数理解析研究所講究録
第1649巻 2009年 181-188

文字列の繰り返し構造の平均解析

草野 一彦† 松原 渉† 石野 明‡ 篠原 歩†

2009年2月3日

概要

文字列 w の部分文字列 $w[i..j]$ において任意の $i \leq k \leq j-p$ で $w[k] = w[k+p]$ が成り立つとき，$w[i..j]$ を周期 p の繰り返し（repetition）と呼ぶ．連（run，maximal repetition）とは左右に延長不可能な繰り返しである．連の周期に対する長さの比 $\frac{j-i+1}{p}$ を連の指数（exponent）と呼ぶ．近年，文字列に含まれる連の個数や指数の和について盛んに研究が進められている．Crochemore と Ilie は長さ n の文字列に含まれる連の指数の和の上界が $2.9n$ であることを示したが，その最大値は未だ知られていない．本稿では任意の長さ任意のアルファベットサイズの文字列に含まれる連の指数の和の平均を表す厳密な閉じた式を示し，文字列の長さが無限大となるときの長さあたりの指数の和の極限値を示す．バイナリ文字列ではこの値はおよそ 1.13103 である．また，この結果を用いて文字列中のちょうど2回の繰り返しである平方の平均個数やその極限値を示す．

1 導入

文字列中の繰り返しはデータ圧縮や遺伝子解析などに応用される重要な問題の一つである．

平方とはちょうど2回の繰り返しである．平方にはいくつかの数え方があるが，本稿では素な根を持つ平方のみを考え，平方の種類数ではなく出現数を数える．このように平方を数えた場合，長さ n の文字列に含まれる平方の個数は $O(n \log n)$ であることが知られている [1].

連とは左右に延長不可能な繰り返しである．Kolpakov と Kucherov は長さ n の文字列に含まれる連の最大個数 $\rho(n)$ が高々 cn であることを示した [5]．彼らは定数 c について具体的な値を与えなかったが，近年この定数を下げるための精力的な研究が活発に行われている [2, 4, 9, 12]．現在，$c \leq 1.029$ であることが示されており [3, 4]，$c < 1$ が予想されている．連の繰り返し回数を指数（exponent）と呼び，指数の和の最大値についても良く研究されている [2, 6, 11]．指数の和の最大値も長さに対して線形であることが示されている．現在の最良の上界は $2.9n$ であり [2]，$2n$ 未満であると予想されている．

連の最大個数 $\rho(n)$ が未だ知られていないにも関わらず，Puglisi と Simpson は長さ n の文字列に含まれる連の平均個数を与える次の式を示した [10].

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

ここで σ はアルファベットサイズ，$\mu(n)$ はメビウス関数である．

本稿ではアルファベットサイズ σ 長さ n の文字列に含まれる指数の和の平均 $e_\sigma(n)$ と平方の平均個数 $s_\sigma(n)$ に注目し，以下の等式が成り立つことを証明する．

$$e_\sigma(n) = \sum_{p=1}^{\frac{n}{2}} L_\sigma(p) \left(2p(n-2p+1)\sigma^{-2p} - (2p-1)(n-2p)\sigma^{-2p-1}\right)$$

$$s_\sigma(n) = \sum_{p=1}^{\frac{n}{2}} L_\sigma(p) \left(p(n-2p+1)\sigma^{-2p}\right)$$

ここで，$L_\sigma(n)$ はアルファベットサイズ σ 長さ n のリンドン文字列の個数を表す．さらに，それぞれの極限値

† 東北大学 大学院情報科学研究科
‡ Google Japan Inc.

が次のようになることを示す.

$$\lim_{n\to\infty}\frac{e_\sigma(n)}{n}=\sum_{d=1}^{\infty}\mu(d)\left(\frac{2(\sigma-1)}{\sigma^{2d}-\sigma}+\frac{1}{d\sigma}\ln\left(\frac{\sigma^{2d}}{\sigma^{2d}-\sigma}\right)\right)$$

$$\lim_{n\to\infty}\frac{s_\sigma(n)}{n}=\sum_{d=1}^{\infty}\mu(d)\left(\frac{\sigma}{\sigma^{2d}-\sigma}\right)$$

