(n \log n)$ [1].

Although the maximal number of runs is unknown, the average number of runs in a string of length n is shown exactly as follows [10]:

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

where σ is alphabet size and $\mu(n)$ is the Möbius function. The average number of squares and the average sum of exponents of runs are also presented [8]:

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

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

* Supported in part by Grant-in-Aid for JSPS Fellows

Kazuhiko Kusano, Ayumi Shinohara: Average Number of Runs and Squares in Necklace, pp. 167–177 .
Proceedings of PSC 2010, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-04597-8

In [9,11], to construct run-rich strings they considered repeated strings or necklace. Therefore we focus on the average number of repetitions in necklace. To obtain the average number of repetitions in necklace we define *circular square* and *circular run* for string and show the average number of them, exactly.

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

2 Preliminary

Let $\Sigma = \{\mathtt{a}, \mathtt{b}, \dots\}$ be an alphabet of size σ. We denote the set of all strings of length n on Σ by Σ^n and the length of a string w by $|w|$. For a string $w = xyz$, strings x, y and z are called *prefix*, *substring* and *suffix* of w, respectively. We denote ith letter of a string w by $w[i]$ and a substring $w[i]w[i+1]\dots w[j]$ of w by $w[i..j]$.

A *necklace* is a word which can be obtained by joining the ends of a string. We denote a necklace of a string w by $\langle w \rangle$.

For a string w of length n and positive integer $p < n$, we say that w has a period p if and only if $w[i] = w[i+p]$ holds for any i, $1 \le i \le n-p$. We denote the set of periods of w by $period\,(w)$. For periods of strings, the following lemma is known [5].

Lemma 1. *Let p and q be periods of a string w. If $|w| \ge p + q - \gcd(p,q)$, w has also period $\gcd(p,q)$.*

A string w is *primitive* if w can not be written as $w = u^k$ by string u and integer $k \ge 2$.

We call a substring $w[i..j]$ a *repetition* if $w[i..j]$ has the smallest period $p \le \frac{j-i+1}{2}$ and denote the substring by triplet $\langle i, j, p \rangle$. We say that $w[i..p]$ is the *root* of the repetition. By Lemma 1, the root of a repetition is primitive. The *exponent* of the repetition is $\frac{j-i+1}{p}$.

A *square* is a repetition whose exponent is exactly 2. We consider only squares which have a primitive root. A *run* is a repetition which has non-extendability, that is, a run $\langle i, j, p \rangle$ in w satisfies the following two conditions:

$$\begin{aligned} i = 1 \quad &\text{or} \quad w[i-1] \neq w[i+p-1], \\ j = n \quad &\text{or} \quad w[j+1] \neq w[j-p+1]. \end{aligned}$$

We denote a string of infinite length, obtained by repeating string w to both left and right, by w^ω. For a string w of length n and integer i, $w^\omega[i] = w[i\%n]$, where the operator $x\%y$ represents a number z such that $1 \le z \le y$ and $z \equiv x \pmod{y}$. In this paper, we define a *circular run* (*circular square*, resp.) for a string w as a run (square, resp.) in w^ω and which starts between 1 and $|w|$. We denote the number of circular squares by $csqr\,(w)$, the number of circular runs in a string w by $crun\,(w)$ and the sum of exponents of runs by $cexp\,(w)$. For a necklace $\langle w \rangle$, we define number of runs $run\,(\langle w \rangle)$, number of squares $sqr\,(\langle w \rangle)$ and sum of exponents of runs $sqr\,(\langle w \rangle)$ as follows:

$$\begin{aligned} run\,(\langle w \rangle) &= crun\,(w)\,, \\ sqr\,(\langle w \rangle) &= csqr\,(w)\,, \\ exp\,(\langle w \rangle) &= cexp\,(w)\,. \end{aligned}$$

