文字列の連の平均個数に関する研究

篠原研究室　4 年　草野　一彦

平成 20 年 3 月 1 日

1　はじめに

文字列中の繰り返し構造はデータ圧縮や遺伝子解析等に応用される重要な問題の一つである．本研究は中でも，左右に延長不可能な 2 回以上の繰り返しである連に着目した．例えば，文字列 $abaababa$ における連は $abaaba$, aa, $ababa$ の 3 つである．長さ n の文字列に含まれる連の最大個数を $\rho(n)$ と表す．$\rho(n)$ は文字列を網羅的に生成して連を数えることで確認できる．

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\rho(n)$	0	1	1	2	2	3	4	5	5	6	7	8	8	10	10	11

$\rho(n)$ の一般項は未だに知られていないが，Kolpakov と Kucherov は 1999 年に $\rho(n)$ が高々 cn であることを証明した [4]．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おり，Puglisi はこの定数 c が $c \leq 3.48$ であることを [10]，Rytter は $c \leq 3.44$ であることをそれぞれ示した [9]．Crochemore と Ilie は $c \leq 1.6$ であることを証明した [1]．この証明は周期 p の連が長さ p の区間には平均的に高々 1 つしか存在しないということを示すものであり，コンピュータで周期の小さな連の個数が文字列の長さ n で抑えられることを検証して $c \leq 1.077$ まで下げている [7]．また，Franek らは下限が $n \to \infty$ で $\rho(n) = 0.927n$ に収束することを示し，上限がこれに等しいと予想している [2]．一方，長さ n の文字列に含まれる連の最小個数は，アルファベットサイズが 2 の場合 $n \leq 3$ で 0，$n \geq 4$ で 1 となり，アルファベットサイズが 3 以上では連を含まない任意の長さの文字列を作ることができるので 0 となる．

連の繰り返し回数を指数と呼ぶ．$abaaba$, aa, $ababa$ の指数はそれぞれ 2, 2, 2.5 である．文字列における連の指数の和の最大値についても研究されていて，Kolpakov と Kucherov が文字列の長さ n に対して線形であることを [3]，Rytter が $25n$ 以下であることを証明した [8]．Crochemore と Ilie は $2.9n$ 以下であることを示した [1]．この値は $2n$ 未満であると予想されている [3]．

本研究では文字列における連の平均的振舞に着目して，長さ n の文字列に含まれる連の個数の平均値 $r(n)$，および指数の和の平均 $e(n)$ が下記の等式で厳密に表現できることを示す（定理 1, 3）．

$$
\begin{aligned}
r(n) &= \sum_{p=1}^{\frac{n}{2}} pL(p)\left((n-2p+1)\,\sigma^{-2p} - (n-2p)\,\sigma^{-2p-1}\right) \\
e(n) &= \sum_{p=1}^{\frac{n}{2}} L(p)\left(2p(n-2p+1)\sigma^{-2p} - (2p-1)(n-2p)\sigma^{-2p-1}\right)
\end{aligned}
$$

また，$\frac{r(n)}{n}, \frac{e(n)}{n}$ が $n \to \infty$ で下式の値に収束することを示す（定理 2, 4）.

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

ここで σ はアルファベットサイズ，$L(n)$ は長さ n のリンドン文字列の個数，$\mu(n)$ はメビウス関数である.

このことにより，文字列の連の個数と指数の和について一般のアルファベットに対して平均値の完全な解析ができたことになる.

2 定義

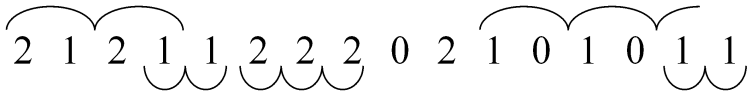
$\Sigma = \{0, 1, 2, \ldots, k-1\}$ をアルファベットとし，その大きさ $|\Sigma| = k$ を σ で表す．文字に対する四則演算は文字を数字とみなして行う．$w = w[0]w[1]\ldots w[l-1] \in \Sigma^*$ を文字列とし，その長さ l を $|w|$ と表わす．文字列 w の部分文字列 $w[i]w[i+1]\ldots w[j]$ を $w[i..j]$ と表す．負整数 i に対して，$w[i] = w[|w|+i]$ と定義する.