2 定義

$\Sigma=\{0,1,2,\ldots,\sigma-1\}$ をサイズ σ のアルファベットとする. すなわち, $|\Sigma|=\sigma$ である. Σ 上の文字列の集合を Σ^* と表し, 長さ n の全ての文字列を Σ^n と書く. 文字列 $w=xyz$ について, x, y, z をそれぞれ w の**接頭辞** (*prefix*), **部分文字列** (*substring*), **接尾辞** (*suffix*) と呼ぶ. 文字列 w の長さを $|w|$ と表す. 長さ 0 の文字列を空文字列と呼び ε で表す. 文字列 w の i 番目の文字を $w[i-1]$ と書く. すなわち, $w=w[0]w[1]\ldots w[|w|-1]$ である. 文字列 w の部分文字列 $w[i]w[i+1]\ldots w[j]$ を $w[i..j]$ と表す. 文字列 w において任意の $i\geq 0$ で $w[i]=w[i+p]$ が成り立つ p を文字列 w の**周期** (*period*) と呼ぶ.

文字列 w がいかなる文字列 u と整数 $k\geq 2$ を用いても $w=u^k$ と書き表せないとき, 文字列 w は**素である** (*primitive*) と言う. 文字列 w が空文字列以外の w の全ての接尾辞の中で辞書順最小であるとき, w を**リンドン文字列** (*Lyndon word*) と呼ぶ [7]. 与えられたアルファベット上の長さ n のリンドン文字列の個数を $L_\sigma(n)$ と定義する. 定義よりリンドン文字列は素であり, 長さ n の素な文字列の個数は $nL_\sigma(n)$ となる. メビウス (Möbius) 関数 $\mu(n)$ は次のように定義される.

$$\mu(n)=\begin{cases}0 & n\text{が平方数で割りきれる}\\ 1 & n\text{が相異なる偶数個の素因数を持つ}\\ -1 & n\text{が相異なる奇数個の素因数を持つ}\end{cases}$$

$L_\sigma(n)$ は以下のように書き表せることが知られている [8].

$$L_\sigma(n)=\frac{1}{n}\sum_{d|n}\mu\left(\frac{n}{d}\right)\sigma^d$$

$d|n$ という表記は d が n の約数であることを示す. いくつかの n についての $\mu(n)$ と $L_2(n)$ の値は次のようになる.

n	1	2	3	4	5	6	7	8	9	10
$\mu(n)$	1	−1	−1	0	−1	1	−1	0	0	1
$L_2(n)$	2	1	2	3	6	9	18	30	56	99

文字列 w が周期 $p\leq\frac{|w|}{2}$ を持つ, すなわち 2 回以上の繰り返しであるとき, w は周期 p で**周期的** (*periodic*) であるという. 部分文字列 $w[i..j]$ が以下の条件を満たすとき $w[i..j]$ は周期 p で左右に**延長不可能** (*non-extendable*) であるという.

$$\begin{array}{lll} i=0 & \text{or} & w[i-1]\neq w[i+p-1]\\ j=n-1 & \text{or} & w[j+1]\neq w[j-p+1]\end{array}$$

w の部分文字列 $w[i..j]$ が周期的であり左右に延長不可能であるとき, $w[i..j]$ を**連** (*run, maximal repetition*) と呼ぶ. 連は周期性と延長不可能性を満たす最小の周期 p で 1 度のみ数える. 周期 p の連 $w[i..j]$ を三項組 $(i.j,p)$ で表す. 連の**根** (*root*) は長さ p の接頭辞 $w[i..i+p-1]$ であり, 連の**指数** (*exponent*) は周期に対する長さの比 $\frac{j-i+1}{p}$ である. 文字列 w に含まれる連の個数を $Runs(w)$, 連の指数の和を $Exp(w)$ と定義する.

部分文字列 v が, 素な文字列 u を用いて $v=u^2$ と書き表せるとき, v を**平方** (*square*) と呼ぶ. 平方の**根** (*root*) は u であり, **周期** (*period*) は $|u|$ である. 文字列 w 中の平方 u^2 を二項組 $(i,|u|)$ で表す. ここで i は w 中での u^2 の出現位置である. w 中の平方の個数を $Sqr(w)$ と定義する.

$w=$ 0101010110112120 に含まれる連と平方を図 1 に例示する. $w[0..7]=$ 01010101 は周期 4 を持つ 2 回繰り返しであるが, 根 $w[0..3]=$ 0101 が素ではないため平方とはみなさない. また, $w[0..7]$ は周期 2 の連として 1 度のみ数える. 平方 0101, 1010, 11 は w 中に複数回現われており本稿ではこれらを全て数える. $Runs(w)=5$, $Exp(w)=\frac{8}{2}+\frac{7}{3}+\frac{2}{1}+\frac{2}{1}+\frac{4}{2}=\frac{37}{3}$, $Sqrt(w)=10$ である.