3 Average number of circular repetitions in a string

For a string of length n and alphabet size σ, the average number of circular squares, the average number of circular runs and the average sum of exponents of circular runs are defined as:

$$S_c(n,\sigma) = \frac{1}{|\Sigma^n|}\sum_{w\in\Sigma^n} csqr(w),$$
$$R_c(n,\sigma) = \frac{1}{|\Sigma^n|}\sum_{w\in\Sigma^n} crun(w),$$
$$E_c(n,\sigma) = \frac{1}{|\Sigma^n|}\sum_{w\in\Sigma^n} cexp(w).$$

3.1 Average number of circular squares

To obtain these values, we count repetitions in all strings of length n. We consider repetitions classified according to their position and period. For the position, it is sufficient to consider only repetitions at one position. The total number of occurrences can be obtained as the product of this number and the length of strings.

Lemma 2. *For a string f and integer i, let $\Sigma_{f,i}$ be the set of string w of length n such that w^ω contains f at i. For any integer i and j, $|\Sigma_{f,i}| = |\Sigma_{f,j}|$.*

Proof. We may assume without loss of generality that $i \le j$. If w is an element of $\Sigma_{f,i}$, then $w[i\%n]w[(i+1)\%n]\dots w[(i+|f|)\%n] = f$. Let $w' = w[n-(j-i)+1..n]w[1..j-i]$. Since w' satisfies the condition $w'[j\%n]w'[(j+1)\%n]\dots w'[(i+|f|)\%n] = f$, w' is in $\Sigma_{f,j}$.

Although the circular repetition is defined as the repetition in an infinity string, the period of the primitive rooted repetition is not so long.

Lemma 3. *Let w be the string of length n. The period of circular square in w is at most n.*

Proof. Let $\langle i,j,p\rangle$ be the circular run in w. If we suppose that $p > n$, the substring $w^\omega[i..j]$ of length $2p$ has two periods n and p. By Lemma 1, it also has period $\gcd(n,p)$, which is less than p and the divisor of p. So the primitive root $w^\omega[i..i+p-1]$ can be written as $w^\omega[i..i+p-1] = u^k$ using a string u and integer $k = \frac{p}{q} > 1$, a contradiction.

The length of the circular square in string of length n can be longer than n. For example, the string `abaab` of length 5 contains the circular square $\langle 1,6,3\rangle$ of length 6.

We consider the number of circular squares in all strings of length n at the position 1. Let $S_{f1}(p,\sigma)$ be the set of squares of period p and alphabet size σ; that is,

$$S_{f1}(p,\sigma) = \{vv : v \in Prim_{p,\sigma}\}.$$

Since a string w^ω may contain at most one elements of $S_{f1}(p,\sigma)$ at the position 1, the number of circular squares of period p in Σ^n equals to the number of the strings w such that w^ω contains the element of $S_{f1}(p,\sigma)$ at the position 1. More generally, we consider the set of strings of length l, instead of $S_{f1}(p,\Sigma)$.

Lemma 4. *Let F be a subset of Σ^l. For the number $N_F(n,\sigma)$ of strings w such that $|w| = n$ and $w^\omega[1..l] \in F$,*

$$N_F(n,\sigma) = \begin{cases} |F|\sigma^{n-l} & \text{if } l \le n, \\ |G| & \text{if } l > n, \end{cases}$$

where the set G is a subset of F and whose elements have the period n; that is

$$G = \{w \in F : n \in period(w)\}.$$

Proof.

1. Case $l \le n$
   The string w^ω contains the element of F at the position 1 if and only if $w[1..l] \in F$. Let C be the set of such strings. We have
   $$C = \{uv : u \in F, v \in \Sigma^{n-l}\}.$$
   The number of elements of C is $|F|\,\sigma^{n-l}$.
2. Case $l > n$
   Since w^ω has the period n, the substring $w^\omega[1..l]$ also has period n. For the element g of G, the suffix $g[n+1..l]$ is the repetition of $g[1..n]$. Therefore, the prefixes of length n of the elements of G are different. So, $N_F(n,\sigma) = |G|$.

