Combinatorics on Words
10th International Conference
WORDS 2015

Florin Manea, Dirk Nowotka
(Editors)

14 - 17 September 2015, Kiel

Kiel Computer Science Series (KCSS) 2015/5 v1.0 dated 2015-8-27

ISSN 2193-6781 (print version)
ISSN 2194-6639 (electronic version)

Electronic version, updates, errata available via https://www.informatik.uni-kiel.de/kcss

The author can be contacted via http://zs.uni-kiel.de

Published by the Department of Computer Science, Kiel University

Department of Computer Science, Kiel University

Please cite as:

▻ F. Manea, D. Nowotka. *Combinatorics on Words, 10th International Conference WORDS 2015. Local Proceedings*. Department of Computer Science, 2015. Kiel University.

```
@proceedings{
booktitle = {Combinatorics on Words, 10th International Conference WORDS 2015.
Local Proceedings},
editor = {Florin Manea, Dirk Nowotka},
publisher = {Department of Computer Science, CAU Kiel},
year = {2015},
number = {2015/5},
series = {Kiel Computer Science Series}
}
```

About this Series

The Kiel Computer Science Series (KCSS) covers dissertations, habilitation theses, lecture notes, textbooks, surveys, collections, handbooks, etc. written at the Department of Computer Science at Kiel University. It was initiated in 2011 to support authors in the dissemination of their work in electronic and printed form, without restricting their rights to their work. The series provides a unified appearance and aims at high-quality typography. The KCSS is an open access series; all series titles are electronically available free of charge at the department's website. In addition, authors are encouraged to make printed copies available at a reasonable price, typically with a print-on-demand service.

Please visit `http://www.informatik.uni-kiel.de/kcss` for more information, for instructions how to publish in the KCSS, and for access to all existing publications.

Preface

This volume contains the Local Proceedings of the Tenth International Conference on WORDS, that took place at the Kiel University, Germany, from the 14th to the 17th September 2015. WORDS is the main conference series devoted to the mathematical theory of words, and it takes place every two years. The first conference in the series was organised in 1997 in Rouen, France, with the following editions taking place in Rouen, Palermo, Turku, Montreal, Marseille, Salerno, Prague, and Turku.

The main object in the scope of the conference, words, are finite or infinite sequences of symbols over a finite alphabet. They appear as natural and basic mathematical model in many areas, theoretical or applicative. Accordingly, the WORDS conference is open to both theoretical contributions related to combinatorial, algebraic, and algorithmic aspects of words, as well as to contributions presenting application of the theory of words, for instance, in other fields of computer science, linguistics, biology and bioinformatics, or physics.

For the second time in the history of WORDS, after the 2013 edition, a refereed proceedings volume was published in Springer's Lecture Notes in Computer Science series. In addition, this local proceedings volume was published in the Kiel Computer Science Series of the Kiel University. Being a conference at the border between theoretical computer science and mathematics, WORDS tries to capture in its two proceedings volumes the characteristics of the conferences from both these worlds. While the Lecture Notes in Computer Science volume was dedicated to formal contributions, this local proceedings volume allows, in the spirit of mathematics conferences, the publication of several contributions informing on current research and work in progress in areas closely connected to the core topics of WORDS. All the papers, the ones published in the Lecture Notes in Computer Science proceedings volume or the ones from this volume, were refereed to high standards by the members of the Program Committee. Following the conference, a special issue of the Theoretical Computer

Science journal will be edited, containing extended versions of papers from both proceedings volumes.

In total, the conference hosted 18 contributed talks. The papers on which 14 of these talks were based, were published in th LNCS volume; the other 4 are published in this volume. In addition to the contributed talks, the conference program included six invited talks given by leading experts in the areas covered by the WORDS conference: Jörg Endrullis (Amsterdam), Markus Lohrey (Siegen), Jean Néraud (Rouen), Dominique Perrin (Paris), Michaël Rao (Lyon), Thomas Stoll (Nancy). WORDS 2015 was the tenth conference in the series, so we were extremely happy to welcome, as invited speaker at this anniversary edition, Jean Néraud, one of the initiators of the series and the main organiser of the first two editions of this conference.

We thank all the invited speakers and all the authors of submitted papers for their contributions to the the success of the conference.

We are grateful to the members of the Program Committee for their work that lead to the selection of the contributed talks, and, implicitly, of the papers published in this volume. They were assisted in their task by a series of external referees, gratefully acknowledged below. The submission and reviewing process used the Easychair system; we thank Andrej Voronkov for this system which facilitated the work of the Programme Committee and the editors considerably. We grateful thank Gheorghe Iosif for designing the logo, poster, and banner of WORDS 2015; the logo of the conference can be seen on the front cover of this book. We also thank the editors of the Kiel Computer Science Series, especially Lasse Kliemann, for their support in editing this volume. Finally, we thank the Organising Committee of WORDS 2015 for ensuring the smooth run of the conference.

Kiel, September 2015

Florin Manea
Dirk Nowotka

Committees

Program Committee

James Currie	University of Winnipeg, Canada
Stepan Holub	Charles University in Prague, Czech Republic
Juhani Karhumäki	University of Turku, Finland
Manfred Kufleitner	University of Stuttgart, Germany
Gad Landau	University of Haifa, Israel
Dirk Nowotka	Kiel University, Germany (PC-Chair)
Wojciech Plandowski	University of Warsaw, Poland
Antonio Restivo	University of Palermo, Italy
Michel Rigo	University of Liège, Belgium
Mikhail Volkov	Ural State University, Russia
Luca Zamboni	University Lyon 1, France