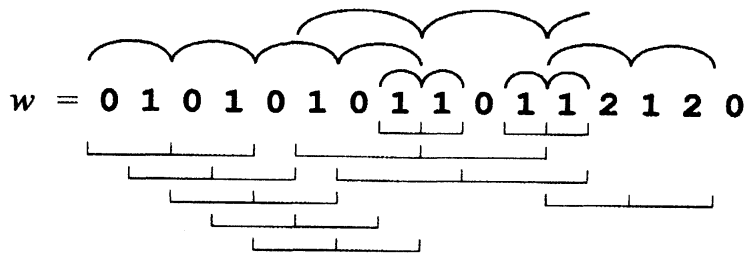

図 1　$w = 0101010110112120$ に含まれる連と平方.

3　結果

アルファベットサイズ σ 長さ n の文字列に含まれる，連の平均個数 $r_\sigma(n)$，連の指数の和の平均 $e_\sigma(n)$，平方の平均個数 $s_\sigma(n)$ をそれぞれ以下のように定義する.

$$r_\sigma(n) = average\{Runs(w) : w \in \Sigma^n\}$$
$$e_\sigma(n) = average\{Exp(w) : w \in \Sigma^n\}$$
$$s_\sigma(n) = average\{Sqr(w) : w \in \Sigma^n\}$$

Puglisi と Simpson は連の平均個数について以下の等式が成り立つことを示した.

定理 1 (Puglisi and Simpson [10]).

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

本稿では連の指数の和の平均と平方の平均個数について次の等式が成り立つことを証明する.

定理 2.

$$e_\sigma(n) = \sum_{p=1}^{\frac{n}{2}} L_\sigma(p)\left(2p(n-2p+1)\sigma^{-2p} - (2p-1)(n-2p)\sigma^{-2p-1}\right)$$

系 1.

$$s_\sigma(n) = \sum_{p=1}^{\frac{n}{2}} L_\sigma(p)\left(p(n-2p+1)\sigma^{-2p}\right)$$

アルファベットサイズ σ 長さ n の文字列 w と正数 p について長さ $n-p$ の文字列 $d_\sigma(w,p)$ を以下のように定義する.

$$d_\sigma(w,p)[i] = w[i+p] - w[i] \pmod{\sigma} \qquad \text{for } 0 \le i < n-p$$

ここで演算子 '$-$' と 'mod' は文字を数字と見なして適用する．例えば，$\Sigma = \{0,1,2\}$ 上の文字列 $w = 21010$ について，$d_\sigma(w,1) = 2212$ であり $d_\sigma(w,2) = 100$ である.

文字列 $w \in \Sigma^n$ の部分文字列 $w[i..j]$ が 0 の最大ブロックである，すなわち任意の t $(i \le t \le j)$ について $w[t] = 0$ が成り立ち以下の条件を満たすとき，$w[i..j]$ を 0-セグメントと呼ぶ.

$$\begin{array}{lll} i = 0 & \text{or} & w[i-1] \neq 0 \\ j = n-1 & \text{or} & w[j+1] \neq 0 \end{array}$$

二項組 (i,j) で 0-セグメント表す.

例 1. 文字列 0012000102 に含まれる 0-セグメントは $(0,1)$, $(4,6)$, $(8,8)$ である.

補題 1. 任意の文字列 w について，$d_\sigma(w,p)[i..j]$ が長さ p 以上の 0-セグメントであるときかつその場合に限り，部分文字列 $w[i..j+p]$ は周期 p の連である．

証明. 0-セグメント (i,j) が $d_\sigma(w,p)$ に存在するとき，$w[t] = w[t+p]$ $(i \le t \le j)$ が成り立つ．すなわち，$w[i..j+p]$ は周期 p を持つ．$|d_\sigma(w,p)[i..j]| = j-i+1 \ge p$ が成り立つときかつこの場合に限り $|w[i..j+p]| = j+p-i+1 \ge 2p$ である．また，$w[i..j+p]$ は左右に延長不可能である．よって，$w[i..j+p]$ は連である．$w[i..j+p]$ は周期 p の連であるときに $d_\sigma(w,p)[i..j]$ が 0-セグメントであることは，自明である． □