Let $S_{f2}(n,p,\sigma)$ be the set of squares of length p and alphabet size σ and which can be contained in a repetition of string of length n; that is

$$S_{f2}(n,p,\sigma) = \{w : w \in S_{f1}(p,\sigma), n \in period\,(w)\}.$$

To obtain the size of $S_{f2}(n,b,\sigma)$, for integer d of divisor of p, we define $S_{f3}(n,p,d,\sigma)$ and $S_{f4}(n,p,d,\sigma)$ as follows:

$$S_{f3}(n,p,d,\sigma) = \left\{u^{\frac{2p}{d}} : u \in Prim_{d,\sigma}, n \in period\left(u^{\frac{2p}{d}}\right)\right\},$$
$$S_{f4}(n,p,d,\sigma) = \left\{u^{\frac{2p}{d}} : u \in \Sigma^d, n \in period\left(u^{\frac{2p}{d}}\right)\right\}.$$

We see that $S_{f2}(n,p,\sigma) = S_{f3}(n,p,p,\sigma)$. Since any string can be written uniquely as an integer power of a primitive string, $S_{f4}(n,p,d,\sigma) = \bigcup_{d|p} S_{f3}(n,p,d,\sigma)$.

First, we consider $S_{f4}(n,p,d,\sigma)$.

Lemma 5. *For the element $u^{\frac{2p}{d}}$ of $S_{f4}(n,p,d,\sigma)$, $u^{\frac{p}{d}}$ has a period $n-p$.*

Proof. Let $v = u^{\frac{p}{d}}$. By the definition of $S_{f4}(n,p,d,\sigma)$, for any position $1 \le i \le 2p-n$, $v^2[i] = v^2[i+n]$. For any position $1 \le j \le p-(n-p)$, $v[j] = v^2[j] = v^2[j+n] = v[j+n-p]$.

For $u^{\frac{2p}{d}} \in S_{f4}(n,p,d,\sigma)$, the string $u^{\frac{p}{d}}$ of length p has two periods d and $n-p$. If $d+(n-p) \le p$ such that $d \le 2p-n$, from lemma 1, $u^{\frac{p}{d}}$ also has a period $\gcd(d,n-p)$. In the other case, $u^{\frac{p}{d}}$ has another period.

Lemma 6. *For the element $u^{\frac{2p}{d}}$ of $S_{f4}(n,p,d,\sigma)$, if $d > 2p-n$, $u^{\frac{p}{d}}$ has a period $d-(2p-n)$.*

Proof. For any position $1 \leq i \leq d-(d-(2p-n))$, $u[i] = u^{\frac{2p}{d}}[i] = u^{\frac{2p}{d}}[i+n] = u[i+n-(2p-d)] = u[i+d-(2p-n)]$.

Therefore,

$$S_{f4}(n,p,d,\sigma) = \begin{cases} \{s^{\frac{2p}{\gcd(d,n-p)}} : s \in \Sigma^{\gcd(d,n-p)}\} & \text{if } d \leq 2p-n, \\ \{s^{\frac{2p}{d-2p+n}} : s \in \Sigma^{d-2p+n}\} & \text{if } d > 2p-n. \end{cases}$$

The number of elements of $S_{f4}(n,p,d,\sigma)$ can be written as

$$|S_{f4}(n,p,d,\sigma)| = \delta_s(n,p,d,\sigma),$$

where

$$\delta_s(n,p,d,\sigma) = \begin{cases} \sigma^{\gcd(d,n-p)} & \text{if } d \leq 2p-n, \\ \sigma^{d-2p+n} & \text{if } d > 2p-n. \end{cases}$$

Lemma 7. *The number of elements of $S_{f2}(n,p,\sigma)$ is as follows:*

$$|S_{f2}(n,p,\sigma)| = \sum_{d|p} \mu\left(\frac{p}{d}\right) \delta_s(n,p,d,\sigma).$$