長さ n の文字列 w について，任意の i $(0 \le i < n-p)$ に対して $w[i] = w[i+p]$ を満たす整数 p を文字列 w の**周期**と呼ぶ．w の部分文字列 $w[i..j]$ で，最小の周期 p が $|w[i..j]| = j-i+1$ の半分以下であり，次の式を満たす（左右に延長不可能である）とき，$w[i..j]$ を**連**と呼び，開始位置 i，終了位置 j，周期 p を用いて (i, j, p) と表す.

$$((i=0) \vee (w[i-1] \neq w[i+p-1])) \wedge ((j=n-1) \vee (w[j+1] \neq w[j-p+1]))$$

連 (i, j, p) について，連の長さと最小の周期の比 $\frac{j-i+1}{p}$ を**指数**といい，最初の p 文字 $w[i..i+p-1]$ を**根**と呼ぶ．文字列 w に含まれる全ての連の集合を $Run(w)$ と表し，文字列 w に含まれる全ての連の指数の和を $Exp(w)$ と定義する.

例 1. $w = 2121122202101011$ とすると $Run(w) = \{(0,3,2), (3,4,1), (5,7,1), (10,14,2), (14,15,1)\}$ である．指数はそれぞれ 2, 2, 3, 2.5, 2 であるから，$Exp(w) = 2+2+3+2.5+2 = 11.5$ となる.



長さ n の文字列 w と $p < n$ に対して，長さ $n-p$ の文字列 d_w を次のように定義する.

$$d_w[i] = w[i+p] - w[i] \pmod{\sigma} \text{ for } 0 \le i < n-p$$

例 2. $w = 22012012112020$, $p = 3$ のとき，$d_w = 20000100211$ である.

	2	2	0	1	2	0	1	2	1	1	2	0	2	0			w
$-$				2	2	0	1	2	0	1	2	1	1	2	0	2	0	$w \gg 3$
				2	0	0	0	0	1	0	0	2	1	1				d_w

長さ n の文字列に含まれる連の平均個数 $average\{|Run(w)| : |w| = n\}$ を $r(n)$ と定義する．長さ n のすべての文字列 Σ^n に含まれる連の総数は $\sigma^n r(n)$ である．また，長さ n の文字列に含まれる連の指数の和の平均 $average\{Exp(w) : |w| = n\}$ を $e(n)$ と定義する．Σ^n に含まれる連の指数の総和は $\sigma^n e(n)$ である．

長さ n のすべての文字列 Σ^n に含まれる p 個以上の 0 の連続の総数を $c(p,n)$，p 個の 0 の連続の総数を $c'(p,n)$ と定義する．定義より，$c(p,n) = \sum_{i=p}^{n} c'(i,n)$ である．Σ^n に含まれる p 個以上の 0 の連続それぞれについてその長さを l として，$\frac{l}{p}$ の総和を $c_e(p,n)$ と定義する．

例 3. $\sigma = 2$ のとき，$z(5) = 80$, $c(2,5) = 20$, $c'(2,5) = 12$ である．2, 3, 4, 5 個の 0 の連続がそれぞれ 12, 5, 2, 1 個含まれているので，$c_e(2,5) = 12 \cdot \frac{2}{2} + 5 \cdot \frac{3}{2} + 2 \cdot \frac{4}{2} + \frac{5}{2} = 26$ である．

00000	**00**100	01**000**	011**00**	1**0000**	101**00**	11**000**	111**00**
00001	**00**101	01**00**1	01101	1**000**1	10101	11**00**1	11101
00010	**00**110	01010	01110	1**00**10	10110	11010	11110
00011	**00**111	01011	01111	1**00**11	10111	11011	11111

文字列 w が文字列 u と整数 $x \geq 2$ を用いて，$w = u^x$ と表せないとき，w は**プリミティブ**であるという．

文字列 w が巡回シフトした文字列の中において辞書順で最小であるとき，w を**リンドン文字列**と呼び，長さ n のリンドン文字列の個数を $L(n)$ と定義する [5]．例えば $w = aabab$ と巡回シフトした文字列を辞書順に整列すると $aabab$, $abaab$, $ababa$, $baaba$, $babaa$ となるので，w はリンドン文字列である．最小であるということからリンドン文字列はプリミティブであり，巡回シフトしたいずれの文字列とも異なるので，長さ n のプリミティブな文字列の個数は $nL(n)$ と表せる．