なお，この補題の中での周期 p は最小であるとは限らない．

アルファベットサイズ σ 長さ n の全ての文字列 Σ^n 中の長さ p の 0-セグメントの個数を $c_\sigma(n,p)$ で，Σ^n 中の長さが p 以上の 0-セグメントの個数を $C_\sigma(n,p)$ で表す．定義から，$C_\sigma(n,p) = \sum_{i=p}^{n} c_\sigma(n,i)$ である．Σ^n 中の長さ p 以上の 0-セグメントについて，$\frac{l}{p}$ の合計を $E_\sigma(n,p)$ で表す．ここで，l は各 0-セグメント長さである．すなわち $E_\sigma(n,p) = \sum_{i=p}^{n} c_\sigma(n,i)\frac{i}{p}$ である．

例 2. $\sigma = 2$ のとき $c_\sigma(5,2)$ は 12 である．以下に長さ 5 の全てのバイナリ文字列に含まれる長さ 2 の 0-セグメントを下線で示す．

00000	00100	01000	01100	10000	10100	11000	11100
00001	00101	01001	01101	10001	10101	11001	11101
00010	00110	01010	01110	10010	10110	11010	11110
00011	00111	01011	01111	10011	10111	11011	11111

同様に，$c_\sigma(5,3) = 5$, $c_\sigma(5,4) = 2$ であり，$c_\sigma(5,5) = 1$ である．よって，$C_\sigma(5,2) = 12+5+2+1 = 20$ となる．$E_\sigma(5,2) = 12\cdot\frac{2}{2} + 5\cdot\frac{3}{2} + 2\cdot\frac{4}{2} + \frac{5}{2} = 26$ である．

補題 2. 任意の正数 n と $p \le n$ について以下の式が成り立つ．

$$C_\sigma(n,p) = (n-p+1)\sigma^{n-p} - (n-p)\sigma^{n-p-1}$$
$$E_\sigma(n,p) = \frac{1}{p}\left(p(n-p+1)\sigma^{n-p} - (p-1)(n-p)\sigma^{n-p-1}\right)$$

証明. アルファベットサイズ σ 長さ n の文字列を長さ p の 0-セグメントで分割した文字列の組の集合 $Q_{\sigma,n,p}$ を次のように定義する．

$$Q_{\sigma,n,p} = \{(u,w) \mid uvw \in \Sigma^n \text{ and } v \text{ は } uvw \text{ 中の 0-セグメント }\}.$$

長さ p の 0-セグメントと $Q_{\sigma,n,p}$ は一対一に対応するので $c_\sigma(n,p) = |Q_{\sigma,n,p}|$ である．

例 3. $\sigma = 2, n = 3, p = 1$ について，Σ^n に含まれる長さ p の 0-セグメントは以下のようになる．

$$\Sigma^3 = \{000, 001, \underline{0}1\underline{0}, \underline{0}11, 100, 1\underline{0}1, 11\underline{0}, 111\}$$

0-セグメントの個数 $c_\sigma(3,1)$ は 5 であり，$Q_{\sigma,3,1}$ は次のようになる．

$$Q_{\sigma,3,1} = \{(\varepsilon,10), (01,\varepsilon), (\varepsilon,11), (1,1), (11,\varepsilon)\}$$

$c_\sigma(n,p)$ の表式は $Q_{\sigma,n,p}$ に含まれる要素の個数を 2 つの場合について考えることで得られる．

(1) $p \le n-1$

$Q_{\sigma,n,p}$ を 0-セグメントの位置によって次のように $Q^l_{\sigma,n,p}$, $Q^m_{\sigma,n,p}$, $Q^r_{\sigma,n,p}$ の 3 つの集合に分ける．

$$Q^l_{\sigma,n,p} = \{(u,w) \in Q_{\sigma,n,p} \mid u = \varepsilon\},$$
$$Q^m_{\sigma,n,p} = \{(u,w) \in Q_{\sigma,n,p} \mid u \ne \varepsilon \text{ and } w \ne \varepsilon\},$$
$$Q^r_{\sigma,n,p} = \{(u,w) \in Q_{\sigma,n,p} \mid w = \varepsilon\}.$$

$u = \varepsilon$ であるとき，$|w| = n-p$ かつ $w[0] \ne 0$ であるから，$w[0]$ について $(\sigma-1)$ 通りの $w[1..n-p-1]$ について σ^{n-p-1} 通りの選び方がある．よって，$|Q^l_{\sigma,n,p}| = (\sigma-1)\sigma^{n-p-1}$．同様に，$w = \varepsilon$ のとき

