ネックレス文字列中の繰り返し構造について

草野　一彦*　　篠原　歩*

2009 年 7 月 22 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 を持つ繰り返しと呼ぶ．連とは左右に延長不可能な繰り返しである．本稿では文字列の両端を繋いで輪にしたネックレス文字列における連を考え，任意の長さと任意のアルファベットサイズのネックレス文字列に含まれる連の平均個数を示す．

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．中でも左右に延長不可能な繰り返しは連と呼ばれ，盛んに研究が行われている．

Kolpakov と Kucherov はある長さ n の文字列が含みうる連の最大個数が高々cn であることを示した [6]．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．現在，Crochemore らによって $c \leq 1.029$ であることが示されており [2, 3]，$c < 1$ が予想されている．一方，具体的な文字列や文字列の構成法を示すことで最大個数の下限を与えようという試みもなされている．現在，Simpson と松原らがそれぞれ文字列の長さに対する連の個数の比が $0.94457571235\ldots$ に漸近する文字列を与えている [5, 8]．連の最大個数の厳密な値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厳密に与える次の式を示した [10]．

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

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．

文字列に含まれる連の繰り返し回数（指数）の和も研究されている．Crochemore らによって長さ n の文字列に含まれる連の指数和の最大値が $2.9n$ 未満であることが示されている [1]．草野らは連の指数和の平均を厳密に与える次の式を示した [7]．

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

ここで，$L(p)$ はリンドン文字列の個数であり，$L(p) = \sum_{d|p} \mu(d) \sigma^{\frac{p}{d}}$ である．

本稿では文字列の両端を繋いで輪にしたネックレス文字列における連を考える（図 1）．連の最大個数の下限の更新においては，連の多い文字列を繰り返すことでさらに連を増やしている [9]．ネックレス文字列における連は繰り返しによって増える連の個数に対応している（補題 2）．また，ネックレス文字列には端が無いため連の解析が容易になる可能性があり，ネックレス文字列における連の解析は通常の文字列中の連の解析に繋がることが期待される．

本稿では長さ n の文字列の両端を繋いだネックレス文字列に含まれる連の平均個数と長さ n のネックレス文字列に含まれる連の平均個数がそれぞれ以下の値となることを示す．

*東北大学大学院情報科学研究科

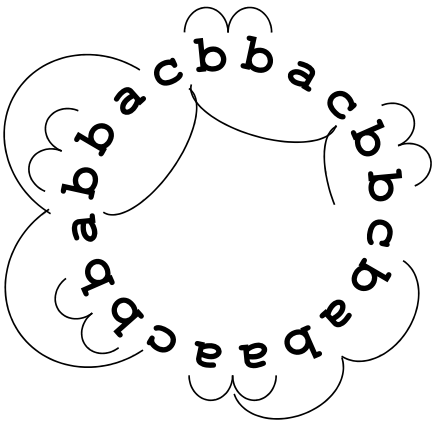

図 1: ネックレス文字列 ⟨bacbbcbabaacbbabbacb⟩ と含まれる連

$$\tau_1(n) = \sum_{p=1}^{n-1} n(\sigma-1)\sigma^{-2p-1} \sum_{d|p} \mu\left(\frac{p}{d}\right)\sigma^d \delta(d, 2p-n)$$

$$\tau_2(n) = \frac{\sum_{d|n} \phi\left(\frac{n}{d}\right)\sigma^d \frac{n}{d}\tau_1(d)}{\sum_{d|n} \phi\left(\frac{n}{d}\right)\sigma^d}$$

ここで $\phi(n)$ はオイラーのトーティエント関数,

$$\delta(d,n) = \begin{cases} 1 & d > n \text{ and } d \not\equiv 0 \pmod{d-n} \\ 0 & \text{elsewhere} \end{cases}$$

である.

2 定義

2.1 文字列

$\Sigma = \{a, b, c, \dots\}$ を有限のアルファベットとし，そのサイズ $|\Sigma|$ を σ で表す．Σ 上の文字列の集合を Σ^* とし，長さ n の文字列の集合を Σ^n とする．文字列 w の i 番目の文字を $w[i]$ と表す．文字列 w の部分文字列 $w[i]w[i+1]\dots w[j]$ を $w[i..j]$ と表す.