Additional Reviewers

Allouche, Jean-Paul	Glen, Amy	Saarela, Aleksi
Amit, Mika	Hadravova, Jana	Sebastien, Labbe
Badkobeh, Golnaz	Leroy, Julien	Sheinwald, Dafna
Bucci, Michelangelo	Manea, Florin	Smyth, William F.
Charlier, Emilie	Mantaci, Sabrina	Stipulanti, Manon
De Luca, Alessandro	Mercas, Robert	Szabados, Michal
Dekking, Michel	Prodinger, Helmut	Sýkora, Jiří
Epifanio, Chiara	Puzynina, Svetlana	Widmer, Steven
Ferenczi, Sebastien	Rozenberg, Liat	Zaroda, Artur
Fici, Gabriele		

Organising Committee

Maike Bradler	Dirk Nowotka
Florin Manea	Philipp Sieweck

Contents

Primitive Roots of Bi-periodic Infinite Pictures

Nicolas Bacquey

GREYC - Université de Caen Basse-Normandie / ENSICAEN / CNRS
Campus Côte de Nacre, Boulevard du Maréchal Juin
CS 14032 CAEN cedex 5, FRANCE

Abstract

This paper defines and studies the notion of *primitive root* of a bi-periodic infinite picture, that is a rectangular pattern that tiles the bi-periodic picture and contains exactly one representative of each equivalence class of its pixels. This notion extends to dimension 2 the notion of primitive root of a bi-infinite periodic word.

We prove that, for each bi-periodic infinite picture P,

- ▻ there exists at least one primitive root of P;
- ▻ there are at most two ordered pairs of positive integers (m, n) such that every primitive root of P has size $m \times n$;
- ▻ for each such pair (m, n), every rectangular pattern of size $m \times n$ extracted from P is a primitive root of P.

We also discuss some additional properties of primitive roots.

1 Introduction: Primitive words in dimension 1 and 2

In the field of formal languages, primitive words are finite words that are not a power of a smaller word. These words are a well studied subject, with an array of open problems related to them. For instance, it is unknown if the language of all primitive words is context-free (see *e.g.* [5] or [7] for more matter on primitive languages). In this paper, we will define an extension of that notion over words of dimension 2, *i.e.* pictures over a finite alphabet. It is important to note that we will consider rectangular words *as part of a bi-periodic, infinite picture* instead of independently from their surroundings. Informally speaking, we will say that primitive rectangular words will be the smallest rectangular words with which we will be able to rebuild the whole infinite picture by translation.

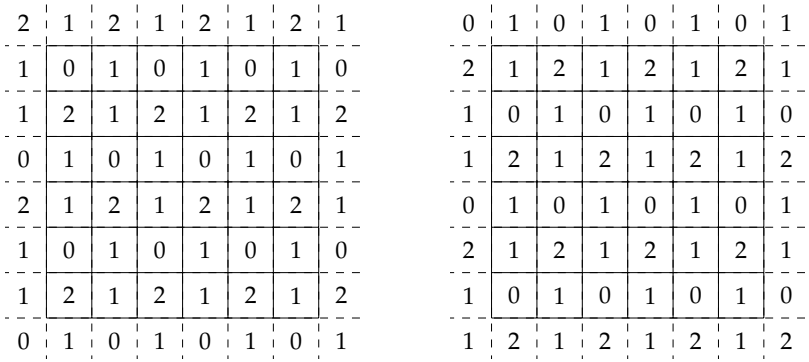
A trivial extension of the notion of primitive words would be to say that a rectangular word is primitive if it is primitive in both directions. However, our twist in the definition will allow us to consider a broader array of primitive words than this trivial extension.

The work presented in this paper originated from the field of Cellular Automata, which are a massively parallel computational model (see [4] or [6]). Our initial goal was to design an algorithm able to perform leader election over infinite periodical pictures [1, 2]. We noticed that the set of leaders of an infinite picture constitute a lattice, and that this lattice could be used to delimit finite rectangular words.

Those particular words, which are the *primitive roots* we will discuss in this article, appear to have very interesting properties that closely relate to formal language theory. We will first formally define those primitive roots, then we will give a tight upper bound to their number (they are not unique up to a shift, as it happens with languages of dimension 1). Finally, we will discuss some of their most interesting properties.

2 Context and definitions

We will now introduce a few definitions that will lead to the proper definition of a primitive root of a bi-dimensional picture.



(a) The original picture P

(b) A new picture $\sigma_h^{-1} \circ \sigma_v^{-1}(P)$

Figure 1. Illustration of the shift function over a picture.

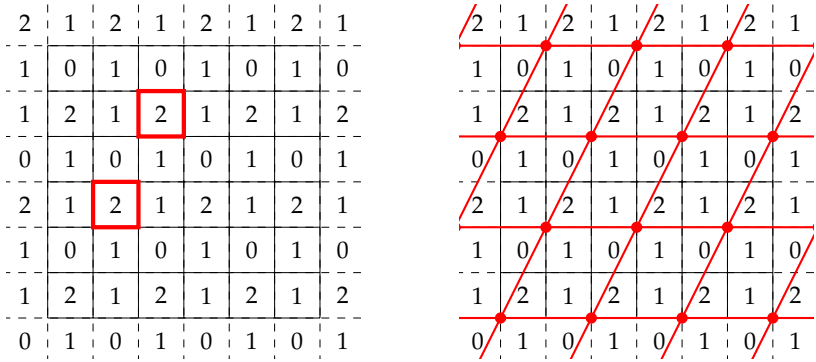
2.1 Definition (pictures)**.** Let Σ be a finite alphabet, we call *picture* a function $P : \mathbb{Z}^2 \to \Sigma$. We say that a picture P is *bi-periodic* if there is a pair of non-collinear vectors $(x_0, y_0), (x_1, y_1) \in \mathbb{Z}^2$ called a *period* of P such that $\forall (x, y) \in \mathbb{Z}^2$:
$P(x + x_0, y + y_0) = P(x, y)$
$P(x + x_1, y + y_1) = P(x, y)$.
In the context of pictures, an element $p \in \mathbb{Z}^2$ is called a *pixel*.