Proof. Since

$$S_{f4}(n,p,p,\sigma) = \bigcup_{d|p} S_{f3}(n,p,d,\sigma),$$

we see that

$$|S_{f4}(n,p,p,\sigma)| = \sum_{d|p} |S_{f3}(n,p,d,\sigma)|.$$

Applying the Möbius inversion formula to this equation we have that

$$\begin{aligned} |S_{f2}(n,p,\sigma)| &= |S_{f3}(n,p,p,)| \\ &= \sum_{d|p} \mu\left(\frac{p}{d}\right) |S_{f4}(n,p,d,\sigma)| \\ &= \sum_{d|p} \mu\left(\frac{p}{d}\right) \delta_s(n,p,d,\sigma). \end{aligned}$$

By Lemma 4 and 7 we can derive the following theorem.

Theorem 8. *For any positive integer n and σ, the average number of circular squares in a string of length n and alphabet size σ is*

$$S_c(n,\sigma) = \frac{n}{\sigma^n} \sum_{p=1}^{n} \sum_{d|p} \mu\left(\frac{p}{d}\right) \delta_s(n,p,d,\sigma).$$

3.2 Average number of circular runs

In this subsection, we show the average number of circular runs in string of length n and alphabet size σ.

Unlike circular squares, whether a substring is a circular run or not depends on the characters next to the substring. For example, the repetition $\langle 2, 5, 2\rangle$ is a run in `aabab`, while the repetition is not a run in `babab` since the repetition can be extended to left. Instead of the set $S_{f1}(n,p)\sigma$, we consider a set $R_{f1}(n,p)\sigma$ of string such that

$$R_{f1}(p,\sigma) = \{cvv : c \neq v[p], v \in Prim_{p,\sigma}\}.$$

There is a circular run in w at the position i if and only if w^ω contains an element of $R_{f1}(n,p)\sigma$ at the position $i-1$.

The Lemma 2 can be applied to the element of $R_{f1}(p,\sigma)$. From Lemma 3 the period of a circular run in a string of length n does not exceed n, since a circular run contains at least one circular square. We consider the number of occurrences of the element of $R_{f1}(p,\sigma)$ in w^ω for all strings w of length n.

We define, for d of divisor of p, $R_{f2}(n,p,\sigma)$, $R_{f3}(n,p,d,\sigma)$ and $R_{f4}(n,p,d,\sigma)$ as follows:

$$\begin{aligned}
R_{f2}(n,p,\sigma) &= \{w \in S_{f1}(p,\sigma) : n \in period\,(w)\},\\
R_{f3}(n,p,d,\sigma) &= \left\{cu^{\frac{2p}{d}} : c \neq u[d], u \in Prim_{d,\sigma}, n \in period\left(cu^{\frac{2p}{d}}\right)\right\},\\
R_{f4}(n,p,d,\sigma) &= \left\{cu^{\frac{2p}{d}} : c \neq u[d], u \in \Sigma^d, n \in period\left(cu^{\frac{2p}{d}}\right)\right\}.
\end{aligned}$$

Since d is a divisor of p, we see that $u^{\frac{p}{d}} = u[d]$.

The Lemma 5 and 6 also hold for $R_{f4}(n,p,d,\sigma)$. The condition $c \neq u[d]$ sometimes makes $R_{f4}(n,p,d,\sigma)$ be empty.

Lemma 9. *If either $d \leq 2p-n$ or $d \equiv 0 \pmod{d-(2p-n)}$, the set $R_{f4}(n,p,d,\sigma)$ is empty.*

Proof. For the element $cu^{2p}d \in R_{f4}(n,p,d,\sigma)$, $u^{\frac{p}{d}}$ has the period $n-p$ and d. If $d \leq 2p-n$ that is $d+(n-p) \leq p$, from Lemma 1, $u^{\frac{p}{d}}$ also has period $t = gcd(d, n-p)$. In this case, $c = u^{\frac{p}{d}}[n-p] = u[d]$ and the condition $c \neq u[d]$ does not hold.

