Some results on the number of periodic factors in words

R. Kolpakov

Lomonosov Moscow State University, Dorodnicyn Computing Centre, Russia

8 February 2018

Repetitions

$w = a_1 \dots a_n$, $|w| = n$ — the length of w

Def: p is a *period* of w if $a_1 \dots a_{n-p} = a_{p+1} \dots a_n$

$p(w)$ — the minimal period of w

$e(w) = \frac{|w|}{p(w)}$ — the *exponent* of w

$re(w) = e(w) - 1$ — the *reduced exponent* of w

Ex: $w = aabaa$

3, 4 and 5 — periods of w

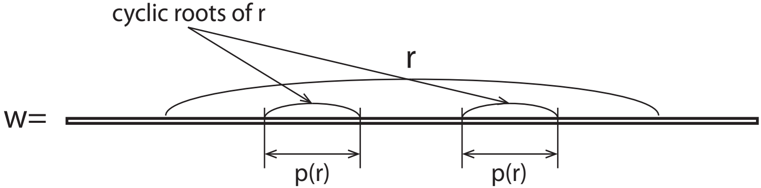
3 — minimal period of w

$\frac{5}{3}$ — exponent of w, $\frac{2}{3}$ — reduced exponent of w

w — *repetition* if $e(w) \geq 2$

Repetitions

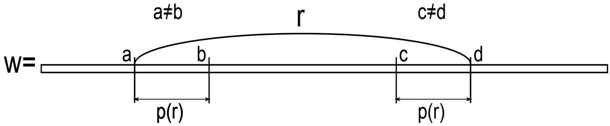
r — repetition in w



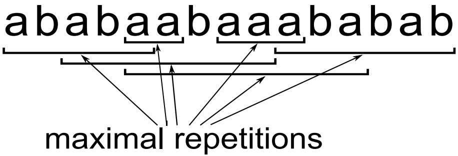
aba, *baa*, *aab* — cyclic roots in repetition *abaabaab*
abaabaab, *aabaabaaba* — repetitions with the same cyclic roots

Maximal repetitions

a repetition r in a word w is maximal (run) if



Ex:



Number of maximal repetitions

$R(n)$ — maximum number of maximal repetitions in words of length n
$E(n)$ — maximum sum of exponents of maximal repetitions in words of length n
$2R(n) \leq E(n)$

R.Kolpakov, G.Kucherov 1999: $E(n) = \Theta(n)$

H.Bannai,T.I, S.Inenaga, Y.Nakashima, M.Takeda, K.Tsuruta 2014: $R(n) < n$, $E(n) < 3n$

$R^{(2)}(n)$ — maximum number of maximal repetitions in binary words of length n
J.Fischer, S.Holub, T.I, M.Lewenstein 2015: $R^{(2)}(n) \leq \frac{22}{23}n$

Number of maximal repetitions

$\lambda = 1, 2, \ldots$

$R_{\geq\lambda}(w)$ — number of maximal repetitions with minimal periods $\geq \lambda$ in word w

$R_{\geq\lambda}(n) = \max_{|w|=n} R_{\geq\lambda}(w)$ — maximum number of maximal repetitions with minimal periods $\geq \lambda$ in words of length n

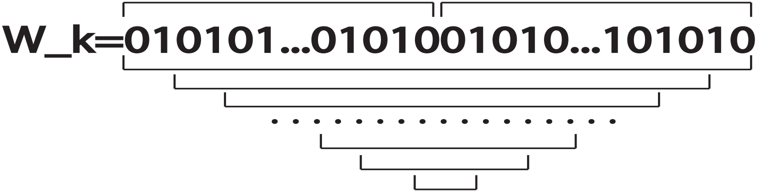
$R(n) = R_{\geq 1}(n) \geq R_{\geq 2}(n) \geq R_{\geq 3}(n) \geq \ldots$

Conjecture1: $R_{\geq\lambda}(n) \leq cn$ where $c \to 0$ as $\lambda \to infty$
Conjecture2: the same letter of word w is contained in $o(|w|)$ maximal repetitions of w

The conjectures are wrong!