長さ n のプリミティブな文字列の個数 $nL(n)$ は $\sum_{d|p} \mu(\frac{p}{d})\sigma^d$ とも表せることが知られている [6]．ここで，$d|p$ は d が p の約数であることを表し，$\mu(n)$ はメビウス関数である．メビウス関数 $\mu(n)$ は n が平方因子を持つとき 0，n が相異なる k 個の素因数に分解されるとき $(-1)^k$ となる．

n	1	2	3	4	5	6	7	8	9	10	11	12
$\mu(n)$	1	−1	−1	0	−1	1	−1	0	0	1	−1	0

3 平均数の解析

3.1 連の平均個数

補題 1. p 個以上の 0 の連続が $d_w[i..j]$ $(j-i+1 \geq p)$ に存在するときかつこのときに限り，$w[i..j+p]$ に周期 p を持つ連がある．ただし，周期 p がこの連の最小の周期であるとは限らない．

例 4. 文字列 $w = 2121122202111110$ について $p = 2$ のとき，$d_w = 00211010120002$ である．d_w 中の p 個以上の 0 の連続は $d_w[0..1]$, $d_w[10..12]$ であり，w 中の周期 p を持つ連は $(0,3,2)$, $(10,14,1)$ である．

証明. 0 の連続が $d_w[i..j]$ にあるとき，$d_w[i] = d_w[i+1] = \cdots = d_w[j] = 0$ で d_w は次式を満たす．

$$((i = 0) \vee (w[i-1] \neq 0)) \wedge ((j = |d_w| - 1) \vee (w[j+1] \neq 0))$$

このことと $w[t+p] = w[t]$ iff $d_w[t] = 0$ $(i \leq t < j-p)$ より，$w[i..j+p]$ は左右に延長不可能で周期 p を持つ．また，$|d_w[i..j]| \geq p$ であるならば $|w[i..j+p]| \geq 2p$ である． □

補題 2. $p \leq n$ を満たす任意の整数 n, p について，$c(p,n) = (n-p+1)\sigma^{n-p} - (n-p)\sigma^{n-p-1}$ である．

証明. $Q_{p,n}$ を p 個の 0 の連続で分けられた文字列の組の集合として，次のように定義する．

$$Q_{p,n} = \{(\alpha, \beta) : \alpha 0^p \beta \in \Sigma^n, (\alpha = \varepsilon) \vee (\alpha[-1] \neq 0), (\beta = \varepsilon) \vee (\beta[0] \neq 0)\}$$

Σ^n 中の p 個の 0 の連続と $q \in Q_{p,n}$ が一対一に対応するので，$c'(p,n) = |Q_{p,n}|$ である．

例 5. $\sigma = 2, n = 3, p = 1$ について，

$$\begin{aligned} \Sigma^n &= \{\underline{\mathbf{000}}, \underline{\mathbf{00}}1, \underline{\mathbf{0}}1\underline{\mathbf{0}}, \underline{\mathbf{0}}11, 1\underline{\mathbf{00}}, 1\underline{\mathbf{0}}1, 11\underline{\mathbf{0}}, 111\} \\ c'(1,3) &= 8 \\ Q_{p,n} &= \{(\varepsilon,\varepsilon), (\varepsilon,1), (\varepsilon,10), (01,\varepsilon), (\varepsilon,11), (1,\varepsilon), (1,1), (11,\varepsilon)\} \end{aligned}$$

(1) $p \leq n-2$ のとき
$\alpha = \varepsilon$ のとき $|\beta| = n-p, \beta[0] \neq 0$ であることから，$|Q_{p,n}| = (\sigma-1)\sigma^{n-p-1}$ となる．$\beta = \varepsilon$ のとき，同様に $|Q_{p,n}| = (\sigma-1)\sigma^{n-p-1}$ である．$\alpha, \beta \neq \varepsilon$ のとき，$|\alpha| + |\beta| = n-p, \alpha[-1] \neq 0, \beta[0] \neq 0$ であるから，$|Q_{p,n}| = (n-p-1)(\sigma-1)^2\sigma^{n-p-2}$ となる．

$$\begin{aligned} c'(p,n) &= |Q_{p,n}| \\ &= 2(\sigma-1)\sigma^{n-p-1} + (n-p-1)(\sigma-1)^2\sigma^{n-p-2} \\ &= (n-p+1)\sigma^{n-p} - 2(n-p)\sigma^{n-p-1} + (n-p-1)\sigma^{n-p-2} \end{aligned}$$