For the case $d > 2p-n$, $c = u^{\frac{2p}{d}}[n] = u[n-(2p-d)] = u[d-(2p-n)]$. Lemma 6 says that u has the period $d-(2p-n)$ such that $u[d] = u[d\%(d-(2p-n))]$. If $d \equiv 0 \pmod{d-(2p-n)}$, we have that $d-(2p-n) = d\%(d-(2p-n))$ and $c = u[d-(2p-n)] = [d\%(d-(2p-n))] = u[d]$.

For the case $d > 2p-n$ and $d \not\equiv 0 \pmod{d-(2p-n)}$, the set $R_{f4}(n,p,d,\sigma)$ can be written as:

$$R_{f4}(n,p,d,\sigma) = \left\{cs^{\frac{2p}{d-2p+n}} : c \neq s[d-2p+n], s \in \Sigma^{d-2p+n}\right\}.$$

Therefore,

$$|R_{f4}(n,p,d,\sigma)| = \delta_r\,(n,p,d,\sigma)\,,$$

where

$$\delta_r\,(n,p,d,\sigma) = \begin{cases} (\sigma-1)\sigma^{d-2p+n-1} & \text{if } d > 2p-n \text{ and } d \not\equiv 0 \pmod{d-(2p-n)},\\ 0 & \text{otherwise.}\end{cases}$$

We can derive $|R_{f2}(n,p,\sigma)|$ as follows:

$$\begin{aligned}|R_{f2}(n,p,\sigma)| &= |R_{f3}(n,p,p,)\,| \\ &= \sum_{d|p} \mu\left(\frac{p}{d}\right) |R_{f4}(n,p,d,\sigma)| \\ &= \sum_{d|p} \mu\left(\frac{p}{d}\right) \delta_r\left(n,p,d,\sigma\right).\end{aligned}$$

Theorem 10. *For any positive integers n and σ, the average number of circular runs in a string of length n and alphabet size σ is*

$$R_c\left(n,\sigma\right) = \frac{n}{\sigma^n} \sum_{p=1}^{n} \sum_{d|p} \mu\left(\frac{p}{d}\right) \delta_r\left(n,p,d,\sigma\right).$$

3.3 Average sum of exponents of circular runs

A circular run contains circular squares of same period. For example, for string $w =$ `abaabaab`, a circular run $\langle 1,8,3\rangle$ contains three circular squares $\langle 1,6,3\rangle$, $\langle 2,7,3\rangle$ and $\langle 3,8,3\rangle$. The number of circular squares depends on period and exponent of the run.

Lemma 11. *A circular run of period p and exponent e contains $(e-2)p+1$ circular squares of period p.*

Proof. Let $\langle i,j,p\rangle$ be a circular run in string w. For any position $i \le k \le j-p$, $w[k] = w[k+p]$ and a substring $w[k..k+p]$ is primitive since $w[k..k+p]$ is a conjugate of primitive string $w[i..i+p]$. The number of circular squares contained the run is $j-2p-i+2$. The exponent of the run is $e = \frac{j-i+1}{p}$. The number of circular squares can be written as $(e-2)p+1$.

Although any circular run contains circular squares, some circular squares are not contained in a run of the same period. For example, for a string $w =$ `abc`, there are circular runs $\langle 1,6,3\rangle$, $\langle 2,7,3\rangle$ and $\langle 3,8,3\rangle$ and there are no circular run containing the squares. For a string $w =$ `abab`, such circular squares are $\langle 1,4,2\rangle$, $\langle 2,5,2\rangle$, $\langle 3,6,2\rangle$ and $\langle 4,7,2\rangle$.

