Searching of gapped repeats and subrepetitions in a word

Roman Kolpakov

Moscow State University, Russia

Mikhail Podolskiy

Moscow State University, Russia

Mikhail Posypkin

Institute for Information Transmission Problems, Moscow, Russia

Nickolay Khrapov

Institute for Information Transmission Problems, Moscow, Russia

$w = a_1 \dots a_n, \quad |w| = n$ — 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

Ex: $w = aabaa$

3, 4 and 5 — periods of w

3 — minimal period of w

$\frac{5}{3}$ — exponent of w

w — *repetition* if $e(w) \geq 2$

uu — *square*

Ex: $abcabc = (abc)^2$ — square of abc

uuu — *cube*

Ex: $abcabcabc = (abc)^3$ — cube of abc

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

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

square uu is *primitive* if u is primitive

M.Crochemore 1981: a word of length n contains $O(n \log n)$ primitive squares, and all primitive squares in the word can be found in $O(n \log n)$ time

D.Gusfield, J.Stoye 2004: all primitive squares in a word of length n can be found in $O(n + S)$ time where S is the number of the squares (for the constant alphabet size)

(fractional) repetition:

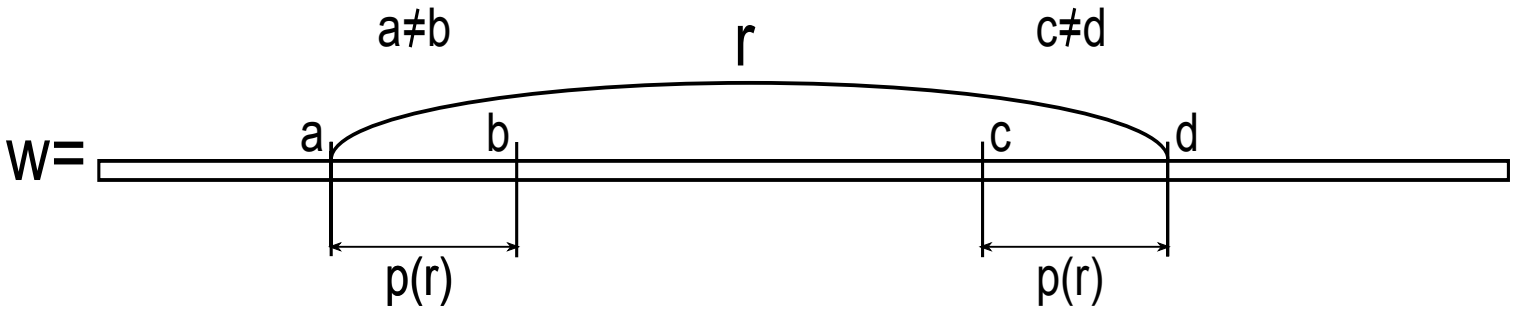
$r = \underbrace{uu \ldots u}_{k} u' = u^k u'$, where $|u| = p(r)$, $k \geq 2$, u' — prefix of u