(2) $p = n-1$ のとき
$|\alpha| + |\beta| = 1$ である．$\alpha = \varepsilon$ のとき，$\beta = b \neq 0$ であることから，$|Q_{p,n}| = \sigma - 1$ となる．$\beta = \varepsilon$ のとき，同様に $|Q_{p,n}| = \sigma - 1$ である．

$$c'(p,n) = |Q_{p,n}| = 2(\sigma-1)$$

(3) $p = n$ のとき
$\alpha = \beta = \varepsilon$ であるから，

$$c'(p,n) = |Q_{p,n}| = 1$$

$p \leq n-1$ で，

$$\begin{aligned} c(p,n) &= \sum_{i=p}^{n} c'(i,n) \\ &= \sum_{i=p}^{n-2} \left((n-i+1)\sigma^{n-i} - 2(n-i)\sigma^{n-i-1} + (n-i-1)\sigma^{n-i-2}\right) + 2(\sigma-1) + 1 \\ &= (n-p+1)\sigma^{n-p} - (n-p)\sigma^{n-p-1} \end{aligned}$$

$c(n,n) = c'(n,n) = 1$ も上式で表せる． □

実際に $\sigma = 2, 3$ のときの $c(p,n)$ の値を計算するとそれぞれ表 1, 2 となる．いずれの場合も $c(n+1, p+1) = c(n,p)$ となっており，この等式は任意の整数 σ について成り立つ．これは長さ n の文字列に含まれる p 個の 0 の連続それぞれが，その連続に 0 を加えた，長さ $n+1$ の文字列中の $p+1$ 個の 0 の連続に一対一に対応するためである．

表 1: $c(p,n)$ $(\sigma=2)$

n \ p	1	2	3	4	5	6
1	1					
2	3	1				
3	8	3	1			
4	20	8	3	1		
5	48	20	8	3	1	
6	112	48	20	8	3	1
7	256	112	48	20	8	3
8	576	256	112	48	20	8

表 2: $c(p,n)$ $(\sigma=3)$

n \ p	1	2	3	4	5	6
1	1					
2	5	1				
3	21	5	1			
4	81	21	5	1		
5	297	81	21	5	1	
6	1053	297	81	21	5	1
7	3645	1053	297	81	21	5
8	12393	3645	1053	297	81	21

例 6. $\sigma=2$ のとき $c(1,3)=c(2,4)=8$ の対応は次のようになる．

$$\begin{array}{lcl} \underline{\mathbf{000}} & \longleftrightarrow & \underline{\mathbf{0000}} \\ \underline{\mathbf{00}}1 & \longleftrightarrow & \underline{\mathbf{000}}1 \\ \underline{\mathbf{0}}10 & \longleftrightarrow & \underline{\mathbf{00}}10 \\ 01\underline{\mathbf{0}} & \longleftrightarrow & 01\underline{\mathbf{00}} \end{array} \qquad \begin{array}{lcl} \underline{\mathbf{0}}11 & \longleftrightarrow & \underline{\mathbf{00}}11 \\ 1\underline{\mathbf{00}} & \longleftrightarrow & 1\underline{\mathbf{000}} \\ 1\underline{\mathbf{0}}1 & \longleftrightarrow & 1\underline{\mathbf{00}}1 \\ 11\underline{\mathbf{0}} & \longleftrightarrow & 11\underline{\mathbf{00}} \end{array}$$

補題 3. d_w と w の先頭 p 文字 $w[0..p-1]$ で，w が一意に定まる．また，$w[0..p-1]$ と w 中の長さ p の任意の区間 $w[i..i+p-1]$ が一対一に対応している．

証明. d_w の定義より，$d_w[0..p-1]$ と $w[0..p-1]$ から $w[p..2p-1]$ が一意に定まる．$d_w[p..2p-1]$ と $w[p..2p-1]$ からは $w[2p..3p-1]$ が定まり，同様にして w が定まる．

$w[i..i+p-1]$ 中の $w[j]$ について，$j=px+y$ $(0\le y<p)$ とすると $w[j]$ は $w[y]$ を用いて，$w[j]=w[y]+\sum_{k=0}^{x} d_w[pk+y] \pmod{\sigma}$ と表せる． □