Lemma 12. *For a primitive string u of length p and integer k, u^k contains n circular squares of period p which is not contained in a circular run of period p.*

Proof. Let $w = u^k$. For any position i, $\langle i, i+2p-1, p\rangle$ is a circular square, since $w[i..i+p-1] = w[i+p..i+2p-1]$ and $w[i..i+p-1]$ is primitive by the definition of w. There is n circular squares of length p. There is no circular run of period p in w, since, for any position i, $w^\omega[i] = w^\omega[i+p]$ and a repetition of period p cannot satisfy non-extendability.

By contradiction, we show that there is no circular square of period $p' \neq p$ which contained in a circular run of period p'. Assume that w contains a circular square $\langle i,j,p'\rangle$. If there is no circular run of period p' containing $\langle i,j,p'\rangle$, the repetition $\langle i,j,p'\rangle$ can extend to both left and right infinitely. It mean that w^ω has the period p'. From Lemma 1 w^ω also has a period $t = \gcd(p',p)$. The period p' is not multiple of p since p' is the period of a circular square. The period t is less than p and a divisor of p, a contradiction.

From Lemma 11 and 12, we can derive the average sum of exponents of circular runs.

Theorem 13. *For positive integers n and σ, the average sum of exponents of circular runs in a string of length n and alphabet size σ is*

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

Proof. Consider a string w of length n. A string w can be uniquely written as $w = u^k$ where u is primitive string and k is integer. Let $csqr_p(w)$, $crun_p(w)$ and $cexp_p(w)$ be the number of circular squares of period p in w, the number of circular runs of period p in w and the sum of exponents of circular runs of period p in w, respectively. Applying Lemma 11 for each circular runs in w we have

$$csqr_p(w) - n[k=p] = (cexp_p(w) - 2crun_p(w))p + crun_p(w)$$
$$cexp_p(w) = \frac{1}{p}csqr_p(w) + \left(2-\frac{1}{p}\right)crun_p(w) - \frac{1}{p}n[k=p],$$

where $[k=p]$ is defined as 1 if $k=p$ and 0 if $k \neq p$. The number of circular squares not to be contained circular run of the same period is $n[k=p]$. Summing them up for each strings and each periods, from Theorem 8 and 10, we can obtain $E_c(n,\sigma)$. The number of strings of length n which can be written as $w = u^k$ equals to the number of primitive strings of length $p = \frac{n}{k}$. It is known that the number is $\sum_{d|p}\mu\left(\frac{p}{d}\right)\sigma^d$.

4 Average number of repetitions in necklace

Although we defined the number of repetitions in a necklace $\langle w\rangle$ equals to the number of circular repetitions in the string w, the average number of repetitions in a necklace of length n and the average number of circular repetitions in a string of length n are different.

Example 14. Let length $n = 4$ and alphabet size $\sigma = 2$. All strings of length n and the numbers of circular runs they contain are as follows:

aaaa 0	aaab 1	aaba 1	aabb 2	abaa 1	abab 0	abba 2	abbb 1
baaa 1	baab 2	baba 0	babb 1	bbaa 2	bbab 1	bbba 1	bbbb 0

Thus, the average number of circular runs in string is $\frac{16}{16} = 1$.

All necklaces of length n and the numbers of runs they contain are as follows:

$\langle$aaaa$\rangle$ 0 $\langle$aaab$\rangle$ 1 $\langle$aabb$\rangle$ 2 $\langle$abab$\rangle$ 0 $\langle$abbb$\rangle$ 1 $\langle$bbbb$\rangle$ 0

Thus, the average number of runs in necklace is $\frac{4}{6} = \frac{2}{3}$.

If and only if a string w of length n is primitive, there is n strings v such that $\langle v \rangle = \langle w \rangle$. Consider the number of repetitions in non-primitive string.

