Discovering Hidden Repetitions in Words

Pawel Gawrychowki[a], Florin Manea[b], Dirk Nowotka[b]

[a]Max-Planck Institut für Informatik Saarbrücken
[b]Christian-Albrechts-Universität zu Kiel

Ilmenau, September 2013

Notations

Word over V: $w = w[1]...w[n]$ – a finite concatenation of letters $w[i] \in V$

Factor of a word: $w[i..j] = w[i] \cdots w[j]$ – elements from position i to j

Let $f : V^* \to V^*$. We say that f is:

morphism: $f(xy) = f(x)f(y)$ for all $x, y \in V^*$,
in particular: $f(w) = f(w[1]) \cdots f(w[n])$;
non-erasing: $f(a) \neq \lambda$ for all $a \in V$;
uniform: $|f(a)| = k$ for all $a \in V$;
literal: $|f(a)| = 1$ for all $a \in V$;

antimorphism: $f(xy) = f(y)f(x)$ for all $x, y \in V^*$,
in particular: $f(w) = f(w[n]) \cdots f(w[1])$

Length type of f: the array $(|f(a)|)_{a \in V}$

Pseudo-repetitions: Initial motivation

Tandem repeats in DNA sequences :

$$\ldots \mathrm{ACT} \quad \mathrm{ACT} \quad \mathrm{ACT} \ldots$$

Used in genetics to determine an individual's inherited treats, to determine parentage, etc.

Pseudo-repetitions: Initial motivation

Tandem repeats in DNA sequences :

$$\ldots\mathsf{ACT}\quad\mathsf{ACT}\quad\mathsf{ACT}\ldots$$

Used in genetics to determine an individual's inherited treats, to determine parentage, etc.

Inverted repeats in DNA sequences:

$$\ldots\mathsf{AAATCGG}\ldots\mathsf{CCGATTT}\ldots$$

Important genetic elements for genome instability; may play role in DNA rearrangement reactions.
A sequence and its complement encode (almost) the same information.

Pseudo-repetitions: Initial motivation

Tandem repeats in DNA sequences :

. . . ACT ACT ACT. . .

Used in genetics to determine an individual's inherited treats, to determine parentage, etc.

Inverted repeats in DNA sequences:

. . . AAATCGG . . . CCGATTT. . .

Important genetic elements for genome instability; may play role in DNA rearrangement reactions.
A sequence and its complement encode (almost) the same information.

Czeizler, Kari, Seki. On a special class of primitive words. TCS, 2010.:
Pseudo-repetitions: generalised tandem repeats, one sequence is followed by consecutive occurrences of either its copy or of its reversed complement.

. . . AAATCGG AAATCGG CCGATTT AAATCGG . . .

Pseudo-repetitions

A word w is

- *repetition*: $w = t^n$, for some proper prefix t (called root)
  primitive word: not a repetition.
- f is an anti-/morphism
  f-repetition: $w \in t\{t, f(t)\}^*$, for some proper prefix t (called root)
  f-primitive word: not an f-repetition.

Pseudo-repetitions

A word w is

- *repetition*: $w = t^n$, for some proper prefix t (called root)
  primitive word: not a repetition.
- f is an anti-/morphism
  f-repetition: $w \in t\{t, f(t)\}^*$, for some proper prefix t (called root)
  f-primitive word: not an f-repetition.

Example

ACGTAC

- *primitive* from the classical point of view

Pseudo-repetitions

A word w is

- *repetition*: $w = t^n$, for some proper prefix t (called root)
  primitive word: not a repetition.
- f is an anti-/morphism
  f-repetition: $w \in t\{t, f(t)\}^*$, for some proper prefix t (called root)
  f-primitive word: not an f-repetition.

Example

ACGTAC

- *primitive* from the classical point of view
- *f-primitive* for morphism f: $f(A) = T$, $f(C) = G$, $f^2 = 1$.

Pseudo-repetitions

A word w is

- *repetition*: $w = t^n$, for some proper prefix t (called root)
  primitive word: not a repetition.
- f is an anti-/morphism
  f-repetition: $w \in t\{t, f(t)\}^*$, for some proper prefix t (called root)
  f-primitive word: not an f-repetition.

