Counting maximal-exponent factors in words

Robert Mercaș

Loughborough University

joint work with Golnaz Badkobeh and Maxime Crochemore

DACS 2016

What is it?!

tin

What is it?!

tin^2

What is it?!

$$tin^2 = tintin$$

What is it?!

$$tin^3 = tintintin$$

What is it?!

$$tin^3 = tintintin$$

$$an$$

What is it?!

$$tin^3 = tintintin$$

$$an^{\frac{3}{2}} = ana$$

What is it?!

$$tin^3 = tintintin$$

$$an^{\frac{5}{2}} = anana$$

What is it?!

$$tin^3 = tintintin$$

$$b(an)^{\frac{5}{2}} = banana$$

What is it?!

$$tin^3 = tintintin$$

$$b(an)^{\frac{5}{2}} = banana$$

$$mississippi$$

What is it?!

$$tin^3 = tintintin$$

$$b(an)^{\frac{5}{2}} = banana$$

$$mississippi = m(iss)^{\frac{7}{3}}ppi$$

Motivation

Motivation

Theoretical

Motivation

Theoretical: combinatorics on words

Motivation

Theoretical: combinatorics on words
Applications

Motivation

Theoretical: combinatorics on words
Applications: compression, coding (run-length or Ziv–Lempel compression)

Motivation

Theoretical: combinatorics on words
Applications: compression, coding (run-length or Ziv–Lempel compression)
Real life

Motivation

Theoretical: combinatorics on words
Applications: compression, coding (run-length or Ziv–Lempel compression)
Real life: analysis of genetic sequences (tandem repeats), prediction of the secondary structure of RNA (palindromic repeats)

Motivation

Theoretical: combinatorics on words
Applications: compression, coding (run-length or Ziv–Lempel compression)
Real life: analysis of genetic sequences (tandem repeats), prediction of the secondary structure of RNA (palindromic repeats)

Repetitions have been thoroughly investigated

Motivation

Theoretical: combinatorics on words
Applications: compression, coding (run-length or Ziv–Lempel compression)
Real life: analysis of genetic sequences (tandem repeats), prediction of the secondary structure of RNA (palindromic repeats)

Repetitions have been thoroughly investigated:

- avoidability

Motivation

Theoretical: combinatorics on words
Applications: compression, coding (run-length or Ziv–Lempel compression)
Real life: analysis of genetic sequences (tandem repeats), prediction of the secondary structure of RNA (palindromic repeats)

Repetitions have been thoroughly investigated:

- avoidability
- runs

Motivation

Theoretical: combinatorics on words
Applications: compression, coding (run-length or Ziv–Lempel compression)
Real life: analysis of genetic sequences (tandem repeats), prediction of the secondary structure of RNA (palindromic repeats)

Repetitions have been thoroughly investigated:

- avoidability
- runs
- α-gapped repeats

What we do

We consider factors uvu where u is their longest border.

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.
We focus on factors having the maximal exponent among all factors occurring in a square-free word

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.
We focus on factors having the maximal exponent among all factors occurring in a square-free word: maximal-exponent factors (MEF).

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.
We focus on factors having the maximal exponent among all factors occurring in a square-free word: maximal-exponent factors (MEF).

Existing bounds on the number of occurrences of MEFs:

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.
We focus on factors having the maximal exponent among all factors occurring in a square-free word: maximal-exponent factors (MEF).

Existing bounds on the number of occurrences of MEFs:

- upper bound: $2.25n$

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.
We focus on factors having the maximal exponent among all factors occurring in a square-free word: maximal-exponent factors (MEF).

Existing bounds on the number of occurrences of MEFs:

- upper bound: $2.25n$
- lower bound: $0.66n$

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.
We focus on factors having the maximal exponent among all factors occurring in a square-free word: maximal-exponent factors (MEF).

Existing bounds on the number of occurrences of MEFs:

- upper bound: $2.25n$
- lower bound: $0.66n$

Results in this paper on the number of occurrences of MEFs::

- upper bound: $1.8n$

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.
We focus on factors having the maximal exponent among all factors occurring in a square-free word: maximal-exponent factors (MEF).

Existing bounds on the number of occurrences of MEFs:

- upper bound: $2.25n$
- lower bound: $0.66n$

Results in this paper on the number of occurrences of MEFs::

- upper bound: $1.8n$
- upper bound: n (special cases)

What we do

We consider factors uvu where u is their longest border.
Their number may be quadratic with respect to the word length.
We focus on factors having the maximal exponent among all factors occurring in a square-free word: maximal-exponent factors (MEF).