$|Q^r_{\sigma,n,p}| = (\sigma-1)\sigma^{n-p-1}$ である．$u \neq \varepsilon$ かつ $w \neq \varepsilon$ である場合，$u[|u|-1]$ と $w[0]$ は 0 ではなく，$|u|+|w| = n-p$ である．0^p の位置が $(n-p-1)$ 通りあり，$u[|u|-1]$ と $w[0]$ について $(\sigma-1)$ 通り，その他の文字について σ^{n-p-2} 通りの選び方がある．すなわち $|Q^m_{\sigma,n,p}| = (n-p-1)(\sigma-1)^2\sigma^{n-p-2}$ となる．$p = n-1$ のとき，u もしくは w は ε であり，$(n-p-1)(\sigma-1)^2\sigma^{n-p-2}$ は 0 となる．ゆえに，

$$\begin{aligned} c_\sigma(n,p) &= |Q_{\sigma,n,p}| = |Q^l_{\sigma,n,p}| + |Q^m_{\sigma,n,p}| + |Q^r_{\sigma,n,p}| \\ &= (n-p+1)\sigma^{n-p} - 2(n-p)\sigma^{n-p-1} + (n-p-1)\sigma^{n-p-2}. \end{aligned}$$

(2) $p = n$

$u = w = \varepsilon$ であるから，$Q_{\sigma,n,p} = \{(\varepsilon,\varepsilon)\}$ となり，$c_\sigma(n,p) = |Q_{\sigma,n,p}| = 1$ である．

$p \leq n-1$ のとき，

$$C_\sigma(n,p) = \sum_{i=p}^{n} c_\sigma(n,i) = (n-p+1)\sigma^{n-p} - (n-p)\sigma^{n-p-1}.$$

この等式は $C_\sigma(n,n) = 1$ についても成り立つ．

$p \leq n-1$ のとき，

$$E_\sigma(n,p) = \sum_{i=p}^{n} c_\sigma(n,i)\frac{i}{p} = \frac{1}{p}\left(p(n-p+1)\sigma^{n-p} - (p-1)(n-p)\sigma^{n-p-1}\right).$$

この等式は $E_\sigma(n,n) = 1$ についても成り立つ． □

補題 3. $d_\sigma(w,p) = d_\sigma(v,p)$ を満たすような任意の整数 p と長さ n の文字列 w と v について，ある i で $w[i..i+p-1] = v[i..i+p-1]$ が成り立つときその場合に限り $w = v$ である．

証明. ($\Longrightarrow$) 数学的帰納法により示す．$i \leq j < i+p$ と k を整数とする．$k = 0$ のとき $w[j+kp] = v[j+kp]$ が成り立つ．$k \geq 1$ のとき，$w[j+kp] = v[j+kp]$ が成り立つならば，$w[j+(k+1)p] = w[j+kp]+d_\sigma(w,p)[j+kp] \pmod\sigma = v[j+kp]+d_\sigma(v,p)[j+kp] \pmod\sigma = v[j+(k+1)p]$ となる．よって，$w[i..n-1] = v[i..n-1]$．同様に，$w[0..i+p-1] = v[0..i+p-1]$．ゆえに，$w = v$．
($\Longleftarrow$) 自明である． □

$w \in \Sigma^n$ ついて $d_\sigma(w,p)$ の長さは $n-p$ であり，Σ^{n-p} に含まれる 0-セグメントの個数は $C_\sigma(n-p,p)$ である．補題 1 と 3 から Σ^n には周期 p の連が $\sigma^p C_\sigma(n-p,p)$ 含まれている．しかし，連は複数の周期を持ちうる．例えば，0101010101 は 2 と 4 の両方の周期を持っている．このような連を 2 回以上数えることを防ぐために，連をその最小の周期で数えることを考える．

補題 4. 最小の周期が p である連の個数と周期 p を持つ連の個数の比は $\frac{pL(p)}{\sigma^p}$ である．

証明. 周期 p を持つ連の個数は $\sigma^p C_\sigma(n-p,p)$ である．周期性補題 [7] より，もし周期 p を持つ連が異なる周期 $q < p$ を持つならば，その連は $gcd(p,q)$ も周期として持つ．つまり $w[i..j]$ が $q < p$ を周期として持たないならば，その根は素である．長さ p の素な文字列の個数は $pL_\sigma(p)$ である．ゆえに，最小の周期が p である連の個数は $pL_\sigma(p)C_\sigma(n-p,p)$ である． □