文字列 w において，任意の $1 \le i \le |w| - p$ で $w[i] = w[i+p]$ が成り立つとき，p を文字列の**周期**という．部分文字列 $w[i..j]$ が周期 $p \le \frac{j-i+1}{2}$ を持つ，すなわち 2 回以上の繰り返しであるとき，$w[i..j]$ は**周期的**であるという．部分文字列が $w[i..j]$ 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**であるという.

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

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，$w[i..j]$ を連と呼ぶ．周期性と延長不可能性を満たす最小の周期 p を持つ連 $w[i..j]$ について，$w[i..i+p-1]$ を**根**といい，最小の周期に対する連の長さの比 $\frac{j-i+1}{p}$ を**指数**という．文字列 w が文字列 u と整数 $k \ge 2$ を用いて，$w = u^k$ と表せないとき，w は**素である**という．連の根は常に素である．文字列 w が含む連の個数を $run(w)$ で表す.

2.2 ネックレス文字列

文字列 w の両端を繋いで輪にしたものを**ネックレス文字列**と呼び，$\langle w \rangle$ と書く．ネックレス文字列は反転したものは区別するが，回転したものは区別しない．すなわち，⟨abc⟩ と ⟨acb⟩ は異なるネックレス文字列であるが，⟨abc⟩ と ⟨bca⟩ は同一のネックレス文字列である．長さ n のネックレス文字列の集合を NL_n とする．ネックレス文字列 $\langle w \rangle$ 上の位置を元になった文字列 w の対応する位置を用いて $w[i]$ のように表す．w の長さ $|w|$ をネックレス文字列の長さ $|\langle w \rangle|$ とする.

長さ n のネックレス文字列 $\langle w \rangle$ 上の部分文字列 $w[i..j]$ を $w[i\%n]w[(i+1)\%n]\dots w[j\%n]$ と定義する．ここで $i\%n$ は 1 以上で $x \equiv i \pmod{n}$ を満たす最小の x を表す．部分文字列 $w[i..j]$ は（ネックレスではない）文字列であり，ネックレス文字列 $\langle w \rangle$ 以上の長さにもなりうる．ネックレス文字列 $\langle w \rangle$ 上の部分文字列 $w[i..j]$ 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であるという.

$$\begin{array}{lcl} w[(i-1)\%n] & \neq & w[(i+p-1)\%n] \\ w[(j+1)\%n] & \neq & w[(j-p+1)\%n] \end{array}$$

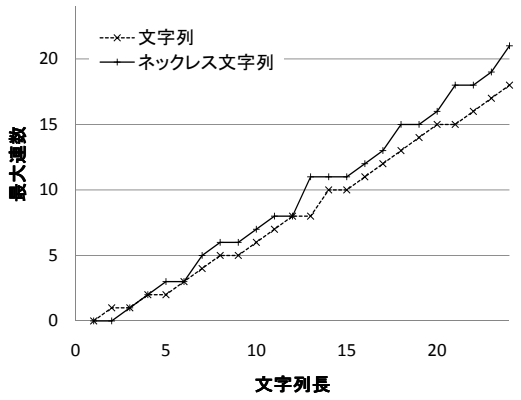

図 2: 連の最大個数

ネックレス文字列の周期性を文字列と同様に定義し，周期性と延長不可能性を満たすものを，ネックレス文字列上の連と呼ぶ．本稿では，ネックレス文字列中の延長不可能性を満たさない繰り返しを連とはみなさない．例えば，ネックレス文字列 $\langle$abab$\rangle$ において，繰り返し abab は . . . abababab . . . と延長できるため連ではない．ネックレス文字列 $\langle w\rangle$ が含む連の個数も $run(\langle w\rangle)$ で表す．

図 2 に，長さ n の文字列とネックレス文字列が含む連の最大個数を示す．これは計算機で文字列とネックレス文字列を網羅的に生成することによって得られたものである．$n \leq 24$ では，$n = 1$ のときを除いて，ネックレス文字列は文字列以上の個数の連を含むことができる．

補題 1. 長さ n のネックレス文字列中に周期 $p \geq n$ を持つ連は存在しない．

証明. 背理法によって示す．長さ n のネックレス文字列中に周期 $p > n$ を持つ連が存在すると仮定し，そのような連を持つネックレス文字列を w とする．周期 p が文字列長よりも長いため任意の位置 $1 \leq i \leq n$ で $w[i] = w[(i+p)\%n]$ が成り立つが，これは連の延長不可能性，すなわち $w[(i-1)\%n] \neq w[(i+p-1)\%n]$ を満たす i が存在することに矛盾している． □