例 7. $\sigma=3$, $p=2$ のとき，$d_w=20000100211$ と $w[0..1]\in\Sigma^2$ で次のように w が定まる．w 中の長さ 2 の任意の区間には，Σ^2 が一度ずつ現れている．

$w[0..1]$	w
00	0020202121122
01	0121212222102
02	0222222020112
10	1000000101220
11	1101010202200
12	1202020000210
20	2010101111021
21	2111111212001
22	2212121010011

連はその長さの半分以下の周期を複数持ちうる．例えば，0101010101 は周期 2, 4 を持つ連である．補題 1, 3 から Σ^n 中の周期 p を持つ連の個数は $\sigma^p c(p,n-p)$ となるが，このような複数の周期を持つ連を重複して数えることを防ぐため，連を最小の周期でのみ数えることを考える．

補題 4. 周期 p を持つ連のうち，より短い周期 q を持たないものは $\frac{pL(p)}{\sigma^p}$ である．

証明. 補題 3 より，ある d と p，$w[0..p-1]$ から生成される文字列 w のある区間 $w[i..j]$ に存在する σ^p 個の連の根 $w[i..i+p-1]$ には Σ^p が一度ずつ現れる．同一の区間が異なる周期 p, q を持つとき p と q の最大公約数 $\gcd(p,q)$ もまた周期となるため [5]，根がプリミティブであるとき $w[i..j]$ は $q < p$ となる周期 q を持たない．長さ p のプリミティブな文字列の個数は $pL(p)$ である． □

$\sigma = 2$ について $L(p)$ と σ^p は下表のようになる．プリミティブでない文字列は少なく，特に p が素数であるとき $w = u^n$ と表せるのは $|u| = 1$ の場合のみで，$\sigma^p - L(p) = \sigma$ となる．

p	1	2	3	4	5	6	7	8	9	10	11	12
$pL(p)$	2	2	6	12	30	54	126	240	504	990	2046	4020
σ^p	2	4	8	16	32	64	128	256	512	1024	2048	4096

補題 1, 2, 3, 4 より，長さ n の文字列に含まれる連の総数 $\sigma^n r(n)$ について次式を得る．

$$\begin{aligned}
\sigma^n r(n) &= \sum_{p=1}^{\frac{n}{2}} pL(p)c(p, n-p) \\
&= \sum_{p=1}^{\frac{n}{2}} pL(p)\left((n-2p+1)\,\sigma^{n-2p} - (n-2p)\,\sigma^{n-2p-1}\right) \\
&= \sum_{p=1}^{\frac{n}{2}} \sum_{d|p} \mu\left(\frac{p}{d}\right)\sigma^d\left((n-2p+1)\,\sigma^{n-2p} - (n-2p)\,\sigma^{n-2p-1}\right)
\end{aligned}$$

定理 1. 任意の整数 n について，長さ n の文字列の平均連数 $r(n)$ は次の等式で表わされる．

$$r(n) = \sum_{p=1}^{\frac{n}{2}} pL(p)\left((n-2p+1)\,\sigma^{-2p} - (n-2p)\,\sigma^{-2p-1}\right)$$

$\sigma = 2, 3, \ldots, 6$ について，$r(n)$ を図示すると図 1 となる．この傾き $\frac{r(n)}{n}$ は図 2 となり，n が増加するにしたがい一定の値に漸近している．