補題 1 より，$d_\sigma(w,p)$ 中の長さ l $(l \geq p)$ の 0-セグメントは w 中の長さ $l+p$ の連に対応している．これらの連の指数は $\frac{l}{p}+1$ である．このことと補題 2, 3, 4 から Σ^n に含まれる連の指数の和 $\sigma^n e_\sigma(n)$ は以下のように導かれる．

$$\begin{aligned} \sigma^n e_\sigma(n) &= \sum_{p=1}^{\frac{n}{2}} pL_\sigma(p)\left(E_\sigma(n-p,p) + C_\sigma(n-p,p)\right) \\ &= \sum_{p=1}^{\frac{n}{2}} L_\sigma(p)\left(2p(n-2p+1)\sigma^{n-2p} - (2p-1)(n-2p)\sigma^{n-2p-1}\right) \end{aligned}$$

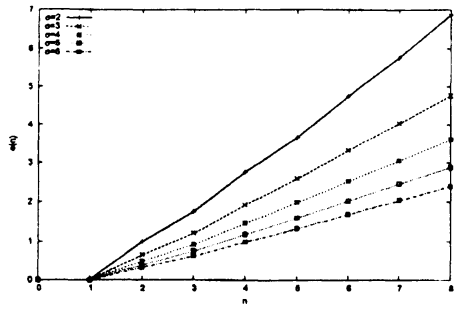

図 2　文字列中の連の指数の和の平均値 $e_\sigma(n)$.

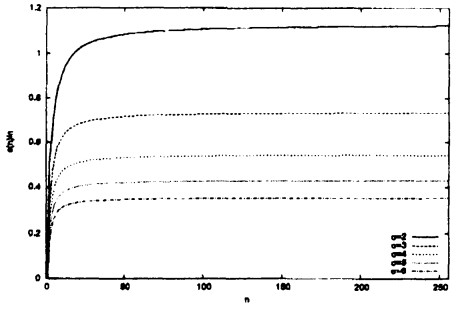

図 3　文字列中の長さあたりの連の指数の和の平均値 $\frac{e_\sigma(n)}{n}$.

これより定理 2 が得られる．図 2 より連の指数の和 $e_\sigma(n)$ は n が大きくなるにつれほぼ線形に増加することがわかる．$\frac{e_\sigma(n)}{n}$ の収束の様子を図 3 に示す．$e_\sigma(n)$ の極限について次の定理が成り立つ．

定理 3. $n \to \infty$ となるときの $\frac{e_\sigma(n)}{n}$ の極限は以下の値となる．

$$\sum_{d=1}^{\infty} \mu(d) \left(\frac{2(\sigma-1)}{\sigma^{2d}-\sigma} + \frac{1}{d\sigma} \ln \left(\frac{\sigma^{2d}}{\sigma^{2d}-\sigma} \right) \right)$$

この定理を証明するために $\frac{e_\sigma(n)}{n}$ を書き換える．

主張 1.

$$\frac{e_\sigma(n)}{n} = \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \sigma^{-2pd+p-1} \left(2(\sigma-1) - \frac{4pd}{n}(\sigma-1) + \frac{1}{pd} + \frac{2}{n}(\sigma-1) \right)$$

証明. $2p(n-2p+1)\sigma^{-2p} - (2p-1)(n-2p)\sigma^{-2p-1}$ を $f(p)$ と置く．

$$\begin{aligned}
e_\sigma(n) &= \sum_{p=1}^{\frac{n}{2}} L_\sigma(p) f(p) \\
&= \sum_{p=1}^{\frac{n}{2}} \sum_{d|p} \mu\left(\frac{p}{d}\right) \sigma^d \frac{f(p)}{p} \\
&= \begin{array}{l}
\mu(1)\sigma^1 \frac{f(1)}{1} \\
+\mu(2)\sigma^1 \frac{f(2)}{2} + \mu(1)\sigma^2 \frac{f(2)}{2} \\
+\mu(3)\sigma^1 \frac{f(3)}{3} \qquad\quad + \mu(1)\sigma^3 \frac{f(3)}{3} \\
+\mu(4)\sigma^1 \frac{f(4)}{4} + \mu(2)\sigma^2 \frac{f(4)}{4} \qquad + \mu(1)\sigma^4 \frac{f(4)}{4} \\
+\mu(5)\sigma^1 \frac{f(5)}{5} \qquad\qquad\quad + \mu(1)\sigma^5 \frac{f(5)}{5} \\
+\mu(6)\sigma^1 \frac{f(6)}{6} + \mu(3)\sigma^2 \frac{f(6)}{6} + \mu(2)\sigma^3 \frac{f(6)}{6} \qquad + \mu(1)\sigma^6 \frac{f(6)}{6} \\
\qquad\qquad\qquad\qquad \vdots
\end{array} \\
&= \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \sigma^p \frac{f(pd)}{pd} \qquad (\mu(d) \text{ を括り出した}) \\
&= \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \frac{1}{pd} \left(2pd(n-2pd+1)\sigma^{-2pd+p} - (2pd-1)(n-2pd)\sigma^{-2pd+p-1} \right) \\
\frac{e_\sigma(n)}{n} &= \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\frac{n}{2d}} \sigma^{-2pd+p-1} \left(2(\sigma-1) - \frac{4pd}{n}(\sigma-1) + \frac{1}{pd} + \frac{2}{n}(\sigma-1) \right)
\end{aligned}$$