長さ n のネックレス文字列中に周期 $p \geq n$ を持つ連は存在しないため，以降は n より短い周期のみを考える．一方，ネックレスの長さ n より長い連は存在しうる．例えば，長さ 5 ネックレス文字列 $\langle$abaab$\rangle$ は長さ 6 の連 abaaba を含んでいる．

ネックレス文字列が含む連の個数は元になった文字列を繰り返したときの連の個数に対応している．

補題 2. 任意の文字列 w，整数 $k \geq 2$ について以下の等式が成り立つ．

$$run(\langle w\rangle) = run\left(w^{k+1}\right) - run\left(w^{k}\right)$$

証明. 長さ n のネックレス文字列 $\langle w\rangle$ が含む連を $w[i..j]$ $(1 \leq i < j)$ とし，この連の最小の周期を p とする．この連に対応する連が w^k の末尾に w を連結したときに増えることを示す．$j \leq n$ ならば連 $w^k w[nk+i..nk+j]$ が新たに連となる．$j > n$ のとき，$w[i..n]$ が 2 回以上の繰り返しを含む，すなわち $i + 2p \leq n$ ならば，連 $w^k w[nk+i..n(k+1)]$ が増える．$j > n$ かつ $i + 2p > n$ の場合，$w^k[n(k-1)+i..nk]$ は繰り返し回数が 2 回に満たないため連ではないが，末尾に w を連結することにより，$w^k w[n(k-1)+i..n(k-1)+j]$ が連となる．よって，$run(\langle w\rangle) \leq run\left(w^{k+1}\right) - run\left(w^{k}\right)$．

長さ w の文字列の k 回繰り返し w^k の末尾に w を連結することによって増える連 $w^k w[i..j]$ を考える．このとき連 $w^k w[i..j]$ は元の文字列 w^k に 2 回以上の繰り返しを含まず，連結した w 内で終了する．すなわち，$nk < i + 2p \leq j$ である．$j < n(k+1)$ であれば，$\langle w\rangle$ は連 $w[i..j]$ を持つ．$j = n(k+1)$ であるとき，$p < n, k \geq 2$ より $\langle w\rangle$ には $w[i] \neq w[(i+p)\%n]$ となる位置 i が存在するため，$\langle w\rangle$ 中の繰り返し $w[i..j]$ は長くともその位置以降に右に延長することはできず，$w[i..j]$ の最大延長は連となる．よって，$run(\langle w\rangle) \geq run\left(w^{k+1}\right) - run\left(w^{k}\right)$ であり，$run(\langle w\rangle) = run\left(w^{k+1}\right) - run\left(w^{k}\right)$． □

2.3 メビウス関数とオイラーのトーティエント関数

メビウス関数は整数 $n > 0$ にたいして次のように

定義される関数である．

$$\mu(n) = \begin{cases} 0 & n \text{ が平方因子を持つ} \\ (-1)^k & n \text{ が } k \text{ 個の素因数を持つ} \end{cases}$$

関数 $f(n)$ と $g(n)$ について以下の 2 つの等式は同値であることが知られている（メビウスの反転公式）．

$$\begin{aligned} f(n) &= \sum_{d|n} g(d) \\ g(n) &= \sum_{d|n} \mu\left(\frac{n}{d}\right) f(d) \end{aligned}$$

$d|n$ という表記は d が n の約数であることを示す．

オイラーのトーティエント関数 $\phi(n)$ は整数 $n > 0$ にたいして，1 から n までの整数のうち n と互いに素なものの個数を表す．トーティエント関数はメビウス関数を用いて次のように表せることが知られている [4]．

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

例 1. $n \leq 8$ にたいして $\mu(n)$ と $\phi(n)$ の値を示す．

n	1	2	3	4	5	6	7	8
$\mu(n)$	1	−1	−1	0	−1	1	−1	0
$\phi(n)$	1	1	2	2	4	2	6	4

3 ネックレス文字列中の連の平均個数

長さ n の文字列 $w \in \Sigma^n$ について $\langle w \rangle$ が含む連の平均個数と，長さ n のネックレス文字列が含む連の平均個数は一致するとは限らない．