All along this article, every picture we talk about will be bi-periodic, except when noted otherwise.

2.2 Definition (shift functions)**.** We introduce the *horizontal shift function* σ_h and the *vertical shift function* σ_v defined over pictures as follows :
$\forall (x, y) \in \mathbb{Z}^2$
$\sigma_h(P)(x, y) = P(x + 1, y)$
$\sigma_v(P)(x, y) = P(x, y + 1)$.

Figure 1 illustrates the action of these functions over a bi-periodic picture. It is immediate to see that those functions are invertible, and that they commute with each other.

2.3 Definition (equivalent pixels)**.** We say that two pixels $p_1 = (x_1, y_1)$ and



(a) Two equivalent pixels. **(b)** The equivalent pixels lattice.

Figure 2. Similar pixels and their induced lattice.

$p_2 = (x_2, y_2)$ of a picture P are *equivalent* if the transformation that translates p_1 onto p_2 leaves the picture unchanged, *i.e.* if $\sigma_h^{x_2-x_1} \circ \sigma_v^{y_2-y_1}(P) = P$. In that case, we note $p_1 \sim p_2$.

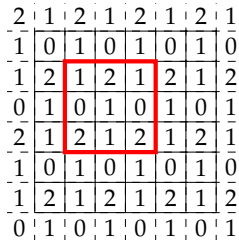
We note that this definition actually corresponds to an equivalence relation. It is easy to see that the following lemma holds for equivalence classes of pixels:

2.4 Lemma. *For any bi-periodic picture P, there exists a finite number of equivalence classes of pixels of that picture. Moreover, each of these equivalence classes contains an infinite number of pixels. Finally, the equivalence class of pixel $(0,0)$ constitutes an* integer lattice of dimension 2, i.e. *a sub-group of $(\mathbb{Z}^2, +)$ (see Figure 2b).*

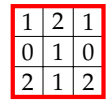
2.5 Definition (rectangular patterns)**.** Let P be a picture over Σ, we call *rectangular pattern* of size $m \times n$ extracted at (x_0, y_0) the following function:

$$\begin{array}{rccc} R_{x_0,y_0} & : & [\![0, m-1]\!] \times [\![0, n-1]\!] & \to & \Sigma \\ & & (x, y) & \mapsto & P(x + x_0, y + y_0) \end{array}$$

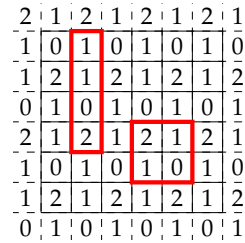
2.6 Definition (primitive root)**.** We say that a rectangular pattern R_{x_0,y_0} of size $m \times n$ is a *primitive root* of the picture P if it contains exactly one



(a) The original picture P



(b) The rectangular pattern of size 3×3 extracted at $(1, 2)$



(c) Primitive roots of picture P

Figure 3. Rectangular patterns in a bi-periodic picture.

representative of each equivalence class of pixels of the picture, *i.e.* $\forall (x, y) \in \mathbb{Z}^2;\ \exists! (x', y') \in [\![0, m-1]\!] \times [\![0, n-1]\!]$ such that $(x, y) \sim (x_0 + x', y_0 + y')$.

Figure 3 gives an example of primitive roots of a picture. The following lemma holds directly from the previous definition :

2.7 Lemma. *All primitive roots of a given picture have the same area, which is the number of equivalence classes of that picture.*

Let us notice that these definitions can be adapted to pictures of dimension 1, *i.e.* words over Σ. It can be easily seen that primitive roots exactly are primitive words in that context.

We also notice that our definition of primitive root is non-constructive. Our first non-trivial result is that these primitive roots indeed exist.

2.8 Theorem (existence of primitive roots)**.** *Let P be a bi-periodic picture, then primitive roots can be extracted from P.*

Theorem 2.8. This proof will use *Hermite normal form* of square matrices, which are a well studied tool of linear algebra. Simply put, an integer matrix H is said to be in Hermite normal form if

- it is lower triangular
- its diagonal entries are positive

▷ in every column, the entries below the diagonal are non-negative and smaller than the ones on the diagonal.

For any integer matrix M, it is known that there exists a unique integer matrix H in Hermite normal form such that $H = U \times M$, where U is unimodular with integer coefficients. We will also use notions related to *integer lattices*, such as the fundamental domain of a lattice. More references about Hermite normal form and lattices can be found in [3].

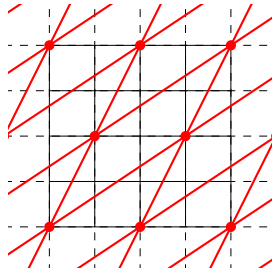
Let us consider the integer lattice $\mathcal{L}$ formed by the equivalence class of pixel $(0,0)$ (See Lemma 2.4). Let us call $\mathcal{B}$ a basis of that lattice (see Figure 4a), and let $\mathcal{B}'$ be the family of vectors whose matrix is the Hermite normal form of the matrix associated with $\mathcal{B}$ (see [3]). Without loss of generality we can assume that $\mathcal{B}' = \begin{pmatrix} \alpha & 0 \\ \beta & \gamma \end{pmatrix}$.
Because of the properties of Hermite transformation (*i.e.* the matrix U is unimodular), it is clear that $\{(\alpha,0),(\beta,\gamma)\}$ is also a basis of $\mathcal{L}$ (see Figure 4b).

Let $\mathcal{F}$ be the fundamental domain of $\mathcal{L}$ associated with basis $\mathcal{B}'$ (see [3]). More precisely, let us consider the pixels within $\mathcal{F}$. Clearly, because $\mathcal{B}'$ is a basis of $\mathcal{L}$, there must be exactly one representative of each equivalence class among them (see Figure 4c).