u — *minimal root* of r

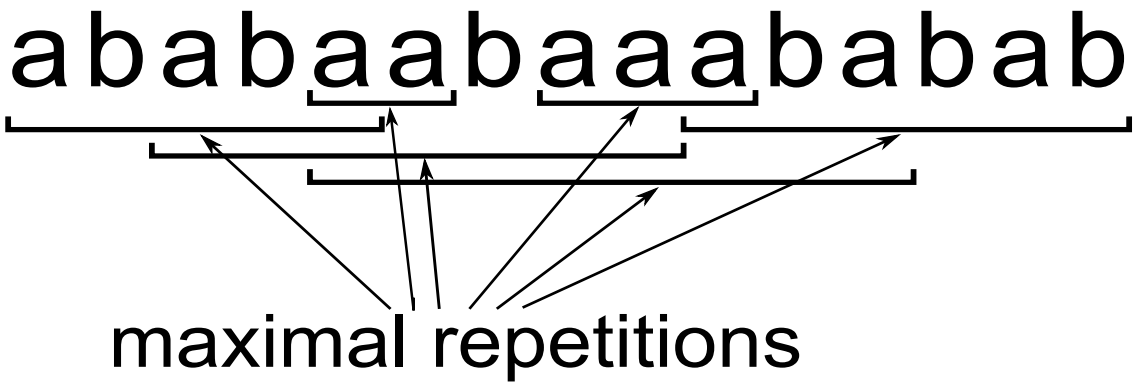
Ex: $r = abcabcabcabcab = (abc)^4 ab$

abc — *minimal* root of r

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



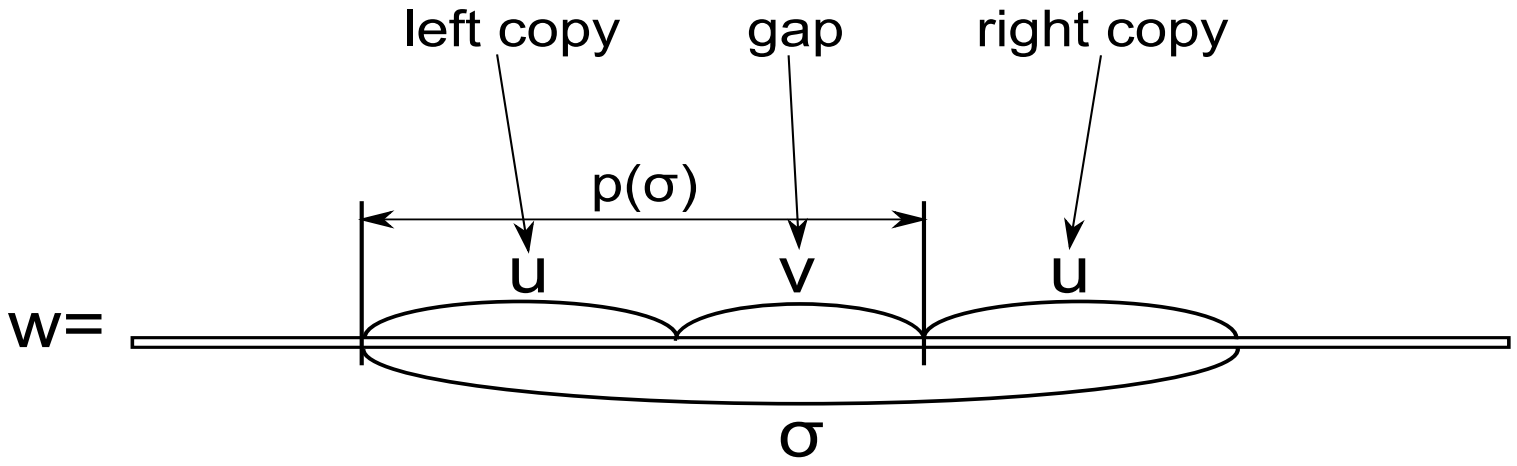
Ex:



R.Kolpakov, G.Kucherov 1999: a word of length n contains $O(n)$ maximal repetitions, and all maximal repetitions in the word can be found in $O(n)$ time (for the constant alphabet size)

M.Crochemore, L.Ilie, L.Tinta 2008: a word of length n contains no more than $1.029n$ maximal repetitions