□

定理 3 を証明する.

証明. (定理 3) $\frac{pd}{n}$ を考慮すべきときは $\sigma^{-2pd+p-1}$ が充分小さくなるため, 項 $\sigma^{-2pd+p-1}\frac{4pd}{n}$ は $n \to \infty$ のとき無視できる.

$$\begin{aligned}\lim_{n\to\infty}\frac{e_\sigma(n)}{n} &= \lim_{n\to\infty}\sum_{d=1}^{\frac{n}{2}}\mu(d)\sum_{p=1}^{\frac{n}{2d}}\sigma^{-2pd+p-1}\left(2(\sigma-1)+\frac{1}{pd}\right)\\ &= \lim_{n\to\infty}\sum_{d=1}^{\frac{n}{2}}\mu(d)\left(\frac{2\sigma^{-2d}(\sigma-1)}{1-\sigma^{1-2d}}-\frac{1}{d\sigma}\ln(1-\sigma^{1-2d})\right)\\ &= \sum_{d=1}^{\infty}\mu(d)\left(\frac{2(\sigma-1)}{\sigma^{2d}-\sigma}+\frac{1}{d\sigma}\ln\left(\frac{\sigma^{2d}}{\sigma^{2d}-\sigma}\right)\right)\end{aligned}$$

□

平方の個数は連の個数と指数の和に関係しているため, $r_\sigma(n)$ と $e_\sigma(n)$ の式を用いて $s_\sigma(n)$ を導ける.

補題 5. 任意の文字列 w と任意の文字列 p について, w 中の周期 p の平方の個数 $sqr(w,p)$ と連の個数 $run(w,p)$, 指数の和 $exp(w,p)$ は次の公式を満たす.

$$sqr(w,p) = p\ exp(w,p) - (2p-1)\ run(w,p)$$

例 4. 文字列 001001101101100 に含まれる周期 3 の連は $(0,5,3)$ と $(4,13,3)$ である. 連の個数は 2 であり, 指数の和は $\frac{16}{3}$ である. この文字列に含まれる周期 3 の平方は $(0,3)$, $(4,3)$, $(5,3)$, $(6,3)$, $(7,3)$, $(8,3)$ である.

証明. w 中の連 (i,j,p) について, 連の定義より $w[l] = w[l+p]$ $(i \le l \le j-p)$ であるから $w[k..k+p-1] = w[k+p..k+2p-1]$ $(i \le k \le j-2p+1)$ が成り立つ. 各 $w[k..k+p-1]$ は $w[i..i+p-1]$ と共役であり素である. 指数 $e = \frac{j-i+1}{p}$ の連は $j-i-2p+2$ 個の平方を含む. 平方は周期的なので周期 p を持つ平方は必ず周期 p の連に含まれる. よって, 連 (i,j,p) に含まれる平方の個数 s について次の等式が成り立つ.

$$s = j-i-2p+2 = p\,e-(2p-1)$$

w 中の各連について両辺を足し合わせることで, 補題 5 が得られる. □

Puglisi と Simpson の結果 [10] を我々の形に合わせて書き換えると以下のようになる.

$$r_\sigma(n) = \sum_{p=1}^{\frac{n}{2}} L_\sigma(p)\left(p(n-2p+1)\sigma^{-2p} - (n-2p)\sigma^{-2p-1}\right)$$

補題 5 を各文字列 $w \in \Sigma^n$ に適用することで系 1 を得る.