Lemma 15. *For string w and integer k, $csqr\left(w^k\right) = k\ csqr\left(w\right)$, $crun\left(w^k\right) = k\ crun\left(w\right)$ and $cexp\left(w^k\right) = k\ cexp\left(w\right)$.*

Proof. By the definition of w^ω, w^ω and $\left(w^k\right)^\omega$ are the same strings. Shifting w^ω to left or right by $|w|$, we get the same string. If there is a repetition in w^ω at the position i, repetitions also exist at the positions $i + |w|$, $i + 2|w|$, ..., $i + (k-1)|w|$.

It is known that the number $|NL_{n,\sigma}|$ of necklaces of length n and alphabet size σ is

$$|NL_{n,\sigma}| = \frac{1}{n} \sum_{d|n} \phi\left(\frac{n}{d}\right) \sigma^d,$$

where $\phi\left(n\right)$ is the Euler's phi function. The function $\phi\left(n\right)$ is defined to be the number of integers less or equal to n which are coprime to n and can be written as:

$$\phi\left(n\right) = \sum_{d|n} \mu\left(\frac{n}{d}\right) d.$$

Using the method calculating the number of necklaces and Lemma 15, we can obtain the number of squares in all necklaces.

Lemma 16. *The number of squares in all necklaces of length n and alphabet size σ is*

$$|NL_{n,\sigma}|\, S_n\left(d, \sigma\right) = \frac{1}{n} \sum_{d|n} \phi\left(\frac{n}{d}\right) \frac{n}{d} \sigma^d S_c\left(d, \sigma\right).$$

Proof. Let T be a multi set of strings obtained by cutting necklaces $NL_{n,\sigma}$ in n ways. For example, for

$$NL_{4,2} = \{\langle \mathtt{aaaa} \rangle, \langle \mathtt{aaab} \rangle, \langle \mathtt{aabb} \rangle, \langle \mathtt{abab} \rangle, \langle \mathtt{abbb} \rangle, \langle \mathtt{bbbb} \rangle\}.$$

T is as follows:

$$T = \left\{ \begin{array}{ll} \mathtt{aaaa\ aaaa\ aaaa\ aaaa} & \mathtt{abab\ baba\ abab\ baba} \\ \mathtt{aaab\ aaba\ abaa\ baaa} & \mathtt{abbb\ bbba\ bbab\ babb} \\ \mathtt{aabb\ abba\ bbaa\ baab} & \mathtt{bbbb\ bbbb\ bbbb\ bbbb} \end{array} \right\}.$$

We see that $|T| = n|NL_{n,\sigma}|$ and the number of circular squares in T is $n\,|NL_{n,\sigma}|S_n\left(d, \sigma\right)$. The number of $w \in \Sigma^n$ in T equals to the number of k such that $1 \le k \le n$ and $w = w[k+1..n]w[1..k]$. This equations holds if w can be written as $w = u^{\frac{n}{|u|}}$ using $u \in \Sigma^{\gcd(k,n)}$. Thus, from Lemma 15,

$$n\,|NL_{n,\sigma}|\, S_n\left(d, \sigma\right) = \sum_{k=1}^{n} \frac{n}{\gcd(k,n)} \sigma^{\gcd(k,n)} S_c\left(\gcd(k,n), \sigma\right).$$

Since $\gcd(k,n)$ is a divisor of n, this equation can be transformed, with $d = \gcd(k,n)$, as follows:

$$\begin{aligned}
|NL_{n,\sigma}|\, S_n(p,\sigma) &= \frac{1}{n}\sum_{d|n}\sum_{k=1}^{n}\frac{n}{d}\sigma^d S_c(d,\sigma)\,[d=\gcd(k,n)] \\
&= \frac{1}{n}\sum_{d|n}\left(\frac{n}{d}\sigma^d S_c(d,\sigma)\sum_{k=1}^{n}\left[\frac{k}{d}\perp\frac{n}{d}\right]\right) \\
&= \frac{1}{n}\sum_{d|n}\left(\frac{n}{d}\sigma^d S_c(d,\sigma)\sum_{k'=1}^{\frac{n}{d}}\left[k'\perp\frac{n}{d}\right]\right) \\
&= \frac{1}{n}\sum_{d|n}\frac{n}{d}\sigma^d S_c(d,\sigma)\,\phi\left(\frac{n}{d}\right).
\end{aligned}$$

From the number of necklaces and Lemma 16, we can derive the following theorem.