Let also $\mathcal{R}$ be the rectangular pattern of size $\alpha \times \gamma$ extracted from P at position $(0,0)$ (see Figure 4c). It appears that $\mathcal{R}$ contains exactly the same equivalence classes as $\mathcal{F}$, because each pixel of $\mathcal{R}$ is either a pixel of $\mathcal{F}$ or the translation by a vector $(-\alpha,0)$ of a pixel of $\mathcal{F}$ (that translation preserves equivalence classes, since $(\alpha,0)$ is a vector of $\mathcal{B}'$).

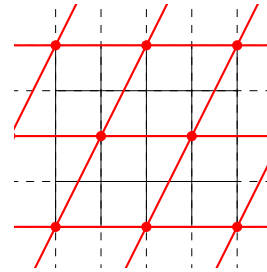
We therefore know that $\mathcal{R}$ contains exactly one representative of each equivalence class of P, which makes it a primitive root. □

We note that the construction of such primitive roots is non-trivial in the general case: The naive algorithm that takes an arbitrary rectangular pattern that contains at least (resp. at most) one representative of each equivalence class and shrinks it (resp. expands it) until it contains exactly one representative does not work. Figure 5 gives counter-examples, in the form of rectangular patterns that contain at least (resp. at most) one representative of each class, and cannot be shrunk (resp. expanded).



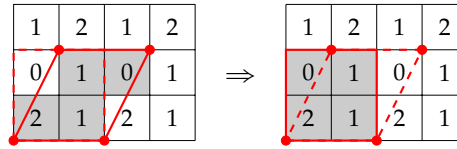
$$\mathcal{B} = \begin{pmatrix} 3 & 2 \\ 1 & 2 \end{pmatrix}$$

(a) The original lattice $\mathcal{L}$ with an unspecified base



$$\mathcal{B}' = \begin{pmatrix} 2 & 0 \\ 1 & 2 \end{pmatrix}$$

(b) The same lattice with an Hermite normal form base



(c) Transforming the fundamental domain $\mathcal{F}$ of $\mathcal{B}'$ into a rectangular pattern $\mathcal{R}$

Figure 4. Illustration of primitive root construction.

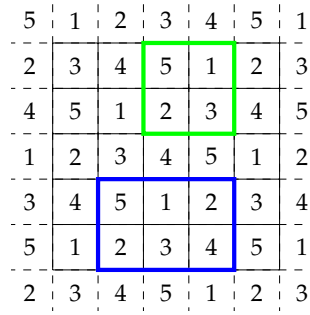


Figure 5. Counter-examples to the naive algorithm.

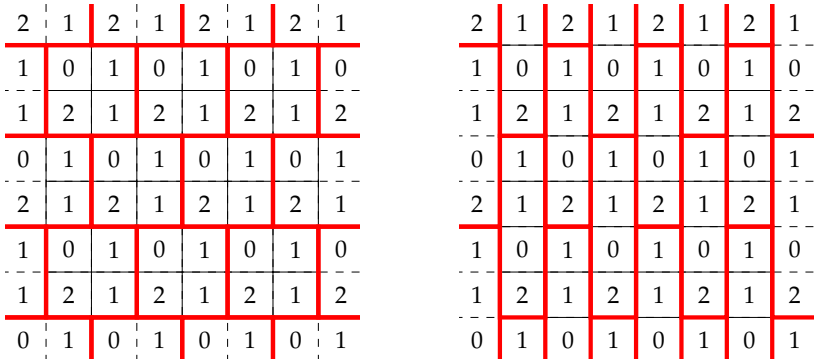


Figure 6. Tilings of a bi-periodic picture by its primitive roots.

3 Main result

Now that we have proved that there exist primitive roots in any bi-periodic picture, we will describe them exactly. In order to achieve this, we will need the following lemma:

3.1 Lemma. *Let $\mathcal{R}$ be a primitive root of a bi-periodic picture P, then $\mathcal{R}$ tiles P by translation.*

This lemma is illustrated by Figure 6, and can be easily proved by noticing that a translation of $\mathcal{R}$ can be constructed around each pixel of P (because $\mathcal{R}$ contains at least one representative of each equivalence class), and that these translations cannot overlap (those representatives must be unique in $\mathcal{R}$). This tiling can also be obtained by copying $\mathcal{R}$ on each point of the lattice $\mathcal{L}$ defined previously.

We can now introduce our second theorem, which gives us a more precise characterization of the primitive roots of a picture.

3.2 Theorem. *let P be a bi-periodic picture, and let $S \subset \mathbb{N}^2$ be the set of all possible sizes for primitive roots of P (more precisely, $S = \{(m,n);\ \exists R_{x,y}$ a primitive root of P of size $m \times n\}$), then:*

$\rhd$ $|S| \leqslant 2$ *(there are at most two different sizes for primitive roots).*

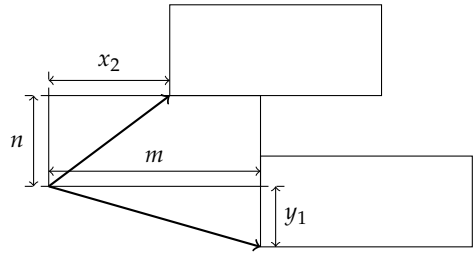


Figure 7. The particular tiling vectors associated to a given primitive root.

▷ $\forall (m,n) \in S;\ \forall (x,y) \in \mathbb{Z}^2$ *if* $R_{x,y}$ *is the rectangular pattern of size* $m \times n$ *extracted from* P *at* (x,y)*, then* $R_{x,y}$ *is a primitive root of* P *(that is, primitive roots can be extracted from anywhere provided that they are of appropriate size).*

Theorem 3.2. We will prove the first point of Theorem 3.2 by associating a matrix in Hermite normal form to each primitive root of picture P, and by considering what it implies.