証明. (系 1)

$$\begin{aligned}\sigma^n s_\sigma(n) &= \sum_{p=1}^{\frac{n}{2}}\Big(pL_\sigma(p)\left(2p(n-2p+1)\sigma^{n-2p} - (2p-1)(n-2p)\sigma^{n-2p-1}\right)\\ &\qquad -(2p-1)L_\sigma(p)\left(p(n-2p+1)\sigma^{n-2p} - p(n-2p)\sigma^{n-2p-1}\right)\Big)\\ &= \sum_{p=1}^{\frac{n}{2}} L_\sigma(p)\left(p(n-2p+1)\sigma^{n-2p}\right)\\ s_\sigma(n) &= \sum_{p=1}^{\frac{n}{2}} L_\sigma(p)\left(p(n-2p+1)\sigma^{-2p}\right)\end{aligned}$$

□

平方の平均個数 $s_\sigma(n)$ もまた n が大きくなるにつれて線形に増加する. $\frac{e_\sigma(n)}{n}$ と同様にして $\frac{s_\sigma(n)}{n}$ の極限値が得られる.

表 1 $\frac{e_\sigma(n)}{n}$, $\frac{r_\sigma(n)}{n}$, $\frac{s_\sigma(n)}{n}$ の極限値.

σ	2	3	4	5	6
$\lim_{n\to\infty} \frac{e_\sigma(n)}{n}$	1.13103	0.73822	0.54459	0.43039	0.35536
$\lim_{n\to\infty} \frac{r_\sigma(n)}{n}$	0.41165	0.30491	0.23736	0.19329	0.16268
$\lim_{n\to\infty} \frac{s_\sigma(n)}{n}$	0.82330	0.45736	0.31648	0.24161	0.19522

系 2.

$$\lim_{n\to\infty} \frac{s_\sigma(n)}{n} = \sum_{d=1}^{\infty} \mu(d) \left(\frac{\sigma}{\sigma^{2d} - \sigma} \right)$$

表 1 に $\frac{e_\sigma(n)}{n}$, $\frac{r_\sigma(n)}{n}$, $\frac{s_\sigma(n)}{n}$ の極限値を示す.

4 結論

本稿では上界が未解決問題であるにも関わらず，文字列中の連の指数の和の平均および平方の平均個数を厳密に求めた．これは，1 つの文字列の中で連がどのように組み合うかはわからないものの，全ての文字列に含まれるある周期を持つ連の個数やその中で根が素であるものの割合が計算できるためである．さらに本稿では文字列長が無限大に近づくときの長さあたりのこれらの値の極限値を与えた．無限長のバイナリアルファベットでは，長さあたりの指数の和の平均はおよそ 1.13103 であり，平方の平均個数はおよそ 0.82330 である．

参考文献

[1] M. Crochemore. An optimal algorithm for computing the repetitions in a word. *Information Processing Letters*, 12:244–250, 1981.

[2] M. Crochemore and L. Ilie. Analysis of Maximal Repetitions in Strings. In *Proc. 32nd International Symposium on Mathematical Foundations of Computer Science (MFCS 2007)*, volume 4708 of *LNCS*, pages 465–476, 2007.

[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)*, volume 5029 of *LNCS*, pages 290–302, 2008.

[5] R. Kolpakov and G. Kucherov. Finding Maximal Repetitions in a Word in Linear Time. In *Proc. 40th Annual Symposium on Foundations of Computer Science (FOCS 1999)*, pages 596–604, 1999.

[6] R. Kolpakov and G. Kucherov. On the sum of exponents of maximal repetitions in a word. Technical Report 99-R-034, LORIA, France, 1999.

[7] M. Lothaire. *Algebraic combinatorics on words.* Cambridge University Press New York, 2002.

[8] M. Lothaire. *Applied Combinatorics on Words.* Cambridge University Press New York, 2005.

[9] S. Puglisi, J. Simpson, and W. Smyth. How many runs can a string contain? *Theoretical Computer Science*, 401(1-3):165–171, 2008.

[10] S. J. Puglisi and J. Simpson. The expected number of runs in a word. *Australasian Journal of Combinatorics*, 42:45–54, 2008.

[11] W. Rytter. The number of runs in a string: Improved analysis of the linear upper bound. In *Proc. 23rd Annual Symposium on Theoretical Aspects of Computer Science (STACS 2006)*, volume 3884 of *LNCS*, pages 184–195, 2006.

[12] W. Rytter. The number of runs in a string. *Information and Computation*, 205(9):1459–1469, 2007.