Example

$$ACGTAC$$

- *primitive* from the classical point of view
- *f-primitive* for morphism f: $f(A) = T$, $f(C) = G$, $f^2 = 1$.
- *f-power* for antimorphism f: $f(A) = T$, $f(C) = G$, $f^2 = 1$:

$$ACGTAC = AC \cdot f(AC) \cdot AC$$

Why Pseudo-repetitions?

Pseudo-repetitions: words with intrinsic (yet, hidden) repetitive structure. Extend both repetitions and palindromic structures.

Repetitions and palindromes: central in combinatorics on words and applications!

Why Pseudo-repetitions?

Pseudo-repetitions: words with intrinsic (yet, hidden) repetitive structure. Extend both repetitions and palindromic structures.
Repetitions and palindromes: central in combinatorics on words and applications!

Originated from biology (Watson-Crick complement: antimorphic involution)

Why Pseudo-repetitions?

Pseudo-repetitions: words with intrinsic (yet, hidden) repetitive structure. Extend both repetitions and palindromic structures.
Repetitions and palindromes: central in combinatorics on words and applications!

Originated from biology (Watson-Crick complement: antimorphic involution)
Such structures appear also in music: ternary song form, same fragment repeated on different pitches.

Why Pseudo-repetitions?

Pseudo-repetitions: words with intrinsic (yet, hidden) repetitive structure. Extend both repetitions and palindromic structures.
Repetitions and palindromes: central in combinatorics on words and applications!

Originated from biology (Watson-Crick complement: antimorphic involution)
Such structures appear also in music: ternary song form, same fragment repeated on different pitches.

[Kari, Seki. An improved bound for an extension of Fine and Wilf theorem, and its optimality. Fundam. Informat. 2010.]
[Chiniforooshan, Kari, Xu. Pseudopower avoidance. Fundam. Informat., 2012.]
[Blondin Massé, Gaboury, Hallé. Pseudoperiodic words. DLT 2012]
[M., Müller, Nowotka. The avoidability of cubes under permutations. DLT 2012.]
[M., Mercas, Nowotka. F & W theorem and pseudo-repetitions. MFCS 2012.]
[M., Müller, Nowotka. On the Pseudoperiodic Extension of $u^\ell = v^m w^n$. FSTTCS 2013.]
[Xu. A Minimal Periods Algorithm with Applications. CPM 2010]
[Gawrychowski, M., Mercas, Nowotka, Tiseanu. Finding Pseudo-Repetitions. STACS 2013.]
[Gawrychowski, M., Nowotka. Discovering Hidden Repetitions. CiE 2013.]

Pseudo-repetitions

Given $w \in V^*$ and an anti-/morphism f, decide whether w is an f-repetition.

Pseudo-repetitions

Given $w \in V^*$ and an anti-/morphism f, decide whether w is an f-repetition.

Theorem (Gawrychowski, Manea, Mercaş, Nowotka, Tiseanu, STACS 2013)

Given $w \in V^*$ and $f : V^* \to V^*$ a constant size anti-/morphism, we decide whether $w \in t\{t, f(t)\}^+$ in $\mathcal{O}(n \log n)$ time. If f is uniform we only need $\mathcal{O}(n)$ time. $\square$

Pseudo-repetitions

Given $w \in V^*$ and an anti-/morphism f, decide whether w is an f-repetition.

Theorem (Gawrychowski, Manea, Mercaş, Nowotka, Tiseanu, STACS 2013)

Given $w \in V^*$ and $f : V^* \to V^*$ a constant size anti-/morphism, we decide whether $w \in t\{t, f(t)\}^+$ in $\mathcal{O}(n \log n)$ time. If f is uniform we only need $\mathcal{O}(n)$ time. □

Theorem (G., M., M., N., T., STACS 2013)

Given $w \in V^*$ and $f : V^* \to V^*$ be a constant size anti-/morphism, we decide whether $w \in \{t, f(t)\}\{t, f(t)\}^+$ in $\mathcal{O}(n^{1+\frac{1}{\log \log n}} \log n)$ time. If f is non-erasing we solve the problem in $\mathcal{O}(n \log n)$ time, while when f is uniform we only need $\mathcal{O}(n)$ time. □

Hidden-repetitions

Given $w \in V^+$, decide whether there exists an anti-/morphism $f : V^* \to V^*$ and a prefix t of w such that $w \in t\{t, f(t)\}^+$.