Let R_{x_0,y_0} be a primitive root of P of size $m \times n$. Thanks to Lemma 3.1, we know that there exists a tiling of P by R. We consider two particular vectors of that tiling, which are illustrated on Figure 7 and defined thereafter:

▷ Let $V_1 = (m, -y_1)$ where y_1 is the smallest positive integer such that $(x_0, y_0) \ \sim \ (x_0 + m, y_0 - y_1)$.

▷ Let $V_2 = (x_2, n)$ where x_2 is the smallest positive integer such that $(x_0, y_0) \ \sim \ (x_0 + x_2, y_0 + n)$.

Because R tiles the picture, it is clear that we have $0 \leqslant y_1 < n$ and $0 \leqslant x_2 < m$. We will now prove that we have either $y_1 = 0$ or $x_2 = 0$.

Indeed, if we have $y_1 \neq 0$ and $x_2 \neq 0$, that would mean there exists a rectangular "hole" of size $x_2 \times y_1$ in the tiling (see Figure 8). Because we have $x_2 < m$ and $y_1 < n$, that means it is impossible to fit a translation of R into that hole, therefore a contradiction with the fact that R tiles the picture. We now have either $V_1 = (m, 0)$ or $V_2 = (0, n)$.

It is important to note that, as they are non-colinear, V_1 and V_2 constitute a basis of the lattice $\mathcal{L}$ associated with P. Up to a re-ordering of the

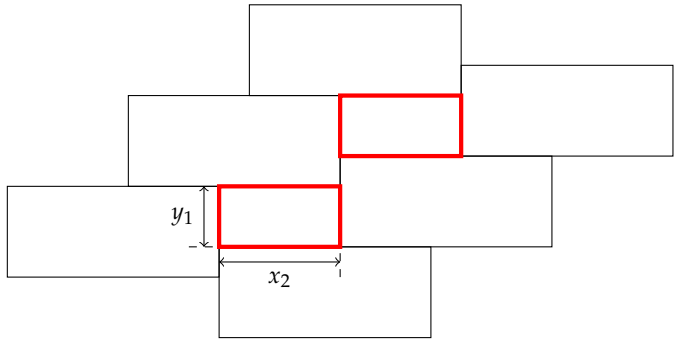


Figure 8. An illustration of what the tiling would look like if $y_1 \neq 0$ and $x_2 \neq 0$.

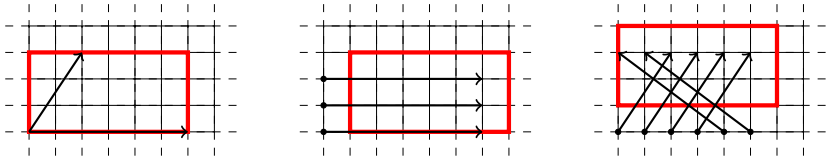
dimensions, we can suppose without loss of generality that $V_1 = (m, 0)$ and $V_2 = (x_2, n)$. Let us consider the matrix $\mathcal{B} = \begin{pmatrix} m & 0 \\ x_2 & n \end{pmatrix}$.

It is the matrix of a basis of $\mathcal{L}$, and it happens to be in Hermite normal form (because $0 \leqslant x_2 < m$). It means that every (m, n) eligible to be the size of a primitive root must be the couple of diagonal coefficients of the matrix of a basis of $\mathcal{L}$ in Hermite normal form (up to a reordering of the dimensions). We know that such a matrix is unique (see [3]), and that there exist 2 reorderings of 2 dimensions ($2! = 2$). Therefore, it means that the couple (m, n) can only have at most two different values, giving us the first point of the theorem.

Now to prove the second point, we will only prove that if R_{x_0,y_0} is a primitive root of size $m \times n$, then the rectangular patterns R'_{x_0+1,y_0} and R''_{x_0,y_0+1} of same size also are. The second point would then automatically follow by induction.

Let therefore R_{x_0,y_0} be a primitive root of P of size $m \times n$. Let also V_1 and V_2 be the particular vectors defined earlier. We assume without loss of generality that $V_1 = (m, 0)$ and $V_2 = (x_2, n)$.

Figure 9 shows that both R'_{x_0+1,y_0} and R''_{x_0,y_0+1} contain the same equivalence classes as R_{x_0,y_0}. Indeed, in both cases there exists a bijection between the equivalence classes of the original pattern and those of the new one; this bijection is a translation by particular vectors which conserve the



(a) The original primitive root and its translation vectors V_1 and V_2.

(b) The new pattern R'_{x_0+1,y_0}

(c) The new pattern R''_{x_0,y_0+1}

Figure 9. Translation of a primitive root preserves its equivalence classes. Here we have $V_1 = (6,0)$ and $V_2 = (2,3)$.

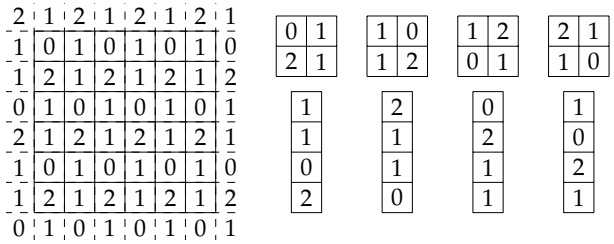


Figure 10. Here are given all the primitive roots of the example picture.

equivalence classes of a pixel.

In the case of R'_{x_0+1,y_0}, the vectors are either V_1 or $(0,0)$ (see Figure 9b). In the case of R''_{x_0,y_0+1}, the vectors are V_2, $V_2 - V_1$ or $(0,0)$ (see Figure 9c). As R'_{x_0+1,y_0} and R''_{x_0,y_0+1} contain the same equivalence classes as R_{x_0,y_0} and also are rectangular patterns, then they also are primitive roots of P. □

Note that the upper bound stated in Theorem 3.2 is tight, as there are pictures for which the primitive roots can have 2 different sizes (in fact, most of them). An example is given on Figure 3c.