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

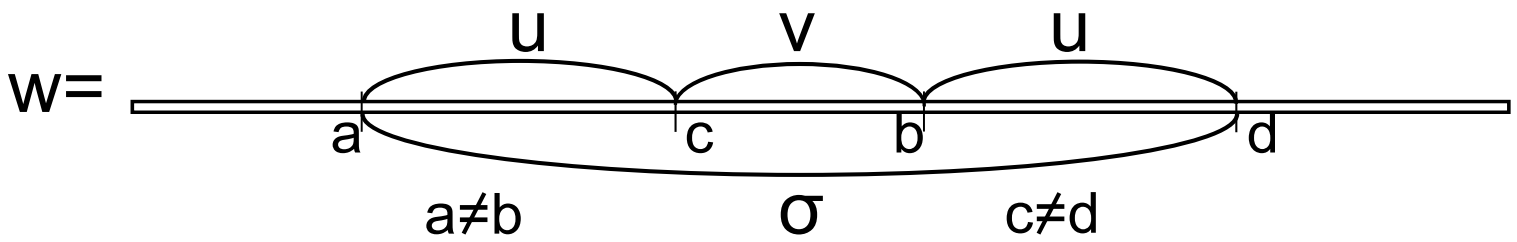


$p(\sigma)$ — the period of σ

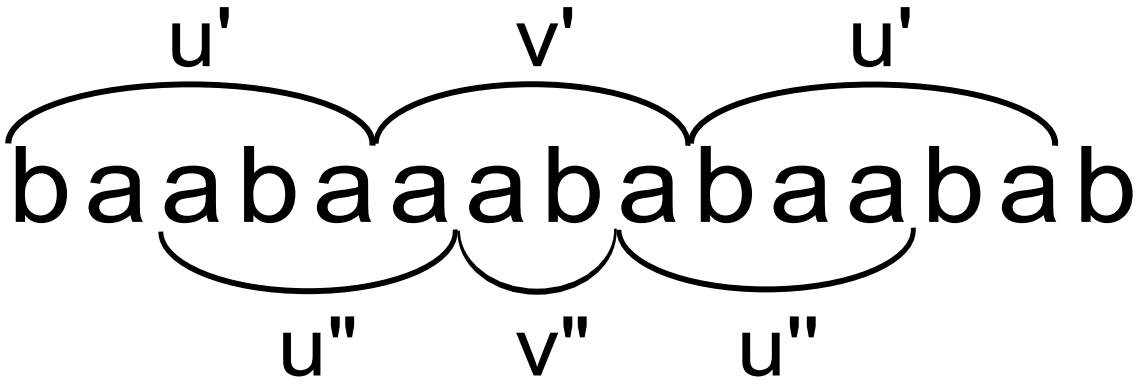
$c(\sigma) = |u|$ — the length of copies of σ

σ is α-gapped repeat if $p(\sigma) \leq \alpha c(\sigma)$

σ is maximal gapped repeat in w if



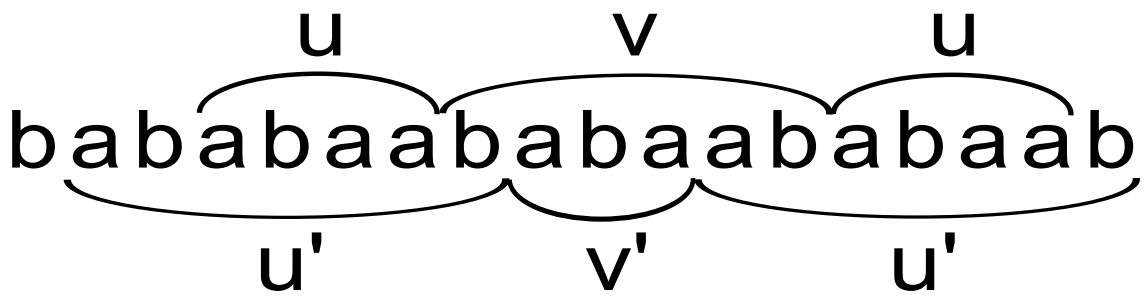
Ex:



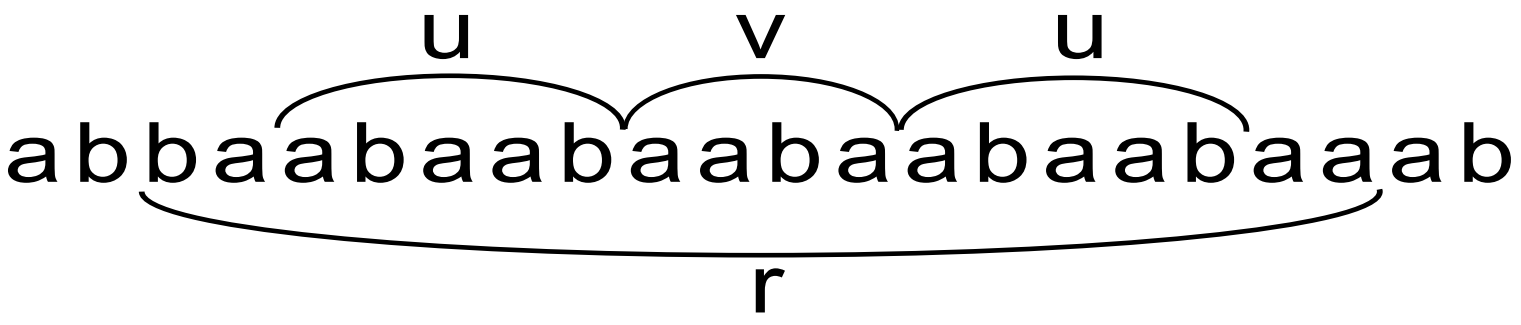
G.Brodal, R.Lyngso, C.Pedersen, J.Stoye 2000: in a word of length n all maximal gapped repeats with gap length from a given interval can be found in $O(n \log n + S)$ time where S is the output size