定理 2. 文字列の長さあたりの平均連数 $\frac{r(n)}{n}$ は $n \to \infty$ で次の値に収束する．

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

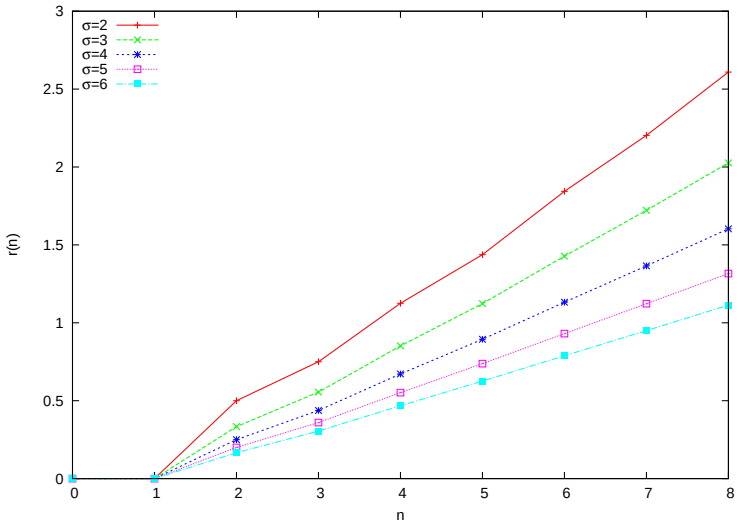


図 1: $r(n)$

証明. $r(n)$ を $f(p) = (n-2p+1)\sigma^{-2p} - (n-2p)\sigma^{-2p-1}$ とおいて計算すると次のようになる.

$$
\begin{aligned}
r(n) &= \sum_{p=1}^{\frac{n}{2}} pL(p)\left((n-2p+1)\,\sigma^{-2p} - (n-2p)\,\sigma^{-2p-1}\right) \\
&= \sum_{p=1}^{\frac{n}{2}} \sum_{d|p} \mu\left(\frac{p}{d}\right)\sigma^d f(p) \\
&= \quad \mu(1)\sigma^1 f(1) \\
&\quad +\mu(2)\sigma^1 f(2)+\mu(1)\sigma^2 f(2) \\
&\quad +\mu(3)\sigma^1 f(3) \qquad\qquad +\mu(1)\sigma^3 f(3) \\
&\quad +\mu(4)\sigma^1 f(4)+\mu(2)\sigma^2 f(4) \qquad\qquad +\mu(1)\sigma^4 f(4) \\
&\quad +\mu(5)\sigma^1 f(5) \qquad\qquad\qquad\qquad +\mu(1)\sigma^5 f(5) \\
&\quad +\mu(6)\sigma^1 f(6)+\mu(3)\sigma^2 f(6)+\mu(2)\sigma^3 f(6) \qquad\qquad +\mu(1)\sigma^6 f(6) \\
&\qquad\qquad\qquad\qquad \vdots
\end{aligned}
$$

$\mu(d)$ で括ると

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

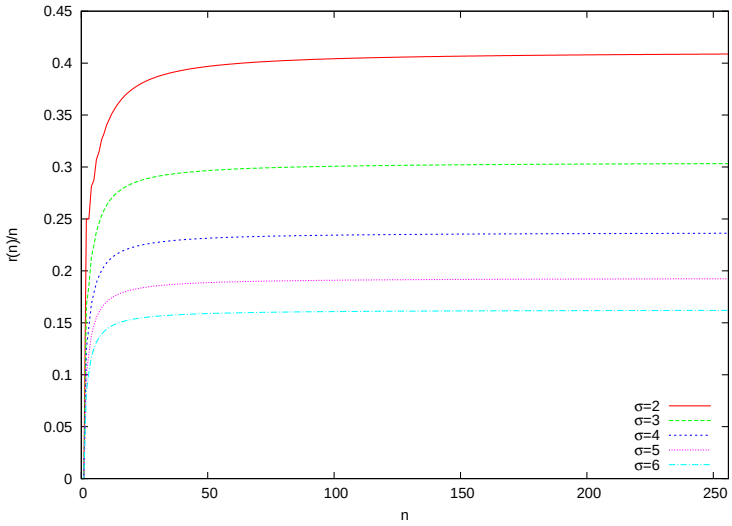


図 2: $\frac{r(n)}{n}$

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

$n\to\infty$ で $\sigma^{(1-2d)\lfloor\frac{n}{2d}\rfloor}\to 0$ であり,

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

□

$\sigma = 2, 3, \ldots, 6$ について，$\lim_{n\to\infty} \frac{r(n)}{n} = \sum_{d=1}^{\infty} \mu(d) \frac{\sigma-1}{\sigma^{2d}-\sigma}$ は次のような値になる．

σ	$\lim_{n\to\infty} \frac{r(n)}{n}$
2	0.41165
3	0.30491
4	0.23736
5	0.19329
6	0.16268

3.2 指数の和の平均

補題 5. $c_e(p,n) = \frac{1}{p}\left(p(n-p+1)\sigma^{n-p} - (p-1)(n-p)\sigma^{n-p-1}\right)$ である．