Number of maximal repetitions

Ex: $w_k = (01)^k 0 (01)^k 0 = (01)^k 00 (10)^k, \quad |w_k| = 4k + 2$



$R_{\geq 1}(w_k) = k + 3, \quad R_{\geq \lambda}(w_k) = k + 3 - \lfloor \lambda/2 \rfloor \gtrsim k \gtrsim |w_k|/4$

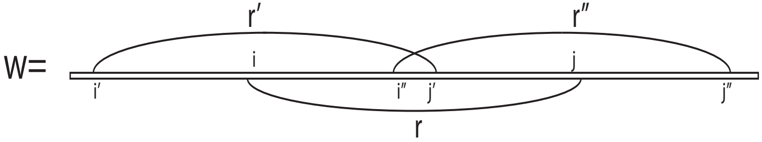
$R_{\geq 1}(n) \geq R_{\geq 2}(n) \geq R_{\geq 3}(n) \geq \ldots \gtrsim n/4$

middle letters are contained in $k + 2$ repetitions

Generation of repetitions

$r' \equiv w[i'..j'], r'' \equiv w[i''..j'']$ — maximal repetitions in w with the same cyclic roots, $p(r') = p(r'') = p$

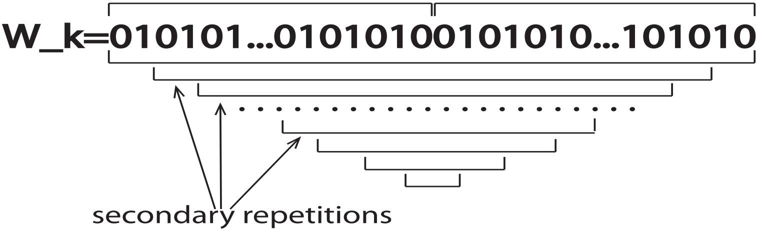
maximal repetition $r \equiv w[i..j]$ is generated by r' and r'' if $p(r) \geq 3p$, $i' < i \leq j'$, $i'' \leq j < j''$



Primary and secondary repetitions

maximal repetition is secondary if it is generated by other maximal repetitions

maximal repetition is primary if it is not secondary



Prop. Any secondary repetition is generated by only one pair of primary repetitions

Primary and secondary repetitions

$Rp_{\geq\lambda}(n)$ — maximum number of primary repetitions with minimal periods $\geq \lambda$ in words of length n
$Ep_{\geq\lambda}(n)$ — maximum sum of exponents of primary repetitions with minimal periods $\geq \lambda$ in words of length n
$Eps_{\geq\lambda}(n)$ — maximum sum of exponents of primary repetitions with minimal periods $\geq \lambda$ and secondary repetitions generated by these primary repetitions in words of length n

$$Eps_{\geq\lambda}(n) \geq Ep_{\geq\lambda}(n) \geq 2Rp_{\geq\lambda}(n)$$

Theorem 1. $Eps_{\geq\lambda}(n) = O(n/\lambda)$

Cor. $Ep_{\geq\lambda}(n) = O(n/\lambda), \quad Rp_{\geq\lambda}(n) = O(n/\lambda)$

Primary and secondary repetitions

Prop. The exponent of any secondary repetition is $< 7/3$, i.e. any maximal repetition with exponent $\geq 7/3$ is primary

$\hat{R}p_{\geq\lambda}(n)$ — maximum number of maximal repetitions with minimal periods $\geq \lambda$ and exponents $\geq 7/3$ in words of length n

Cor. $\hat{R}p_{\geq\lambda}(n) = O(n/\lambda)$

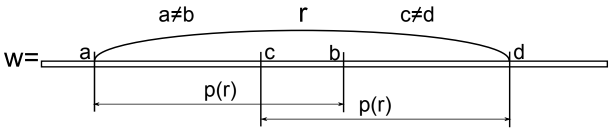
Theorem 2. In a word of length n the same letter is contained in $O(\log \frac{n}{\lambda})$ primary repetitions with minimal periods $\geq \lambda$

Cor. In a word of length n the same letter is contained in $O(\log \frac{n}{\lambda})$ maximal repetitions with minimal periods $\geq \lambda$ and exponents $\geq 7/3$

Subrepetitons

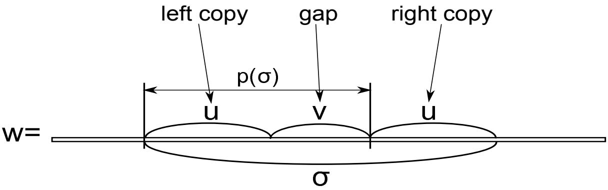
r is a subrepetition (δ-subrepetition) if $e(r) < 2$
$(1 + \delta \leq e(r) < 2 \Leftrightarrow \delta \leq re(r) < 1)$

a subrepetition r in a word w is maximal if



Gapped repeats

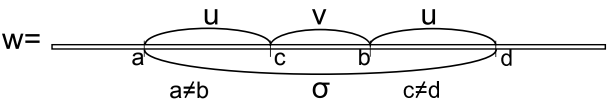
$\sigma = uvu$ — a gapped repeat in w:



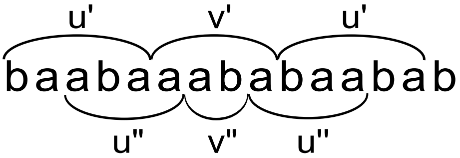
$p(\sigma) = |uv|$ — the period of σ
$c(\sigma) = |u|$ — the length of copies of σ
$\hat{e}(\sigma) = \frac{|\sigma|}{p(\sigma)} = 1 + \frac{|u|}{p(\sigma)}$ — the *exponent* of σ,
$r\hat{e}(\sigma) = \hat{e}(\sigma) - 1 = \frac{|u|}{p(\sigma)}$ — the *reduced exponent* of σ
$\alpha > 1$
σ is α-gapped repeat if $p(\sigma) \leq \alpha c(\sigma)$

Maximal gapped repeats

σ is maximal gapped repeat in w if

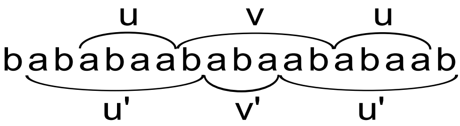


Ex:

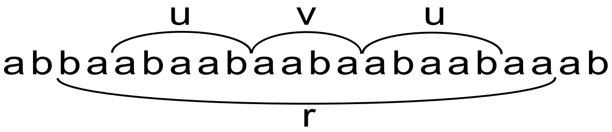


Maximal gapped repeats

any α-gapped repeat $\sigma = uvu$ is contained in either (uniquely defined) maximal α-gapped repeat $\sigma' = u'v'u'$ with the same period, **e.g:**

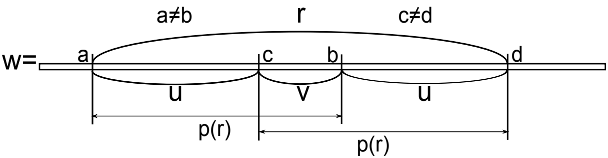


or (uniquely defined) maximal repetiton r such that $p(r)$ is a divisor of $p(\sigma)$, **e.g:**



Maximal gapped repeats and subrepetitions

r — maximal δ-subrepetition in a word w



$\sigma = uvu$ — maximal $\frac{1}{\delta}$-gapped repeat

r and σ are the same factor in w

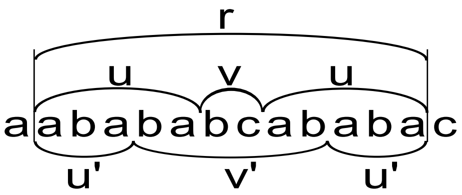
$\Downarrow$

r and σ are uniquely defined by each other

$p(\sigma) = p(r)$, so $\hat{e}(\sigma) = e(r)$ $(r\hat{e}(\sigma) = re(r))$

Maximal gapped repeats and subrepetitions

Ex:



$\sigma = uvu$ — maximal gapped repeat respective to r
$\sigma' = u'v'u'$ — maximal gapped repeat, s.t. r and σ' are same factor but σ' is not respective to r

thus, σ is *principal*, and σ' is not principal

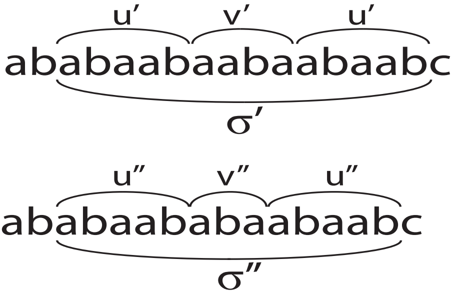
Primitive gapped repeats

$\underbrace{uu\ldots u}_{n}$ — *n-th power* of u, $n \geq 2$

word is *primitive* if it is not a power of some word

gapped repeat uvu is primitive if uv primitive

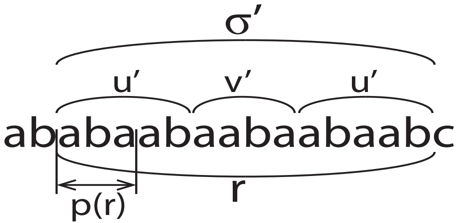
Ex:



$\sigma' = u'v'u'$ — maximal nonprimitive gapped repeat

$\sigma'' = u''v''u''$ — maximal primitive gapped repeat

Primitive gapped repeats



maximal nonprimitive gapped repeat $\sigma' = u'v'u'$ corresponds to maximal repetition r s.t. $p(r) = p(u'v')$

any maximal repetition r corresponds to no more than $\lceil e(r)/2 \rceil$ maximal nonprimitive gapped repeats

$$\Downarrow$$

word of length n contains no more than $O(E(n)) = O(n)$ maximal nonprimitive gapped repeats

Maximal gapped repeats

any principal maximal repeat is primitive, but maximal primitive repeats can be not principal

K^s_δ (K^s) — class of all maximal δ-subrepetitions (subrepetitions) $=$ principal maximal $\frac{1}{\delta}$-gapped repeats (gapped repeats)
K^p_δ (K^p) — class of all maximal primitive $\frac{1}{\delta}$-gapped repeats (gapped repeats)
K^m_δ (K^m) — class of all maximal $\frac{1}{\delta}$-gapped repeats (gapped repeats)

$$K^s_\delta(K^s) \subseteq K^p_\delta(K^p) \subseteq K^m_\delta(K^m)$$

Maximal gapped repeats and subrepetitions

$RE^m(w)$ — sum of reduced exponents of all maximal gapped repeats in word w

R.Kolpakov, G.Kucherov, P.Ochem 2010: $RE^m(w) \leq n \ln n$

a gapped repeat σ is α-gapped if
$\frac{p(\sigma)}{c(\sigma))} \leq \alpha \Leftrightarrow r\hat{e}(\sigma) = \frac{c(\sigma))}{p(\sigma)} \geq 1/\alpha$

Cor. 1 Number of all maximal α-gapped repeats in word w is not greater than $\alpha n \ln n$

$RE^s(w)$ — sum of reduced exponents of all maximal subrepetitons in word w $=$ sum of reduced exponents of all principal maximal gapped repeats $\leq RE^m(w) \leq n \ln n$

Cor. 2 Number of all maximal δ-subrepetitons in word w is not greater than $n \ln n/\delta$

Maximal gapped repeats and subrepetitions

$RE^m_{\geq\lambda}(w)$ ($RE^m_{\leq\lambda}(w)$) — sum of reduced exponents of all maximal gapped repeats with periods $\geq \lambda$ ($\leq \lambda$) in word w

R.Kolpakov, G.Kucherov, P.Ochem 2010:
$RE^m_{\geq\lambda}(w) \leq n \ln(n/\lambda)$, $RE^m_{\leq\lambda}(w) \leq n(1 + \ln \lambda)$

Cor. 1 Number of all maximal α-gapped repeats with periods $\geq \lambda$ ($\leq \lambda$) in word w is not greater than $\alpha n \ln(n/\lambda)$ ($\alpha n(1 + \ln \lambda)$)

Cor. 2 Number of all maximal δ-subrepetitons with minimal periods $\geq \lambda$ ($\leq \lambda$) in word w is not greater than $n \ln(n/\lambda)/\delta$ ($n(1 + \ln \lambda)/\delta$)

Maximal gapped repeats and subrepetitions

$RE^m(n) = \max_{|w|=n} RE^m(w), \quad RE^s(n) = \max_{|w|=n} RE^s(w)$
$RE^p(n) = \max_{|w|=n} RE^p(w)$ where $RE^p(w)$ — sum of reduced exponents of all maximal primitive gapped repeats in word w

Lower bounds for unbounded alphabet:

$w'_k = ab_1ab_2ab_3 \dots ab_ka, \quad |w'_k| = 2k+1$

$$RE^m(w'_k) = RE^p(w'_k) = RE^s(w'_k) > \frac{1}{2}[(k+1)\ln(k+1) - k]$$

$$\gtrsim \frac{1}{4}|w'_k| \ln |w'_k|$$

$$\frac{1}{4} n \ln n \lesssim RE^s(n) \le RE^p(n) \le RE^m(n) \le n \ln n$$

Maximal gapped repeats and subrepetitions

$RE^s(n) = \Theta(n \log n)$
$RE^p(n) = \Theta(n \log n)$
$RE^m(n) = \Theta(n \log n)$

w'_k contains no less than $\lfloor \alpha/2 \rfloor [(k+1) - \lceil \alpha/2 \rceil] = \Omega(\alpha |w'_k|)$ maximal primitive α-gapped repeats and no less than $\lfloor \frac{1}{2\delta} \rfloor [(k+1) - \lceil \frac{1}{2\delta} \rceil] = \Omega(|w'_k|/\delta)$ maximal δ-subrepetitons

w'_k contains a total of $\Theta(|w'_k|^2)$ maximal subrepetitons

Maximal gapped repeats and subrepetitions

$RE^m_{bin}(n) = \max_{w\in\{0,1\}^n} RE^m(w)$
$RE^p_{bin}(n) = \max_{w\in\{0,1\}^n} RE^p(w)$
$RE^s_{bin}(n) = \max_{w\in\{0,1\}^n} RE^s(w)$

Lower bounds for binary alphabet:

$$w''_k = (0011)^k = \underbrace{00110011\ldots0011}_{k}, \quad |w''_k| = 4k$$

$$RE^p_{bin}(w''_k) > (k-\frac{1}{4})[\ln(4k-1)-\ln 3]-(k-1) \gtrsim \frac{1}{4}|w''_k| \ln |w''_k|$$

$$\frac{1}{4} n \ln n \lesssim RE^p_{bin}(n) \le RE^m_{bin}(n) \le n \ln n$$

$RE^p_{bin}(n) = \Theta(n \log n), \quad RE^m_{bin}(n) = \Theta(n \log n)$

Maximal gapped repeats and subrepetitions

w''_k contains no less than $\lfloor\frac{\alpha-1}{4}\rfloor[4k-\lceil\alpha\rceil]=\Omega(\alpha|w''_k|)$ maximal primitive α-gapped repeats

R.Kolpakov, G.Kucherov, P.Ochem 2010: w''_k contains a total of $\Theta(|w''_k|^2)$ maximal primitive gapped repeats

w''_k contains $2k+1$ maximal repetitions and only $2k-2$ maximal subrepetitions (all these subrepetitions are of the form *abba* where $a,b\in\{0,1\}$, $a\neq b$), i.e.
$RE^s(w''_k)=(2k-2)/3\sim|w''_k|/6$

Maximal gapped repeats and subrepetitions

w — word of length n

$RE^m(w) - RE^p(w) \;=\;$ sum of reduced exponents of all maximal nonprimitive gapped repeats in word $w \;<\;$ number of all maximal nonprimitive gapped repeats in word $w = O(n)$

$$RE^m(n) - RE^p(n) = O(n) \quad \Rightarrow \quad RE^m(n) \sim RE^p(n)$$

in analogous way $RE^m_{bin}(n) \sim RE^p_{bin}(n)$

Maximal α-gapped repeats and δ-subrepetitions

M.Crochemore, R.Kolpakov, G.Kucherov 2015: word of length n contains $O(\alpha n)$ maximal α-gapped repeats

P.Gawrychowski, T.I, S.Inenaga, D.Köppl, F.Manea 2015: word of length n contains no more than $18\alpha n$ maximal α-gapped repeats

by the example of w''_k, this bound is asymptotically tight for big enough α for any alphabet

Cor. word of length n contains no more than $18n/\delta$ maximal δ-subrepetitions

by the example of w'_k, this bound is asymptotically tight for small enough δ for unbounded alphabet

Maximal repeats with arbitrary gap

$f : \mathbf{N} \to \mathbf{R}$, $g : \mathbf{N} \to \mathbf{R}$, s.t. $0 < g(x) \leq f(x)$

$\sigma = uvu$ — f, g-repeat if $g(|u|) \leq |v| \leq f(|u|)$

α-gapped repeats are f, g-repeats for $g(x) = \min(1, \alpha - 1)$,
$f(x) = (\alpha - 1)x$

Maximal repeats with arbitrary gap

$f : \mathbf{N} \to \mathbf{R}$, $g : \mathbf{N} \to \mathbf{R}$, s.t. $0 < g(x) \leq f(x)$

$$\partial_f^+(x) = \begin{cases} (f(x+1) - f(x)), & \text{if } f(x+1) \geq f(x); \\ 0, & \text{otherwise;} \end{cases}$$

$$\partial_f^-(x) = \begin{cases} (f(x) - f(x+1)), & \text{if } f(x) \geq f(x+1); \\ 0, & \text{otherwise.} \end{cases}$$