Existing bounds on the number of occurrences of MEFs:

- upper bound: $2.25n$
- lower bound: $0.66n$

Results in this paper on the number of occurrences of MEFs::

- upper bound: $1.8n$
- upper bound: n (special cases)
- lower bound: $\frac{k}{k+1}n$

Notations

$w = abbaabbaababa$

Notations

$$w = abbaabbaababa \text{ with } |w| = 13$$

Notations

$$w = abbaabbaababa \text{ with } |w| = 13$$

$$w = xyz$$

Notations

$$w = abbaabbaababa \text{ with } |w| = 13$$

$$w = xyz$$

p is a period of w if $w[i] = w[i + p]$

Notations

$$w = abbaabbaababa \text{ with } |w| = 13$$

$$w = xyz$$

p is a period of w if $w[i] = w[i+p]$ and $p(w)$ the smallest period

Notations

$$w = abbaabbaababa \text{ with } |w| = 13$$

$$w = xyz$$

p is a period of w if $w[i] = w[i+p]$ and $p(w)$ the smallest period

exponent of w, denoted by $e(y) = \frac{|y|}{p(y)}$

Notations

$$w = abbaabbaababa \text{ with } |w| = 13$$

$$w = xyz$$

p is a period of w if $w[i] = w[i+p]$ and $p(w)$ the smallest period

exponent of w, denoted by $e(y) = \frac{|y|}{p(y)}$

The MEFs are factors uvu of w whose exponents $\frac{uvu}{uv}$ are maximal amongst all exponents of other factors of w.

Notations

$$w = abbaabbaababa \text{ with } |w| = 13$$

$$w = xyz$$

p is a period of w if $w[i] = w[i+p]$ and $p(w)$ the smallest period

exponent of w, denoted by $e(y) = \frac{|y|}{p(y)}$

The MEFs are factors uvu of w whose exponents $\frac{uvu}{uv}$ are maximal amongst all exponents of other factors of w.

abbaabbaab and ababa represent MEFs with $u =$ ab and $v =$ ba, or $u =$ a and $v =$ b, respectively, where the exponent of the MEF is 2.5

Previous Results

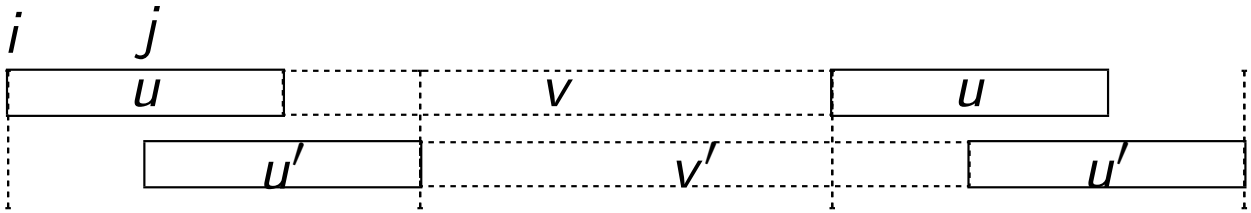
Lemma (Badkobeh, Crochemore)

Consider two occurrences of MEFs with the same border length b starting at respective i and j positions in the word. Then, $|j - i| > b$.

Previous Results

Lemma (Badkobeh, Crochemore)

Consider two occurrences of MEFs with the same border length b starting at respective i and j positions in the word. Then, $|j - i| > b$.



Previous Results

Lemma (Badkobeh, Crochemore)

Consider two occurrences of MEFs with the same border length b starting at respective i and j positions in the word. Then, $|j - i| > b$.

A MEF uvu is a δ-MEF if its border length $b = |u| = |uvu| - p(uvu)$ satisfies $2\delta < b \leq 4\delta$. Then any MEF is a δ-MEF for some $\delta \in \Delta$, where $\Delta = \{1/4, 1/2, 1, 2, 2^2, 2^3, \ldots\}$.

Lemma (Badkobeh, Crochemore)

Let uvu and $u'v'u'$ be occurrences of δ-MEFs in w whose left borders mid-positions are at respective positions i and j on w. Then, $|j - i| \geq \delta$.

Previous Results

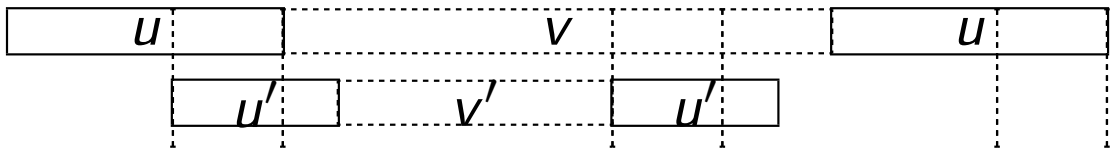
Lemma (Badkobeh, Crochemore)

Consider two occurrences of MEFs with the same border length b starting at respective i and j positions in the word. Then, $|j - i| > b$.