証明. 定義より $c_e(p,n)$ は $c'(p,n)$ を用いて $\sum_{i=p}^{n} \frac{ic'(i,n)}{p}$ と表せる．
$p \le n-1$ のとき，

$$\begin{aligned} c_e(p,n) &= \sum_{i=p}^{n} \frac{ic'(i,n)}{p} \\ &= \sum_{i=p}^{n-2} \frac{i}{p}\left((n-i+1)\sigma^{n-i} - 2(n-i)\sigma^{n-i-1} + (n-i-1)\sigma^{n-i-2}\right) \\ &\quad + \frac{n-1}{p} 2(\sigma-1) + \frac{n}{p} \\ &= \frac{1}{p}\left(p(n-p+1)\sigma^{n-p} - (p-1)(n-p)\sigma^{n-p-1}\right) \end{aligned}$$

$c_e(n,n) = c'(n,n) = 1$ も上式で表せる． □

定理 3. 長さ n の文字列に含まれる連の指数の和の平均 $e(n)$ は以下の等式で表わされる．

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

証明. d_w に含まれる l 個の 0 の連続は $l \ge p$ であるとき，w の長さ $l+p$ の連と対応する．この連の指数は $\frac{l}{p}+1$ であるため，Σ^n に含まれる周期 p の連の指数の総和は $pL(p)\left(c_e(p,n) + c(p,n)\right)$ と表せる．

$$\begin{aligned} \sigma^n e(n) &= \sum_{p=1}^{\frac{n}{2}} pL(p)\left(c_e(p,n-p) + c(p,n-p)\right) \\ &= \sum_{p=1}^{\frac{n}{2}} L(p)\left(2p(n-2p+1)\sigma^{n-2p} - (2p-1)(n-2p)\sigma^{n-2p-1}\right) \end{aligned}$$

□

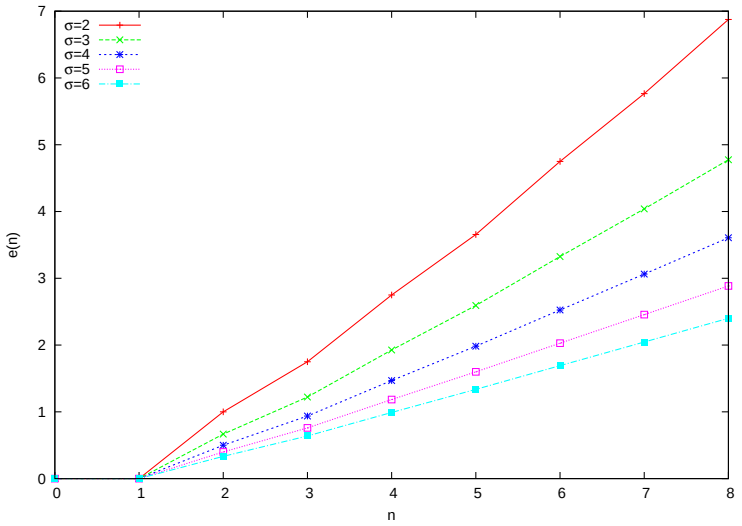


図 3: $e(n)$

定理 4. 文字列の長さあたりの指数の和の平均 $\frac{e(n)}{n}$ は $n \to \infty$ で次の値に収束する.

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

証明.

$$\begin{aligned}
\sigma^n e(n) &= \sum_{p=1}^{\frac{n}{2}} pL(p) \left(c(p, n-p) + c_e(p, n-p) \right) \\
&= \sum_{p=1}^{\frac{n}{2}} \sum_{d|p} \mu\left(\frac{p}{d}\right) \sigma^d \frac{1}{p} \left(2p(n-2p+1)\sigma^{n-2p} - (2p-1)(n-2p)\sigma^{n-2p-1} \right) \\
&= \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\lfloor \frac{n}{2d} \rfloor} \frac{1}{pd} \left(2pd(n-2pd+1)\sigma^{n-2pd+p} - (2pd-1)(n-2pd)\sigma^{n-2pd+p-1} \right) \\
\frac{e(n)}{n} &= \sum_{d=1}^{\frac{n}{2}} \mu(d) \sum_{p=1}^{\lfloor \frac{n}{2d} \rfloor} \frac{1}{npd} \left(2pd(n-2pd+1)\sigma^{-2pd+p} - (2pd-1)(n-2pd)\sigma^{-2pd+p-1} \right) \\
\lim_{n\to\infty} \frac{e(n)}{n} &= \sum_{d=1}^{\infty} \mu(d) \left(\frac{2(\sigma-1)}{\sigma^{2d}-\sigma} + \frac{1}{d\sigma} \ln \left(\frac{\sigma^{2d}}{\sigma^{2d}-\sigma} \right) \right)
\end{aligned}$$