$\partial_f^+ = \sup_x\{\partial_f^+(x)\}, \quad \partial_f^- = \sup_x\{\partial_f^-(x)\}$
(if these supremums exist)

$\partial_{f,g}^a = \max\{\partial_f^+, \partial_g^-\}$ (if ∂_f^+, ∂_g^- exists)
$\partial_{f,g}^b = \max\{\partial_f^-, \partial_g^+\}$ (if ∂_f^-, ∂_g^+ exists)

Maximal repeats with arbitrary gap

$f : \mathbf{N} \to \mathbf{R}$, $g : \mathbf{N} \to \mathbf{R}$, s.t. $0 < g(x) \le f(x)$

if at least one of values $\partial^a_{f,g}$, $\partial^b_{f,g}$ exists

$$\partial_{f,g} = \min\{\partial^a_{f,g}, \partial^b_{f,g}\}$$

$\Delta_{f,g}(x) = \frac{1}{x}(f(x) - g(x)) \ge 0$
$\Delta_{f,g} = \sup_x\{\Delta_{f,g}(x)\}$ (if these supremums exists)

Theorem. Let for $f(x)$, $g(x)$ the both values $\partial_{f,g}$, $\Delta_{f,g}$ exist. Then a word of length n contains no more than $O(n(1 + \max\{\partial_{f,g}, \Delta_{f,g}\}))$ maximal f,g-repeats.

Maximal repeats with arbitrary gap

Ex: $f(x) = \alpha_f + \beta_f x$, $g(x) = \alpha_g + \beta_g x$ where $0 < \alpha_g \le \alpha_f$, $0 \le \beta_g \le \beta_f$.

$\partial_f^+(x) = \beta_f$, $\partial_f^-(x) = 0$, $\partial_g^+(x) = \beta_g$, $\partial_g^-(x) = 0$
$\partial_f^+ = \beta_f$, $\partial_f^- = 0$, $\partial_g^+ = \beta_g$, $\partial_g^- = 0$.
$\partial_{f,g}^a = \beta_f$, $\partial_{f,g}^b = \beta_g$
$\partial_{f,g} = \min\{\beta_f, \beta_g\} = \beta_g$.
$\Delta_{f,g}(x) = \frac{1}{x}(\alpha_f - \alpha_g) + (\beta_f - \beta_g)$
$\Delta_{f,g} = \Delta_{f,g}(1) = (\alpha_f - \alpha_g) + (\beta_f - \beta_g)$.
Thus, the number of maximal f, g-repeats in a word of length n is bounded by

$$O(n(1 + \max\{\beta_g, (\alpha_f - \alpha_g) + (\beta_f - \beta_g)\})) = O(n(1 + \max\{\beta_g, (\alpha_f - \alpha_g), (\beta_f - \beta_g)\})).$$

Open problems

1. Check if for any λ

$$R_{\geq\lambda}(n) - R_{\geq\lambda+1}(n) = \Omega(n)$$

2. $\hat{w}_k = ((01)^k 0)^k$

$R_{\geq\lambda}(\hat{w}_k) \gtrsim k(k-1) \gtrsim |\hat{w}_k|/2$
$R_{\geq 1}(n) \geq R_{\geq 2}(n) \geq R_{\geq 3}(n) \geq ... \gtrsim n/2$

Is the lower bound $n/2$ tight?

Open problems

3. $Rp(n)$ — maximum number of primary repetitions in words of length n
$Ep(n)$ — maximum sum of exponents of primary repetitions in words of length n

$$Rp(n) \stackrel{?}{=} R(n), \quad Ep(n) \stackrel{?}{=} E(n)$$

4. $RE^{s}_{bin}(n) \stackrel{?}{=} \Theta(n \log n)$

Is the $O(n/\delta)$ upper bound on the number of maximal δ-subrepetitions asymptotically tight for words over binary alphabet?

Open problems

5. Is it true that word of length n contains no more than αn maximal α-gapped repeats?

6. $f(x) = \alpha x$, $g(x) = \beta x$ where $\beta < \alpha$
Is it true that word of length n contains $O(n(1 + \alpha - \beta))$ maximal f, g-repeats?