例 2. $\sigma = 2$ のとき，$w \in \Sigma^4$ の両端を繋いだネックレス文字列はそれぞれ次の個数の連を持つ．

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

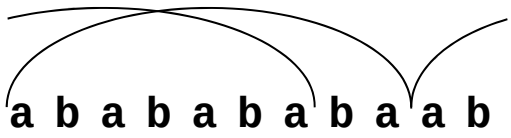

図 3: $w[1]$ から開始する周期 9 を持つ連を含む文字列の例．周期性の制約から $w[1..9]$ は $w[10..11]$ の繰り返しとなっている．

よって $w \in \Sigma^4$ について $\langle w \rangle$ が持つ連の平均個数は $\frac{16}{16} = 1$ である．$\langle$abab$\rangle = \langle$baba$\rangle$ などは同一のネックレス文字列である．長さ 4 のネックレス文字列 NL_4 と含まれる連数を列挙すると次のようになる．

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

連の平均個数は $\frac{4}{6} = \frac{2}{3}$ である．

本節では，まず長さ n の文字列の両端を繋いだときの連の平均個数 $\tau_1(n) = \frac{1}{\sigma^n} \sum_{w \in \Sigma^n} run(\langle w \rangle)$ を求め，そこから長さ n のネックレス文字列が含む連の平均個数 $\tau_2(n) = \frac{1}{|NL_n|} \sum_{w \in NL_n} run(w)$ を導く．

3.1 文字列の両端を繋いだとき

長さ n の文字列 $w \in \Sigma^n$ のうち，両端を繋いでネックレス文字列 $\langle w \rangle$ としたときに $w[1]$ から開始し周期 p を持つ連を含む文字列の集合を $W(n,p) \subseteq \Sigma^n$ とする．

連の延長不可能性と周期性から，$w \in W(n,p)$ は次の条件を満たす．

$$w[n] \neq w[p] \quad (1)$$

$$w[i] = w[(i+p)\%n] \quad (1 \leq i \leq p) \quad (2)$$

連の長さが n より長いとき，周期性に関する条件は繰り返し適用される．このとき，$w[1..p-1]$ は $w[p..n]$ の繰り返しとなる（図 3）．

補題 3. $w \in W(n,p)$ と整数 $1 \le i \le n-p$，整数 k について，以下の等式が成り立つ．

$$w[k(n-p)+i] = w[p+i] \quad (1 \le k(n-p)+i \le p)$$

証明. k に関する数学的帰納法で示す．$k=0$ のとき，条件 (2) より $w[i] = w[p+i]$ である．$w[k(n-p)+i] = w[p+i]$ が成り立つと仮定すると，$(k+1)(n-p)+i \le q$ ならば，条件 (2) より，$w[(k+1)(n-p)+i] = w[((k+1)(n-p)+i+p)\%n] = w[k(n-p)+i] = w[p+i]$ となる．□

長さ n の全ての文字列 $w \in \Sigma^n$ について，w をネックレス文字列 $\langle w \rangle$ としたときに $\langle w \rangle$ に含まれる $w[1]$ から開始し周期 p を持つ連の個数を $cm\,(n,p)$ とする．$\langle w \rangle$ が持つこのような連は高々1 つなので，$cm\,(n,p) = |W(n,p)\,|$ である．

補題 4. 任意の整数 $n,\ p$ について，以下の等式が成り立つ．

$$cm\,(n,p) = \begin{cases} (\sigma-1)\sigma^{n-p-1} & p < n \ \textit{and} \\ & p \not\equiv 0 \pmod{n-p} \\ 0 & \textit{elsewhere} \end{cases}$$

証明. 条件 $(1,2)$ を満たす文字列を列挙することを考える．$p \ge n$ の場合は補題 1 より周期 p を持つ連が無いため，$cm\,(n,p) = 0$ である．$p \equiv 0 \pmod{n-p}$ ならば，補題 3 において $i = n-p$ とすることで $w[n] = w[(k+1)(n-p)] = w[p]$ を得る．これは条件 1 に矛盾するためそのような文字列を作ることはできず，$cm\,(n,p) = |W(n,p)\,| = 0$ となる．$p \not\equiv 0 \pmod{n-p}$ ならば，$w[p+1..n-1]$ を任意に定め，$w[n] \neq w[p] = w[p\%(n-p)+p]$ となるよう $w[n]$ を決めることで，補題 3 から $w[1..p]$ が定まり，条件 $(1,2)$ 満たす文字列を全て列挙することができる．このような文字列の選び方は $\sigma^{(n-1)-(p+1)+1}(\sigma-1) = (\sigma-1)\sigma^{n-p-1}$ 通りであるから，$cm\,(n,p) = |W(n,p)\,| = (\sigma-1)\sigma^{n-p-1}$ である．□