Hidden-repetitions

Given $w \in V^+$, decide whether there exists an anti-/morphism $f : V^* \to V^*$ and a prefix t of w such that $w \in t\{t, f(t)\}^+$.

Theorem (Gawrychowski, Manea, Nowotka, CiE 2013)

Given a word w and a vector T of $|V|$ numbers, we decide whether there exists an anti-/morphism f of length type T such that $w \in t\{t, f(t)\}^+$ in $\mathcal{O}(n(\log n)^2)$ time. If T defines uniform anti-/morphisms: $\mathcal{O}(n)$ time.

Theorem (Gawrychowski, Manea, Nowotka, CiE 2013)

For a word $w \in V^+$, deciding the existence of $f : V^* \to V^*$ and a prefix t of w such that $w \in t\{t, f(t)\}^+$ with $|t| \geq 2$ (respectively, $w \in t\{t, f(t)\}\{t, f(t)\}^+$) takes linear time (respectively, is NP-complete) in the general case, is NP-complete for f non-erasing, and takes $\mathcal{O}(n^2)$ time for f uniform.

Repetitive factors (G.M.M.N.T, STACS 2013)

Given a word $w \in V^*$ and f,
(1) Enumerate all (i, j, ℓ), $1 \leq i, j, \ell \leq |w|$, such that there exists t with $w[i..j] \in \{t, f(t)\}^{\ell}$.
(2) Given ℓ, enumerate all (i, j), $1 \leq i, j \leq |w|$, so there exists t with $w[i..j] \in \{t, f(t)\}^{k}$.

Repetitive factors (G.M.M.N.T, STACS 2013)

Given a word $w \in V^*$ and f,
(1) Enumerate all (i,j,ℓ), $1 \leq i,j,\ell \leq |w|$, such that there exists t with $w[i..j] \in \{t, f(t)\}^\ell$.
(2) Given ℓ, enumerate all (i,j), $1 \leq i,j \leq |w|$, so there exists t with $w[i..j] \in \{t, f(t)\}^k$.

Finding the set of all ℓ-repetitive factors (for all ℓ, resp. for a given ℓ):

- f general: $\mathcal{O}(n^{3.5})$, resp. $\mathcal{O}(n^2\ell)$.
- f non-erasing: $\underline{\Theta(n^3)}$, resp. $\underline{\Theta(n^2)}$.
- f literal: $\underline{\Theta(n^2 \log n)}$, resp. $\underline{\Theta(n^2)}$.

Highlighted bounds: no other algorithm performs better in the worst case.

f-patterns

$f : V^* \to V^*$ anti-/morphism.
An unary f-pattern p: element of $\{x, f(x)\}^*$.
If $p \in \{x, f(x)\}^k$, k is the length of p.

f-patterns

$f : V^* \to V^*$ anti-/morphism.
An unary f-pattern p: element of $\{x, f(x)\}^*$.
If $p \in \{x, f(x)\}^k$, k is the length of p.

Instance of p: word obtained by replacing in p the variable x by $t \in V^+$.

f-patterns

$f : V^* \to V^*$ anti-/morphism.
An unary f-pattern p: element of $\{x, f(x)\}^*$.
If $p \in \{x, f(x)\}^k$, k is the length of p.

Instance of p: word obtained by replacing in p the variable x by $t \in V^+$.

Example

If $f = (\cdot)^R$, the mirror image, then $xf(x) = xx^R$ is a pattern whose instances are all palindromes of even length.
If $f = \mathbf{1}$, the identity morphism, then $xf(x) = x^2$ is a pattern whose instances are all squares.

Avoiding f-patterns

Practice and theory: literal functions!

Problem

Given $w \in V^+$, $|w| = n$, $f : V^* \to V^*$ a literal anti-/morphism, and an f-pattern p, decide whether there exists an instance of p occurring in w.