R.Kolpakov, G.Kucherov 2000: in a word of length n all gapped repeats with a given gap length d can be found in $O(n \log d + S)$ time where S is the output size (for the constant alphabet size)

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:**

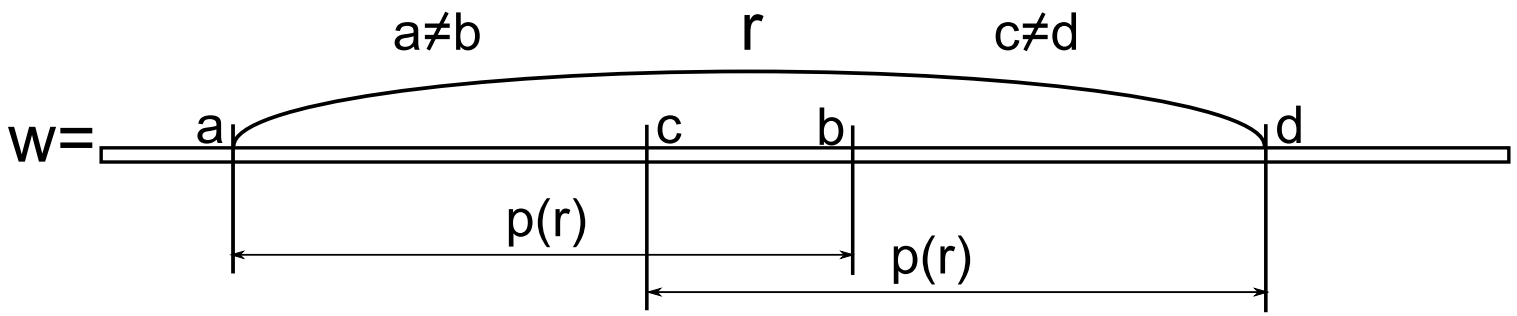


or (uniquely defined) maximal repetiton r such that $p(r)$ is a divisor of $p(\sigma)$, **e.g:**



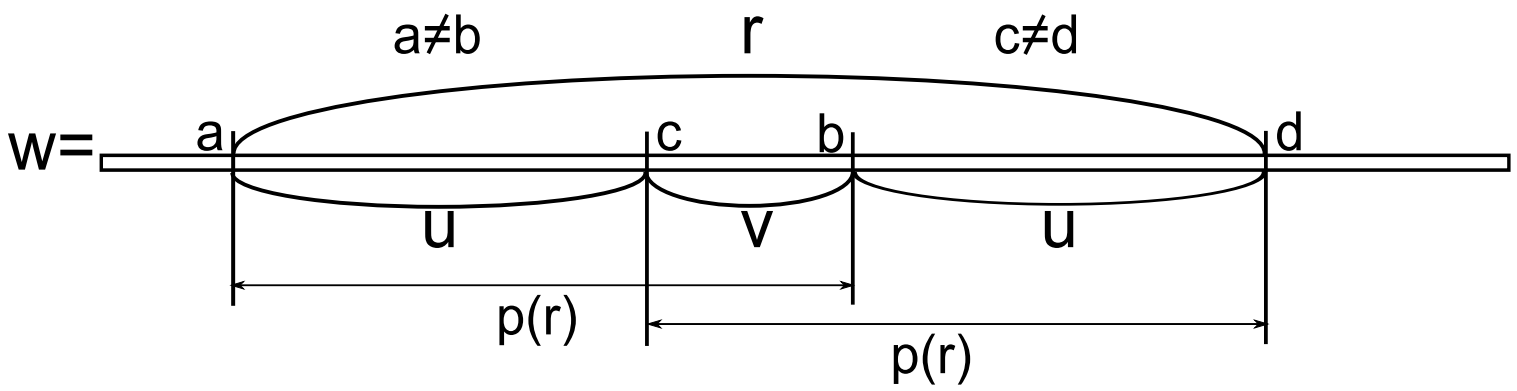
r is δ-subrepetition if $1 + \delta \leq e(r) < 2$