The second point of Theorem 3.2 can give us all the primitive roots of a given picture. An example of its application is given on Figure 10.

4 Discussion about the root extracting function

In this section, we will study some interesting properties of the function $\mathcal{F}$ that maps a bi-periodic picture to the set of its primitive roots, or equivalently (as Theorem 3.2 states) the set of sizes of its primitive roots.

More formally, we can say that $\mathcal{F} : \Sigma^{\mathbb{Z}^2} \to (\mathbb{N}^2)^2$. Note that $\mathcal{F}(P)$ is only defined if its argument P is a *periodic* picture.

4.1 Computability of function $\mathcal{F}$

The first and perhaps most interesting result is that function $\mathcal{F}$ is indeed computable, even if its argument is an infinite object (more precisely, an infinite object with finite support, but whose size is unknown). A few computational models can reasonably process an input whose size is infinite, but the computability of this function has been proved with the model of Cellular Automata, which is one of those models (See *e.g.* [4] or [6] for general matters on Cellular Automata).

This computability result will not be extensively discussed here, as it would easily double the size of this article and has already been proved in [2]. However, note that this result could also adapt to classical computational models with finite input, such as Turing Machines. In that case it becomes quasi trivial, because the only relevant way to describe the input picture would be to give one of its periods (not necessarily the shortest one). If the period is known, then a straightforward application of the construction given in the proof of Theorem 2.8 would immediately give the primitive roots of the picture.

4.2 Injectivity in the general case

It is immediate to see that two bi-periodic pictures which are shifts of each other have exactly the same set of primitive roots, so the injectivity of function $\mathcal{F}$ is clearly disproved. However, it would be interesting to see what happens if the pictures are defined up to a shift, which is a reasonable assumption.

It appears that the function $\mathcal{F}$ is not injective either in that case. Indeed, Figure 11 proves it by giving two different bi-periodic pictures that have

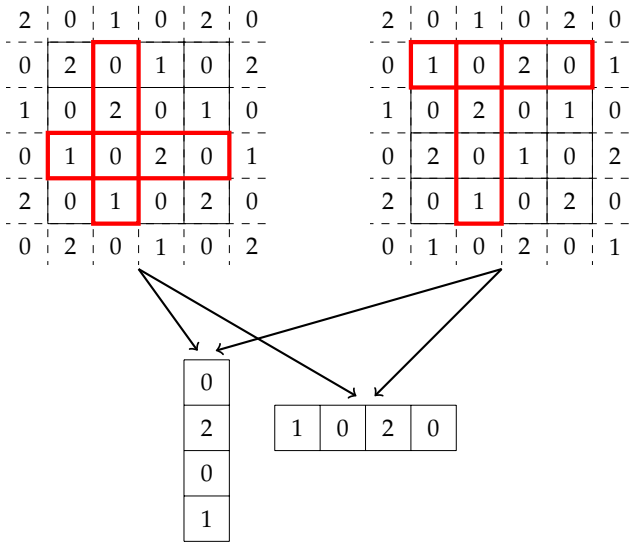


Figure 11. Illustration of the non-injectivity of the root extracting function: both pictures have the same set of primitive roots (all the primitive roots are shifts of the ones presented on the figure).

exactly the same set of primitive roots. We can infer that the mere knowledge of its primitive roots is not sufficient to deduce a whole bi-periodic picture; one would also need the tiling vectors of a given root.

An attentive reader may have noticed that both pictures shown on Figure 11 are rotations of each other, and that the set of their primitive roots is invariant by rotation. This reader could ask if the injectivity of function $\mathcal{F}$ is true if the pictures are defined up to a rotation. Unfortunately, it is not the case, as there exist more complex counterexamples which are not equivalent by rotation.

4.3 Bijectivity in the case of a double root

Now let us consider what happens when there is only one possible size for the primitive root. This case happens when the Hermite normal form of the matrix associated with P is diagonal instead of simply triangular.

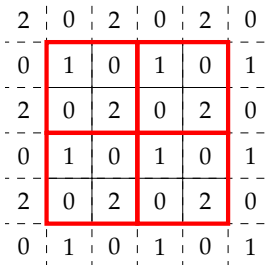


Figure 12. When there is only one possible size for the primitive roots of a picture, the inverse function exists and is trivial.

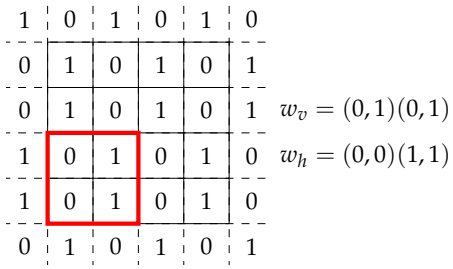


Figure 13. A primitive root can be seen as a one-dimensional vertical word w_v over the alphabet of horizontal tuples of Σ, or as a horizontal word w_h over the alphabet of vertical tuples of Σ.

An example of that case is shown on Figure 12.

If R_{x_0,y_0} is a double root of the picture P of size $m \times n$, it means that the translation vectors associated with R_{x_0,y_0} are horizontal and vetical ($V_1 = (m,0)$ and $V_2 = (0,n)$). In that case, the original picture can be rebuilt P by merely translating R_{x_0,y_0} along V_1 and V_2. This means that $\mathcal{F}$ is bijective if there is a double root, provided the picture is defined up to a translation.

4.4 Properties of the primitive roots

It can be interesting to study the relation between the primitive roots we defined in this article with *primitive words* of dimension 1 (see *e.g.* [7]). In particular, if P is a picture over an alphabet Σ and R_{x_0,y_0} is a root of P of size $m \times n$, R can be seen as a horizontal word over the alphabet Σ^n, or as a vertical word over the alphabet Σ^m (as shown on Figure 13). It appears that at least one of these words is primitive in their particular alphabet (the proof is quite simple, and uses once again the fact that either V_1 or V_2 have a null component).