Problem

Given $w \in V^+$, $|w| = n$, $f : V^* \to V^*$ a literal anti-/morphism, and an integer $k > 0$, decide whether there exists a factor v of w with $v \in \{t, f(t)\}^k$ for some $t \in V^+$.

Basic tools

Computational model: RAM with logarithmic word size.

A word u, with $|u| = n$, over $|V| \in \mathcal{O}(n^c)$.

Build in linear time:
– suffix array data structure for u;
– data structures allowing us to answer in $\mathcal{O}(1)$ queries:
"How long is the longest common prefix of $u[i..n]$ and $u[j..n]$?", denoted $LCPref_u(i, j)$.

Basic tools

Computational model: RAM with logarithmic word size.

A word u, with $|u| = n$, over $|V| \in \mathcal{O}(n^c)$.

Build in linear time:
– suffix array data structure for u;
– data structures allowing us to answer in $\mathcal{O}(1)$ queries:
"How long is the longest common prefix of $u[i..n]$ and $u[j..n]$?", denoted $LCPref_u(i,j)$.

In our case:

- w is the input word,
- f a fixed anti-/morphism,
- $u = wf(w)$, $|u| \in \mathcal{O}(|w|)$.

Basic tools

Computational model: RAM with logarithmic word size.

A word u, with $|u| = n$, over $|V| \in \mathcal{O}(n^c)$.

Build in linear time:
– suffix array data structure for u;
– data structures allowing us to answer in $\mathcal{O}(1)$ queries:
"How long is the longest common prefix of $u[i..n]$ and $u[j..n]$?", denoted $LCPref_u(i,j)$.

In our case:

- w is the input word,
- f a fixed anti-/morphism,
- $u = wf(w)$, $|u| \in \mathcal{O}(|w|)$.
- Constant time: does $w[i..j]$ / $f(w[i..j])$ occur at position s in w?

g-factorisation

$g : V^* \to V^*$ literal anti-/morphism, $w \in V^*$.
The g-factorisation of w is defined as follows. We factor $w = u_1 \cdots u_r$ if the following hold for all $i \geq 1$:

- If letter a occurs in w immediately after $u_1 \cdots u_{i-1}$ and neither a or $g(a)$ appeared in $u_1 \cdots u_{i-1}$, then $u_i = a$.
- Otherwise, u_i is the longest word such that $u_1 \cdots u_{i-1} u_i$ is a prefix of w and u_i or $g(u_i)$ occurs at least once as a factor in $u_1 \cdots u_{i-1}$.

g-factorisation

$g : V^* \to V^*$ literal anti-/morphism, $w \in V^*$.
The g-factorisation of w is defined as follows. We factor $w = u_1 \cdots u_r$ if the following hold for all $i \geq 1$:

- If letter a occurs in w immediately after $u_1 \cdots u_{i-1}$ and neither a or $g(a)$ appeared in $u_1 \cdots u_{i-1}$, then $u_i = a$.
- Otherwise, u_i is the longest word such that $u_1 \cdots u_{i-1} u_i$ is a prefix of w and u_i or $g(u_i)$ occurs at least once as a factor in $u_1 \cdots u_{i-1}$.

Lemma

If g is a literal anti-/morphism we can compute the g-factorisation of a word w of length n in time $\mathcal{O}(n)$.

(Practical consequence: fast identification of inverted repeats in DNA, when g models the Watson-Crick complement.)

f-repetitions and f-factorisations, f morphism

Lemma

Let f be a literal morphism, w a word, and p a pattern of length $k \geq 2$, such that $p \neq x^{k-1}f(x)$. Let $w = u_1 \cdots u_r$ be the f-factorisation of w and consider all instances of p. Then for any instance $w[i..j]$ with $|u_1 \cdots u_{h-1}| < j \leq |u_1 \cdots u_h|$ we have two mutually exclusive possibilities:

1. $i > |u_1 \cdots u_{h-1}|$, *and we call $w[i..j]$ a* secondary *instance, completely contained in u_h,*
2. $j - i + 1 \leq k(|u_{h-1}| + |u_h|)$, *and we call $w[i..j]$ a* crossing *instance.*