a subrepetition r in a word w is maximal if



R.Kolpakov, G.Kucherov, P.Ochem 2010: a word of length n contains $O(\frac{n}{\delta} \log n)$ maximal δ-subrepetitions

r — maximal δ-subrepetition in a word w



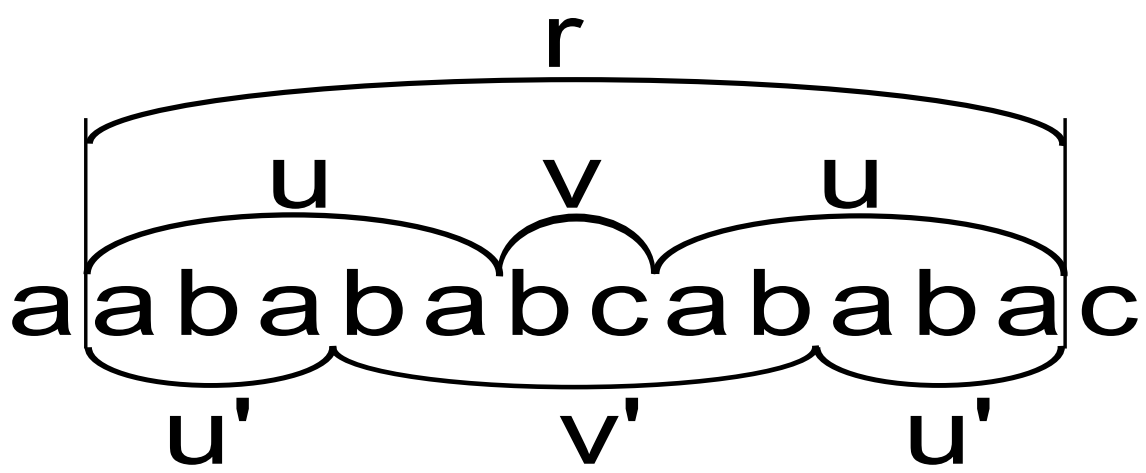
$\sigma = uvu$ — maximal $\frac{1}{\delta}$-gapped repeat *respective to* r

r and σ are the same factor in w

$\Downarrow$

r and σ are uniquely defined by each other

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

number of maximal δ-subrepetitions = number of principal maximal $\frac{1}{\delta}$-gapped repeats

$$\Downarrow$$

number of maximal δ-subrepetitions $\leq$ number of maximal $\frac{1}{\delta}$-gapped repeats

Lemma: A word of length n contains $O(\alpha^2 n)$ maximal α-gapped repeats.

Corollary: A word of length n contains $O(n/\delta^2)$ maximal δ-subrepetitions.

Theorem 1: In a word of length n all maximal α-gapped repeats can be computed in $O(\alpha^2 n)$ time for the constant alphabet size and in $O(n \log n + \alpha^2 n)$ time for the general case.

computing all maximal δ-subrepetitions $\equiv$ selecting all principal repeats from all maximal $\frac{1}{\delta}$-gapped repeats

Proposition 1: A maximal α-gapped repeat σ is principal iff σ is not contained in a maximal α-gapped repeat σ' s.t. $p(\sigma') < p(\sigma)$ or in a maximal repetition r s.t. $p(r) < p(\sigma)$.

Theorem 2: In a word of length n all maximal δ-subrepetitions can be computed in $O(\frac{n \log\log n}{\delta^2})$ time for the constant alphabet size and $O(n \log n + \frac{n \log\log n}{\delta^2})$ time for the general case.

Proposition 2: A maximal gapped repeat $\sigma = uvu$ is principal iff $p(\sigma)$ is the minimal period of uvu.

T.Kociumaka, J.Radoszewski, W.Rytter, T.Walen 2012: a hash table data structure for a given word of length n can be constructed in $O(n \log n)$ expected time and allows to compute the minimal period p of a required factor u of the word in $O(\log(1 + \frac{|u|}{|u|-p}))$ time

Theorem 3: In a word of length n all maximal δ-subrepetitions can be computed in $O(n \log n + \frac{n}{\delta^2} \log \frac{1}{\delta})$ expected time.

Further research:

- improving the bounds for the numbers of maximal α-gapped repeats and maximal δ-subrepetitions in words of fixed length;
- optimal algorithms for computing all maximal α-gapped repeats and all maximal δ-subrepetitions in a word.