□

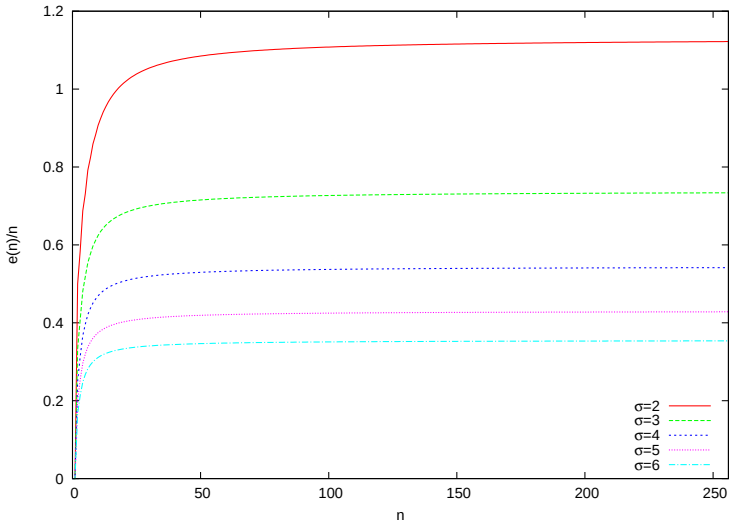


図 4: $\frac{e(n)}{n}$

$\sigma = 2, 3, \ldots, 6$ について，$e(n)$ と $\frac{e(n)}{n}$ はそれぞれ図 3 と 4 のようになり，$\lim_{n\to\infty} \frac{e(n)}{n}$ は下表の値に収束する．

σ	$\lim_{n\to\infty} \frac{e(n)}{n}$
2	1.13103
3	0.73822
4	0.54459
5	0.43039
6	0.35536

4　まとめ

長さ n の文字列に含まれる連の個数の平均 $r(n)$ と連の指数の和の平均 $e(n)$，および $n \to \infty$ での極限を導出した．

本研究で長さ n の文字列の集合 Σ^n に含まれる連の総数を求めることができたが，今後はそれが Σ^n の中でどのように分布しているのか，ひいては長さ n の文字列が含む連の最大個数について研究したい．

また，例えばリンドン文字列などの，特定の性質を持った文字列に含まれる連の平均個数や連の指数の和の平均についても調べたい．

5　謝辞

本研究を行うにあたって多大なご指導を頂きました篠原　歩教授，石野　明助教，篠原研究室の方々，そして学生生活を暖かく見守ってくれた両親に深く感謝いたします．

参考文献

[1] M. Crochemore and L. Ilie. Analysis of Maximal Repetitions in Strings. *LECTURE NOTES IN COMPUTER SCIENCE*, 4708:465, 2007.

[2] F. Franek and Q. Yang. An asymptotic lower bound for the maximal-number-of-runs function. In J. Holub and J. Žďárek eds., *Proceedings of the Prague Stringology Conference '06*, pp. 3–8, Czech Technical University in Prague, Czech Republic, 2006.

[3] R. Kolpakov and G. Kucherov. On the sum of exponents of maximal repetitions in a word.

[4] R. Kolpakov and G. Kucherov. Finding maximal repetitions in a word in linear time. *Proceedings of the 1999 Symposium on Foundations of Computer Science, New York (USA). IEEE Computer Society, October*, pp. 17–19, 1999.

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

[6] M. Lothaire. *Applied Combinatorics on Words*. Cambridge University Press, 2005.

[7] L. I. M. Crochemore and L. Tinta. The "runs" conjecture. http://www.csd.uwo.ca/~ilie/runs.html.

[8] W. Rytter. The number of runs in a string: improved analysis of the linear upper bound. *Proceedings of the 23rd Symposium on Theoretical Aspects of Computer Science, B. Durand and W. Thomas, eds., Lecture Notes in Comput. Sci*, 2884:184–195.

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

[10] W. F. S. Simon J. Puglisi, Jamie Simpson. How many runs can a string contain? *Theoretical Computer Science Accepted*, 2007.