Furthermore, the leftmost instance of the pattern is crossing.

f-repetitions and $\mathbf{1}$-factorisations

Lemma

Let f be a literal anti-/morphism, w a word, and p a pattern of length $k \geq 3$, such that $p \notin \{x^{k-1}f(x), f(x)^{k-1}x\}$. Let $w = u_1 \cdots u_r$ be the $\mathbf{1}$-factorisation of w and consider all instances of the pattern p. Then for any such instance $w[i..j]$ with $|u_1 \cdots u_{h-1}| < j \leq |u_1 \cdots u_h|$ we have two mutually exclusive possibilities:

1. *$i > |u_1 \cdots u_{h-1}|$, and $w[i..j]$ is a secondary instance, completely contained in u_h,*
2. *$j - i + 1 \leq k(|u_{h-1}| + |u_h|)$, and $w[i..j]$ is a crossing instance.*

Furthermore, the leftmost instance of the pattern is crossing.

Finding the instances of p, for antimorphic f

Instances of x^k, $xf(x)$, $f(x)x$, and $f(x)^k$ are found in linear time.

Finding the instances of p, for antimorphic f

Instances of x^k, $xf(x)$, $f(x)x$, and $f(x)^k$ are found in linear time.

Lemma

Let f be a literal antimorphism, w be an **1***-factorized word, and p a pattern of length $k \geq 3$, such that $p \notin \{x^{k-1}f(x), f(x)x^{k-1}, f(x)^{k-1}x, xf(x)^{k-1}, x^k, f(x)^k\}$. We can output a crossing instance (as a pair of indices) of the pattern in $\mathcal{O}(nk^2)$ time.*

Finding the instances of p, for antimorphic f

Instances of x^k, $xf(x)$, $f(x)x$, and $f(x)^k$ are found in linear time.

Lemma

Let f be a literal antimorphism, w be an $\mathbf{1}$-factorized word, and p a pattern of length $k \geq 3$, such that $p \notin \{x^{k-1}f(x), f(x)x^{k-1}, f(x)^{k-1}x, xf(x)^{k-1}, x^k, f(x)^k\}$. We can output a crossing instance (as a pair of indices) of the pattern in $\mathcal{O}(nk^2)$ time.

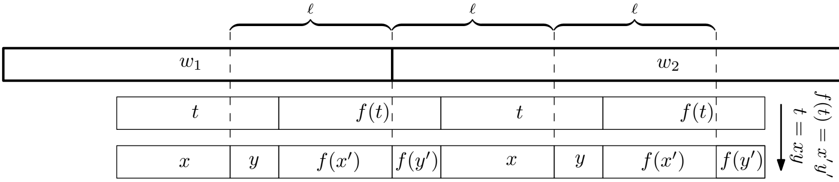

Figure : Finding $tf(t)tf(t)$ in the catenation of two words.

Finding the instances of p, for antimorphic f

The remaining cases: p is $x^{k-1}f(x)$ or $xf(x)^{k-1}$ (and symmetrical).

Finding the instances of p, for antimorphic f

The remaining cases: p is $x^{k-1}f(x)$ or $xf(x)^{k-1}$ (and symmetrical).

Solution ($p = x^{k-1}f(x)$): find a position i such that the pseudopalindromic radius at i is at least as long as the length of the shortest word whose k-th power is a suffix of $w[1..i-1]$.

Finding the instances of p, for antimorphic f

The remaining cases: p is $x^{k-1}f(x)$ or $xf(x)^{k-1}$ (and symmetrical).

Solution ($p = x^{k-1}f(x)$): find a position i such that the pseudopalindromic radius at i is at least as long as the length of the shortest word whose k-th power is a suffix of $w[1..i-1]$.

Lemma

Given a word w of length n and $k \leq 2$, we can compute for each position i the smallest $\ell \leq 1$ such that $w[i-k\ell+1..i]$ is a power of $w[i-\ell+1..i]$, in $\mathcal{O}(n)$ total time.