A MEF uvu is a δ-MEF if its border length $b = |u| = |uvu| - p(uvu)$ satisfies $2\delta < b \leq 4\delta$. Then any MEF is a δ-MEF for some $\delta \in \Delta$, where $\Delta = \{1/4, 1/2, 1, 2, 2^2, 2^3, \ldots\}$.

Lemma (Badkobeh, Crochemore)

Let uvu and $u'v'u'$ be occurrences of δ-MEFs in w whose left borders mid-positions are at respective positions i and j on w. Then, $|j - i| \geq \delta$.



Previous Results

Lemma (Badkobeh, Crochemore)

Consider two occurrences of MEFs with the same border length b starting at respective i and j positions in the word. Then, $|j - i| > b$.

A MEF uvu is a δ-MEF if its border length $b = |u| = |uvu| - p(uvu)$ satisfies $2\delta < b \leq 4\delta$. Then any MEF is a δ-MEF for some $\delta \in \Delta$, where $\Delta = \{1/4, 1/2, 1, 2, 2^2, 2^3, \ldots\}$.

Lemma (Badkobeh, Crochemore)

Let uvu and $u'v'u'$ be occurrences of δ-MEFs in w whose left borders mid-positions are at respective positions i and j on w. Then, $|j - i| \geq \delta$.

The direct consequence of the previous lemma is that if uvu and $u'v'u'$ are two δ-MEFs, then u cannot contain u'. Hence, we get the upper bound:

$$\left(\sum_{b=1}^{b=k} \frac{n}{b+1}\right) + \frac{1}{k}\left(2 + \frac{1}{2} + \frac{1}{2^2} + \cdots\right) n = \left(\sum_{b=1}^{b=k} \frac{n}{b+1}\right) + \frac{4n}{k}.$$

Small tweak

Lemma

Let $S = [r, .., s]$ be an interval of integers such that $r > \frac{2s}{3}$. Then within every $r+1$ positions, there are at most two MEFs with border lengths in S.

Small tweak

Lemma

Let $S = [r, .., s]$ be an interval of integers such that $r > \frac{2s}{3}$. Then within every $r+1$ positions, there are at most two MEFs with border lengths in S.

We introduce the notion of γ-MEFs, for a positive real number γ: a MEF uvu is a γ-MEF if its border length $b = |u|$ satisfies $2\gamma \leq b < 3\gamma$. Then any MEF is a γ-MEF for some $\gamma \in \Gamma$ where $\Gamma = \{\frac{1}{2}, \frac{1}{2} \cdot (\frac{3}{2}), \frac{1}{2} \cdot (\frac{3}{2})^2, \ldots\}$.

Corollary

For three consecutive γ-MEFs starting at positions i, j and k, respectively, $\max\{k-i, j-i\} > 3\gamma$.

Long Borders

Theorem

There are less than $4n/b$ occurrences of MEFs with maximum length border at least b in a length n word.

Long Borders

Theorem

There are less than $4n/b$ occurrences of MEFs with maximum length border at least b in a length n word.

Corollary

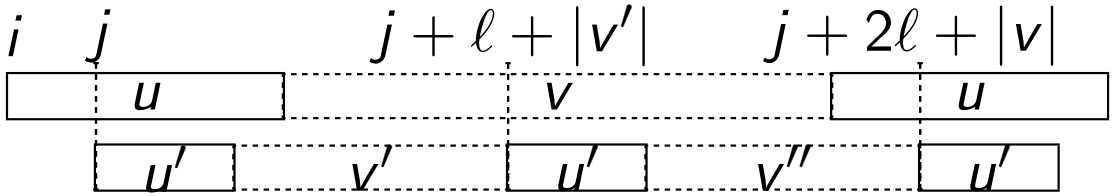
There are less than $n/2$ occurrences of MEFs with border length at least 8 in a word of length n.

Double borders

What happens if a *MEF* with the border double the length of another, contains the latter's border, entirely?

Double borders

What happens if a *MEF* with the border double the length of another, contains the latter's border, entirely?



Double borders

What happens if a *MEF* with the border double the length of another, contains the latter's border, entirely?

Proposition

There are at most $2n/(2\ell+1)$ *MEFs with border lengths* ℓ *and* 2ℓ *in a word of length* n.