Conversely, one could ask if any rectangular pattern that is either "horizontally primitive" or "vertically primitive" can be the primitive root of a certain bi-periodic picture. It seems to be the case, but we were not

able to find a formal proof of that property.

5 Conclusion and perspectives

Although the definition of a primitive root is non-constructive, this article succeeds in exhibiting all of them for every bi-periodic picture, and shows some of their properties. Now let us review the remaining points that could lead to future works.

5.1 Open problems

As stated earlier, we still don't know if every two-dimensional word that fits our naive precondition (*i.e.* being either horizontally primitive or vertically primitive) can be the primitive root of a picture. Deciding that property could lead to a characterization of the 2-dimensional language of potential primitive roots. Even if we don't know much about that language, we still have some lower bounds about its recognizability (it obviously is as hard to recognize as the language of primitive words, which is its restriction to dimension 1).

5.2 Primitive roots in higher dimensions

An immediate extension of our work would be to extend it to pictures of dimension higher than 2. All the definitions scale nicely, up to the definition of a multi-dimensional primitive root, as a hyper-parallelogram containing exactly one representative of each equivalence class of pixels. The existence of such primitive roots also holds, due to the same argument used in the proof of Theorem 2.8, only using the Hermite normal form of matrices of higher dimensions.

However, Theorem 3.2 seems harder to prove; it would state that *for a multi-periodic picture of dimension d, there are at most $d!$ possible sizes for its primitive roots*. We know how to construct pictures that have at least $d!$ different sizes of primitive roots ($d!$ is the number of linear orders of the d dimensions). In order to prove that this bound is a maximum, we miss a statement that would be equivalent to "at least one of the

translating vectors have a null component" in higher dimensions. Note that this problem is a purely geometric one; it only relates to the tiling of the space by translations of a hyper-parallelogram, and does not relate to formal languages. The subject of properties of primitive roots of higher dimensions is also left unexplored.

References

[1] N. Bacquey. Complexity classes on spatially periodic cellular automata. In E. W. Mayr and N. Portier, editors, *31st International Symposium on Theoretical Aspects of Computer Science*, volume 25 of *LIPIcs*, pages 112–124, 2014. doi: 10.4230/LIPIcs.STACS.2014.112. URL http://dx.doi.org/10.4230/LIPIcs.STACS.2014.112.

[2] N. Bacquey. Leader election on two-dimensional periodic cellular automata. *currently submitted*, 2015.

[3] H. Cohen. *A Course in Computational Algebraic Number Theory*, volume 138 of *Graduate Texts in Mathematics*. Springer-Verlag Berlin Heidelberg, 1993. ISBN 978-3-540-55640-4. doi: 10.1007/978-3-662-02945-9.

[4] M. Delorme and J. Mazoyer. *Cellular Automata: a parallel model*, volume 460. Springer Science & Business Media, 1998.

[5] P. Dömösi, S. Horváth, and M. Ito. Formal languages and primitive words. *Publ. Math. Debrecen*, 42(3–4):315–321, 1993.

[6] J. Kari. Theory of cellular automata: A survey. *Theoretical Computer Science*, 334(1-3):3–33, 2005. doi: 10.1016/j.tcs.2004.11.021. URL http://dx.doi.org/10.1016/j.tcs.2004.11.021.

[7] H. Petersen. On the language of primitive words. *Theoretical Computer Science*, 161(1):141–156, 1996.

Periodicity of Generalized Pseudostandard Words

Josef Florian, Lubomíra Dvořáková

Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague,
13 Trojanova, 120 00 Praha 2, Czech Republic
{florijos,lubomira.dvorakova}@fjfi.cvut.cz

Abstract

Generalized pseudostandard words were introduced by De Luca and de Luca in [4]. Recently, they have been studied intensively, nevertheless in comparison to the palindromic and pseudopalindromic closure, there are still a lot of open problems concerning generalized pseudopalindromic closure and the associated generalized pseudostandard words. We present here a necessary and sufficient condition for their periodicity over ternary alphabet. More precisely, we describe how the directive bi-sequence of a generalized pseudostandard word has to look like in order to correspond to a periodic word. We extend thus the result from [1] where we found such a condition over binary alphabet. It is interesting that the conditions on periodicity over binary and ternary alphabet are surprisingly different. We state moreover as a conjecture a necessary and sufficient condition for periodicity over any alphabet.

1 Basics from combinatorics on words

Any finite set of symbols is called an alphabet $\mathcal{A}$, the elements are called letters. A (finite) word w over $\mathcal{A}$ is any finite sequence of letters. Its length $|w|$ is the number of letters it contains. The empty word – the neutral element for concatenation of words – is denoted ε and its length is set $|\varepsilon| = 0$. The symbol $\mathcal{A}^*$ stands for the set of all finite words over $\mathcal{A}$. An infinite word $\mathbf{u}$ over $\mathcal{A}$ is any infinite sequence of letters. A finite word w is a factor of the infinite word $\mathbf{u} = u_0u_1u_2\ldots$ with $u_i \in \mathcal{A}$ if there exists an index $i \geqslant 0$ such that $w = u_iu_{i+1}\ldots u_{i+|w|-1}$. The symbol $\mathcal{L}(\mathbf{u})$ is used for the set of factors of $\mathbf{u}$ and is called the language of $\mathbf{u}$, similarly $\mathcal{L}_n(\mathbf{u})$ stands for the set of factors of $\mathbf{u}$ of length n.