Result:

Theorem

Given a word $w \in V^*$, with $|w| = n$, a literal anti-/morphism $f : V^* \to V^*$, and an f-pattern p of length k, we can decide whether w contains an instance of p in $\mathcal{O}(nk^2)$ time; for a fixed pattern p, the problem can be solved in linear time.

Result:

THEOREM

Given a word $w \in V^*$, with $|w| = n$, a literal anti-/morphism $f : V^* \to V^*$, and an f-pattern p of length k, we can decide whether w contains an instance of p in $\mathcal{O}(nk^2)$ time; for a fixed pattern p, the problem can be solved in linear time.
If f is bijective, then the problem can be solved in $\mathcal{O}(n \log n)$ time.

Finding k-repetitions, for antimorphic f

Lemma

If w contains $\max(k, 3)$ pseudopalindromes of length ℓ starting at positions $s, s + \delta_1, s + \delta_2, \ldots$ with all $\delta_i \leq \frac{\ell}{4}$, then w has a factor r^k with $r = f(r)$. Accordingly, w contains an instance of any pattern of length k.

Finding k-repetitions, for antimorphic f

Lemma

If w contains $\max(k, 3)$ pseudopalindromes of length ℓ starting at positions $s, s + \delta_1, s + \delta_2, \ldots$ with all $\delta_i \leq \frac{\ell}{4}$, then w has a factor r^k with $r = f(r)$. Accordingly, w contains an instance of any pattern of length k.

1. Look for instances of x^k, $f^k(x)$, $f(x)^{k-1}x$, $xf(x)^{k-1}$, $x^{k-1}f(x)$, or $f(x)x^{k-1}$...

Finding k-repetitions, for antimorphic f

Lemma

If w contains $\max(k,3)$ pseudopalindromes of length ℓ starting at positions $s, s+\delta_1, s+\delta_2, \ldots$ with all $\delta_i \leq \frac{\ell}{4}$, then w has a factor r^k with $r = f(r)$. Accordingly, w contains an instance of any pattern of length k.

1. Look for instances of x^k, $f^k(x)$, $f(x)^{k-1}x$, $xf(x)^{k-1}$, $x^{k-1}f(x)$, or $f(x)x^{k-1}$...

2. Using the **1**-factorisation $w = u_1 \cdots u_r$, and the fact that there cannot be too many pseudo-palindromes, we generate all instances of $xf(x)$ "near" the border between u_{h-1} and u_h...

Finding k-repetitions, for antimorphic f

Lemma

If w contains $\max(k, 3)$ pseudopalindromes of length ℓ starting at positions $s, s+\delta_1, s+\delta_2, \ldots$ with all $\delta_i \leq \frac{\ell}{4}$, then w has a factor r^k with $r = f(r)$. Accordingly, w contains an instance of any pattern of length k.

1. Look for instances of x^k, $f^k(x)$, $f(x)^{k-1}x$, $xf(x)^{k-1}$, $x^{k-1}f(x)$, or $f(x)x^{k-1}$...
2. Using the **1**-factorisation $w = u_1 \cdots u_r$, and the fact that there cannot be too many pseudo-palindromes, we generate all instances of $xf(x)$ "near" the border between u_{h-1} and u_h...
3. Try to construct a full instance of the pattern around such an instance of $xf(x)$.

Result:

Theorem

Given a word $w \in V^*$*, with* $|w| = n$*, a literal anti-/morphism* $f : V^* \to V^*$*, and a positive integer* k*, we can decide whether* w *contains a factor of the form* $\{t, f(t)\}^k$*, for some word* t*, in* $\mathcal{O}(nk^2)$ *time; for a constant* k*, the problem can be solved in linear time.*

Result:

Theorem

Given a word $w \in V^*$, with $|w| = n$, a literal anti-/morphism $f : V^* \to V^*$, and a positive integer k, we can decide whether w contains a factor of the form $\{t, f(t)\}^k$, for some word t, in $\mathcal{O}(nk^2)$ time; for a constant k, the problem can be solved in linear time.
If f is bijective, then we can compute the maximum k such that w contains a factor of the form $\{t, f(t)\}^k$, for some word t, in $\mathcal{O}(n \log n)$ time.

Thank you!

Thank you!

10^{th} International Conference on WORDS

Kiel, 2015