Exponential borders

Example

Consider words $u = ab$, $v = ac$ and alphabet $\Sigma = \{a_1, b_1, a_2, b_2, \ldots\}$ for which $a, b, c \notin \Sigma$. Let $S_1 = ua_1vb_1$ and $S_i = S_{i-1}ua_ivb_i$ for $i \geq 2$.

$$ab \cdot ac \cdot ab \cdot ac \cdot ab \cdot ac \cdot ab \cdot ac \cdot ab \cdot ac \cdot ab$$

Exponential borders

Example

Consider words $u = ab$, $v = ac$ and alphabet $\Sigma = \{a_1, b_1, a_2, b_2, \ldots\}$ for which $a, b, c \notin \Sigma$. Let $S_1 = ua_1vb_1$ and $S_i = S_{i-1}ua_ivb_i$ for $i \geq 2$.

$$ab \cdot ac \cdot ab \cdot ac \cdot ab \cdot ac \cdot ab \cdot ac \cdot ab \cdot ac \cdot ab$$

Lemma

Every word of length n contains, for any positive integer $\ell \leq \frac{n}{2}$, at most $\frac{n^2}{\ell n + 2\ell}$ MEFs with the border length of the form $\ell\, 2^i$, for any $0 \leq i \leq \log(\frac{n}{\ell})$.

Small borders

Lemma

Every word of length n contains at most $4n/5$ MEFs with the border length in the set $\{1, 2, 4\}$.

Small borders

Lemma

Every word of length n contains at most $4n/5$ MEFs with the border length in the set $\{1, 2, 4\}$.

Lemma

There are at most $13n/10$ occurrences of MEFs whose border length is at most 7 in a word of length n.

Main Result

Theorem

There exist at most $1.8n$ *number of occurrences of MEFs in a word of length* n.

Special cases

Lemma

There are at most n occurrences of maximal-exponent factors in a word of length n, whenever the maximal exponent is greater than 1.5*.*

Theorem

Every length n word contains at most n occurrences of MEFs whenever the length of these factors is not a multiple of their longest border.

Lower bound – $5n/6$

Lower bound – $5n/6$

Construct the infinite word $\Omega = \prod_{i=1}^{\infty} \left(\prod_{j=1}^{8} u_{(j,i)} \right)$, from the sequence:

$$u_{(1,i)} = \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_1\ \mathtt{a}_2\ \mathtt{d}_1\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_1\ \mathtt{f}_{1,i}$$
$$u_{(2,i)} = \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_2\ \mathtt{a}_2\ \mathtt{d}_2\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_1\ \mathtt{f}_{2,i}$$
$$u_{(3,i)} = \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_3\ \mathtt{a}_2\ \mathtt{d}_3\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_2\ \mathtt{f}_{3,i}$$
$$u_{(4,i)} = \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_4\ \mathtt{a}_2\ \mathtt{d}_4\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_2\ \mathtt{f}_{4,i}$$
$$u_{(5,i)} = \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_1\ \mathtt{a}_2\ \mathtt{d}_5\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_3\ \mathtt{f}_{5,i}$$
$$u_{(6,i)} = \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_2\ \mathtt{a}_2\ \mathtt{d}_6\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_3\ \mathtt{f}_{6,i}$$
$$u_{(7,i)} = \mathtt{a}_1\ \mathtt{b}_1\ \mathtt{c}_3\ \mathtt{a}_2\ \mathtt{d}_7\ \mathtt{a}_3\ \mathtt{b}_2\ \mathtt{e}_4\ \mathtt{f}_{7,i}$$
$$u_{(8,i)} = \mathtt{a}_1\ \mathtt{b}_3\ \mathtt{c}_4\ \mathtt{a}_2\ \mathtt{d}_8\ \mathtt{a}_3\ \mathtt{b}_4\ \mathtt{e}_4\ \mathtt{f}_{8,i}$$

Proposition

The ratio between the length of the prefixes of Ω and the number of occurrences of its maximal-exponent factors they contain tends to $5/6$.

Lower bound – fixed alphabet

Note that the maximal exponent of factors in Ω is $10/9$ and that its construction can be extended to whatever exponent of the form $(2^\ell + 2)/(2^\ell + 1)$, in a similar fashion.

Lower bound – fixed alphabet

Note that the maximal exponent of factors in Ω is $10/9$ and that its construction can be extended to whatever exponent of the form $(2^\ell + 2)/(2^\ell + 1)$, in a similar fashion.

Observe that letters $\mathtt{f}_{j,i}$ occurring in Ω can be drawn from an 11-letter alphabet disjoint from Σ.

Thanks

Supported by:

- P.R.I.M.E. programme of DAAD co-funded by BMBF and EU's 7th Framework Programme (grant 605728)
- Newton International Fellowship with funds from the Royal Society and the British Academy