長さ n の文字列 $w \in \Sigma^n$ のうち，両端を繋いでネックレス文字列 $\langle w \rangle$ としたときに $w[1]$ から開始し最小の周期として p を持つ連を含むものの個数を $c\,(n,p)$ とする．このうち連の指数が t 以上であるものを $c\,(n,p,t)$ とする．連の定義から連の指数は 2 以上なので，$c\,(n,p) = c\,(n,p,2)$ である．

補題 5. 任意の $n,\ p,\ t \ge 2$ について，以下の等式が成り立つ．

$$c\,(n,p,t) = c\,(n-p(t-2),p)$$

証明. 長さ n の文字列 $w \in \Sigma^n$ のうち $w[1]$ から開始し最小の周期 p と t 以上の指数を持つ連を含む文字列の集合を $V\,(n,p,t) \subseteq \Sigma^n$ とする．$V\,(n,p,t)$ は連の延長不可能性と周期性からそれぞれ以下の条件を満たす．

$$\begin{aligned} w[n] &\neq w[p] \\ w[i] &= w[(i+kp)\%n] \\ &\qquad (1 \le i \le p,\ 1 \le k \le t-1) \end{aligned}$$

p が最小の周期であることから，連の根 $w[1..p]$ は素である．

$V_1 = V\,(n,p,t)$ と $V_2 = V\,(n-p(t-2),p,2)$ を考える．V_1 と V_2 の要素には一対一対応が存在する．

$v_1 \in V_1$ にたいして $v_1' = v_1[p(t-2)+1..n]$ とすると，$|v_1'| = n-p(t-2)$ である．$v_1'[n-p(t-2)] = v_1[n] \neq v_1[p] = v_1[p(t-1)] = v_1'[p]$ であり，$1 \le i \le p$ で $v_1'[i] = v_1[i+p(t-2)] = v_1[i+p(t-1)] = v_1'[(i+p)\%(n-p(t-2))]$ である．また，$v_1'[1..p] = v_1[p(t-2)+1..p(t-1)] = v_1[1..p]$ は素なので，$v_1' \in V_2$ となる．

$v_2 \in V_2$ にたいして $v_2' = v_2[1..p]^{t-1}v_2[p+1..n-p(t-2)]$ とすると，$|v_2'| = n$ である．$v_2'[n] = v_2[n-p(t-2)] \neq v_2[p] = v_2'[p]$ である．$1 \le i \le p,\ 1 \le k \le t-2$ で $v_2'[i+kp] = v_2[i] = v_2'[i]$ であり，$k = t-1$ のとき $v_2'[(i+(t-1)p)\%n] = v_2[(i+p)\%(n-p(t-2))] = v_2[i] = v_2'[i]$ である．また，$v_2'[1..p] = v_2[1..p]$ は素なので，$v_2' \in V_1$ となる．□

周期 p を持つ連は唯一の最小の周期 $q \le p$ を持つ．この時，連の長さは $2p$ 以上であるから，連の指数は $\frac{2q}{p}$ 以上である．よって，$cm\,(n,p)$ は $c\,(n,p,t)$ を用

いて次のように表せる．

$$cm\left(n,p\right)=\sum_{d|p}c\left(n,d,\frac{2d}{p}\right)$$

この等式をメビウスの反転公式によって変形することで，$c\left(n,p\right)$ が求められる．

補題 6. 任意の整数 $n,\ p$ について，以下の等式が成り立つ．

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

ただし，

$$\delta\left(d,n\right)=\begin{cases}1 & d>n\ and\ d\not\equiv 0\ (\mathrm{mod}\ d-n)\\ 0 & elsewhere\end{cases}$$

である．

証明.