Theorem 17. *For integers n and σ, the average number of squares in necklace of length n and alphabet size σ is*

$$S_n(p,\sigma) = \frac{\sum_{d|n}\phi\left(\frac{n}{d}\right)\frac{n}{d}\sigma^d S_c(d,\sigma)}{\sum_{d|n}\phi\left(\frac{n}{d}\right)\sigma^d}.$$

Similarly we obtain the average number and the average sum of exponents of runs in necklace.

Theorem 18. *For integers n and σ, the average number of runs in necklace of length n and alphabet size σ is*

$$R_n(p,\sigma) = \frac{\sum_{d|n}\phi\left(\frac{n}{d}\right)\frac{n}{d}\sigma^d R_c(d,\sigma)}{\sum_{d|n}\phi\left(\frac{n}{d}\right)\sigma^d},$$

and the average sum of exponents of runs in necklace is

$$E_n(p,\sigma) = \frac{\sum_{d|n}\phi\left(\frac{n}{d}\right)\frac{n}{d}\sigma^d E_c(d,\sigma)}{\sum_{d|n}\phi\left(\frac{n}{d}\right)\sigma^d}.$$

5 Conclusion

In this paper we defined circular squares and circular runs in a string and considered squares and runs in a necklace. They are useful for analysing ordinary squares and runs, especially a lower bound of the number of them. We showed the average number of runs, the average number of squares and the average number of sum of exponents of runs in a necklace. It would also be interesting problem to analyse the average number of distinct repetitions instead of their occurrences.

References

1. M. Crochemore: *An optimal algorithm for computing the repetitions in a word.* Information Processing Letters, 12 1981, pp. 244–250.
2. M. Crochemore and L. Ilie: *Analysis of maximal repetitions in strings*, in Proceedings of the 32nd International Symposium on Mathematical Foundations of Computer Science (MFCS 2007), vol. 4708 of LNCS, Springer-Verlag, 2007, pp. 465–476.
3. M. Crochemore, L. Ilie, and L. Tinta: *The "runs" conjecture.* http://www.csd.uwo.ca/~ilie/runs.html.
4. M. Crochemore, L. Ilie, and L. Tinta: *Towards a solution to the "runs" conjecture*, in Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching (CPM 2008), vol. 5029 of LNCS, Springer-Verlag, 2008, pp. 290–302.
5. N. Fine and H. Wilf: *Uniqueness theorems for periodic functions.* Proceedings of the American Mathematical Society, 1965, pp. 109–114.
6. R. Kolpakov and G. Kucherov: *Finding maximal repetitions in a word in linear time*, in Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS 1999), IEEE Computer Society, 1999, pp. 596–604.
7. R. Kolpakov and G. Kucherov: *On the sum of exponents of maximal repetitions in a word*, Tech. Rep. 99-R-034, LORIA, France, 1999.
8. K. Kusano, W. Matsubara, A. Ishino, and A. Shinohara: *Average value of sum of exponents of runs in strings*, in Proceedings of the Prague Stringology Conference 2008, Czech Technical University in Prague, 2008, pp. 185–192.
9. W. Matsubara, K. Kusano, H. Bannai, and A. Shinohara: *A series of run-rich strings*, in Proceedings of the 3rd International Conference on Language and Automata Theory and Applications (LATA 2009), Springer, 2009, pp. 578–587.
10. S. J. Puglisi and J. Simpson: *The expected number of runs in a word.* Australasian Journal of Combinatorics, 42 2008, pp. 45–54.
11. J. Simpson: *Modified padovan words and the maximum number of runs in a word.* Australasian Journal of Combinatorics, 46 2010, pp. 129–146.