$$\begin{aligned}cm\left(n,p\right) &= \sum_{d|p}c\left(n,d,\frac{2d}{p}\right)\\ &= \sum_{d|p}c\left(n-2p+2d,d\right)\end{aligned}$$

$n'=n-2p$ とおく

$$\begin{aligned}cm\left(n'+2p,p\right) &= \sum_{d|p}c\left(n'+2d,d\right)\end{aligned}$$

メビウスの反転公式

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

□

ネックレス文字列の回転に対する対称性から，長さ n の全ての文字列 $w\in\Sigma^{n}$ に対して，$\langle w\rangle$ が含む最小の周期 p を持つ連の個数は $n\,c\left(n,p\right)$ である．これを周期 $1\le p\le n-1$ について足し合わせることで連の総数が求められ，$|\Sigma^{n}|=\sigma^{n}$ であるから，連の平均個数を導ける．

定理 1. 長さ n の文字列 $w\in\Sigma^{n}$ の両端を繋いでネックレス文字列 $\langle w\rangle$ としたときの，連の平均個数は次の式で表される．

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

3.2 ネックレス文字列

$\tau_{1}\left(n\right)$ と $\tau_{2}\left(n\right)$ が一致しないのは，$w=\mathtt{ababab}=(\mathtt{ab})^{3}$ のように他の文字列の繰り返しである文字列は文字列とその循環が $n=|w|$ 未満の個数しかないためである．

文字列 w を k 回繰り返した文字列の両端を繋いだネックレス文字列 $\left\langle w^{k}\right\rangle$ が含む連の個数は $run\left(\langle w\rangle\right)$ を用いて表せる．

補題 7. 任意の文字列 w と整数 $k\ge 1$ について，以下の等式が成り立つ．

$$run\left(\left\langle w^{k}\right\rangle\right)=k\ run\left(\langle w\rangle\right)$$

証明. 補題 2 より任意の整数 $t\ge 2$ を用いて，

$$\begin{aligned}run\left(\left\langle w^{k}\right\rangle\right) &= run\left(w^{kt+k}\right)-run\left(w^{kt}\right)\\ &= \sum_{i=0}^{k-1}\left(run\left(w^{kt+i+1}\right)-run\left(w^{kt+i}\right)\right)\\ &= k\ run\left(\langle w\rangle\right).\end{aligned}$$

□

長さ n のネックレス文字列の個数 $|NL_{n}|$ を考える．

補題 8. 任意の整数 n について，以下の等式が成り立つ [4].

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

ネックレス文字列の個数 $|NL_{n}|$ を求めた手法 [4] と同様にして，長さ n の全ての文字列 Σ^{n} の各要素の両端を繋いだ際の連の総数 $SR\left(n\right)=\sigma^{n}\tau_{1}\left(n\right)$ から，

長さ n の全てのネックレス文字列 NL_n に含まれる連の総数 $NR(n) = |NL_n|\,\tau_2(n)$ を導く．この手法では全ての長さ n のネックレス文字列を n 通りに切って文字列としたものを考える．

例 3. $\sigma = 2$ のとき NL_4 の各要素を 4 通りに切って文字列にすると次のようになる．この中には Σ^4 の各要素が少なくとも 1 回現われている．

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

補題 9. 任意の整数 n について，以下の等式が成り立つ．

$$NR(n) = \frac{1}{n}\sum_{d|n}\phi\left(\frac{n}{d}\right)\frac{n}{d}SR(d)$$

証明. 長さ n の全てのネックレス文字列 NL_n を n 通りに切った文字列の多重集合を T とする．$|T| = n\,|NL_n|$ であり，T が含む連の総数は $n\,NR(n)$ である．T に含まれる文字列 $w \in \Sigma^n$ の個数は，$w = w[k+1..n]w[1..k]$ $(0 \le k < n)$ を満たす k の個数と等しい．k を与えたときに $w = w[k+1..n]w[1..k]$ となる $w \in \Sigma^n$ の個数を考える．等式が成り立つのは $u \in \Sigma^{\gcd(k,n)}$ として，$w = u^{\frac{n}{|u|}}$ と表せるときである．よって，補題 7 より，

$$n\ NR(n) = \sum_{k=0}^{n-1}\frac{n}{\gcd(k,n)}SR(\gcd(k,n))\,.$$

$\gcd(k,n)$ は n の約数であるから，この式は $d = \gcd(k,n)$ として次のように変形できる．

$$\begin{aligned}
NR(n) &= \frac{1}{n}\sum_{d|n}\frac{n}{d}SR(d)\sum_{k=0}^{n-1}[d = \gcd(k,n)] \\
&= \frac{1}{n}\sum_{d|n}\frac{n}{d}SR(d)\sum_{k=0}^{n-1}\left[\frac{k}{d} \perp \frac{n}{d}\right] \\
&= \frac{1}{n}\sum_{d|n}\frac{n}{d}SR(d)\sum_{k=0}^{\frac{n}{d}-1}\left[k \perp \frac{n}{d}\right] \\
&= \frac{1}{n}\sum_{d|n}\frac{n}{d}SR(d)\,\phi\left(\frac{n}{d}\right)
\end{aligned}$$

ここで，演算子 $x \perp y$ は x と y が互いに素であることを表す． □

補題 8,9 から定理を導ける．

定理 2. 任意の整数 n について，長さ n のネックレス文字列が含む連の平均個数 $\tau_2(n)$ は次のように表される．

$$\begin{aligned}
\tau_2(n) &= \frac{NR(n)}{|NL_n|} \\
&= \frac{\sum_{d|n}\phi\left(\frac{n}{d}\right)\sigma^d\frac{n}{d}\tau_1(d)}{\sum_{d|n}\phi\left(\frac{n}{d}\right)\sigma^d}
\end{aligned}$$

アルファベットサイズ $\sigma = 2, \sigma = 3$ について $\tau_1(n)$ と $\tau_1(n)$ および通常の文字列中の連の平均個数 $r(n)$ の値をそれぞれ図 4 と図 5 に示す．

4 まとめ

本稿では文字列の両端を繋いだ際の連の平均個数 $\tau_1(n)$ とネックレス文字列が含む連の平均個数 $\tau_2(n)$ を任意の文字列長と任意のアルファベットサイズについて厳密に示した．今後の課題はネックレス文字列中の連の指数和の平均やちょうど 2 回の繰り返しであるスクエアの平均個数を調べ，連の最大個数の解明に繋げることである．

参考文献

[1] 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.

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

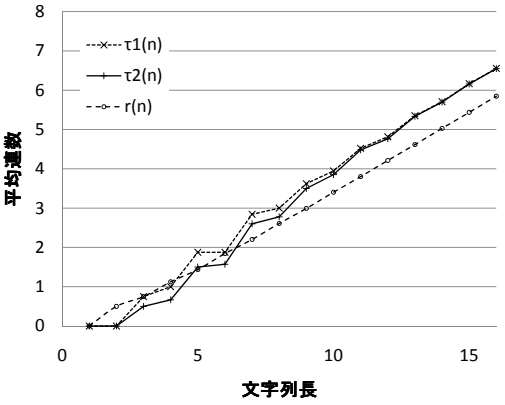

図 4: 連の平均個数（$\sigma = 2$）

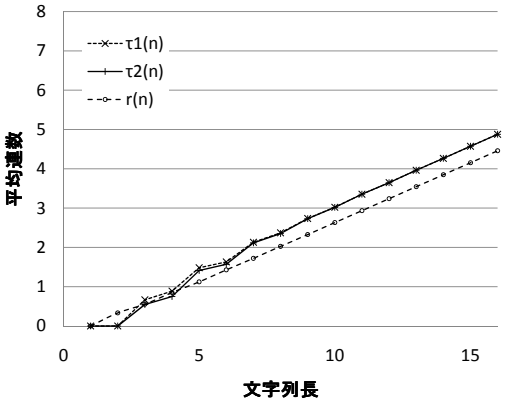

図 5: 連の平均個数（$\sigma = 3$）

[3] 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.

[4] R. Graham, D. E. Knuth, and O. Patashnik. **コンピュータの数学**. 共立出版, 1993.

[5] S. J. Modified Padovan words and the maximum number of runs in a word. *Australasian Journal of Combinatorics (to appear)*.

[6] 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.

[7] 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*, pages 185–192, 2008.

[8] 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*, pages 578–587. Springer, 2009.

[9] W. Matsubara, K. Kusano, A. Ishino, H. Bannai, and A. Shinohara. New lower bounds for the maximum number of runs in a string. In *Proceedings of the Prague Stringology Conference 2008*, pages 140–145, 2008.

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