Let $w \in \mathcal{L}(\mathbf{u})$. A left extension of w is any word $aw \in \mathcal{L}(\mathbf{u})$, where $a \in \mathcal{A}$. The factor w is called left special if w has at least two left extensions. The (factor) complexity of $\mathbf{u}$ is the map $\mathcal{C}_\mathbf{u} : \mathbb{N} \to \mathbb{N}$ defined as

$$\mathcal{C}_\mathbf{u}(n) = \#\mathcal{L}_n(\mathbf{u}).$$

The following results on complexity come from [8]. If an infinite word is eventually periodic, i.e., it is of the form wv^ω, where w, v are finite words (w may be empty – in such a case we speak about a purely periodic word) and ω denotes an infinite repetition, then its factor complexity is bounded. An infinite word is not eventually periodic – such a word is called aperiodic – if and only if its complexity satisfies: $\mathcal{C}(n) \geqslant n+1$ for all $n \in \mathbb{N}$. If an infinite word $\mathbf{u}$ contains for every length n a left special factor of length n, the complexity is evidently strictly growing, hence $\mathbf{u}$ is aperiodic.

An involutory antimorphism is a map $\vartheta : \mathcal{A}^* \to \mathcal{A}^*$ such that for every $v, w \in \mathcal{A}^*$ it holds $\vartheta(vw) = \vartheta(w)\vartheta(v)$ and moreover ϑ^2 equals identity. It is clear that in order to define an antimorphism, it suffices to provide letter images. There are only two involutory antimorphisms over the alphabet $\{0,1\}$: the reversal (mirror) map R satisfying $R(0) = 0,\ R(1) = 1$, and the exchange antimorphism E given by $E(0) = 1,\ E(1) = 0$. We use the notation $\overline{0} = 1$ and $\overline{1} = 0$, $\overline{E} = R$ and $\overline{R} = E$. There are only four involutory antimorphisms over the alphabet $\{0,1,2\}$: the reversal map R satisfying $R(0) = 0,\ R(1) = 1,\ R(2) = 2$, and three exchange

antimorphisms $E_0,\ E_1,\ E_2$ given by

$$\begin{array}{lll} E_0(0)=0, & E_0(1)=2, & E_0(2)=1 \\ E_1(0)=2, & E_1(1)=1, & E_1(2)=0 \\ E_2(0)=1, & E_2(1)=0, & E_2(2)=2\,. \end{array}$$

Consider an involutory antimorphism ϑ over $\mathcal{A}$. A finite word w is a ϑ-palindrome if $w = \vartheta(w)$. The ϑ-palindromic closure w^{ϑ} of a word w is the shortest ϑ-palindrome having w as prefix. For instance, over binary alphabet $011^R = 0110,\ 011^E = 011001$. We say that an infinite word $\mathbf{u}$ is obtained by the ϑ-palindromic closure using a directive sequence $\Delta = \delta_1\delta_2\ldots$ of letters in $\mathcal{A}$ if w_n is a prefix of $\mathbf{u}$ for every $n \in \mathbb{N}$, where

$$w_0 = \varepsilon \quad \text{and} \quad w_{n+1} = (w_n\delta_{n+1})^{\vartheta}.$$

We speak about the palindromic closure if $\vartheta = R$ and the pseudopalindromic closure if we do not need to specify which antimorphism ϑ is used.

2 Definition of generalized pseudostandard Words

Generalized pseudostandard words form a generalization of infinite words obtained by the palindromic, resp. pseudopalindromic closure; such constructions were described and studied in [3], [5], [7], [4].

2.1 Definition. Let $\mathcal{A}$ be an alphabet and G be the set of all involutory antimorphisms on $\mathcal{A}^*$. Let $\Delta = \delta_1\delta_2\ldots$ and $\Theta = \vartheta_1\vartheta_2\ldots$, where $\delta_i \in \mathcal{A}$ and $\vartheta_i \in G$ for all $i \in \mathbb{N}$. The infinite *generalized pseudostandard word* $\mathbf{u}(\Delta,\Theta)$ is the word whose prefixes w_n are obtained from the recurrence relation

$$\begin{gathered} w_{n+1} = (w_n\delta_{n+1})^{\vartheta_{n+1}}, \\ w_0 = \varepsilon. \end{gathered}$$

The sequence $\Lambda = (\Delta,\Theta)$ is called the *directive bi-sequence* of the word $\mathbf{u}(\Delta,\Theta)$.

Examples of generalized pseudostandard words include: episturmian words (they are obtained by the palindromic closure, thus $\Theta = R^{\omega}$),

the Thue–Morse word [4], standard Rote words [2], and a subclass of generalized Thue–Morse words [6].

In contrast to the palindromic and pseudopalindromic closure, the sequence of prefixes w_n does not have to contain all ϑ-palindromic prefixes of $\mathbf{u}(\Delta, \Theta)$, where ϑ is an involutory antimorphism over $\mathcal{A}$. However, if it is the case, we say that the directive bi-sequence is normalized. In [2], the authors provide an algorithm over binary alphabet for normalization of any directive bi-sequence in such a way that the obtained generalized pseudostandard word remains unchanged.

2.2 Theorem. *Let $\Lambda = (\Delta, \Theta)$ be a directive bi-sequence, where Δ is a sequence over $\{0, 1\}$ and Θ is a sequence over $\{E, R\}$. Then there exists a normalized directive bi-sequence $\widetilde{\Lambda} = (\widetilde{\Delta}, \widetilde{\Theta})$ such that $\mathbf{u}(\Delta, \Theta) = \mathbf{u}(\widetilde{\Delta}, \widetilde{\Theta})$.*

Moreover, in order to normalize the sequence Λ, it suffices firstly to execute the following changes of its prefix (if it is of the corresponding form):

$\triangleright$ $(a\bar{a}, RR) \to (a\bar{a}a, RER)$,

$\triangleright$ $(a^i, R^{i-1}E) \to (a^i\bar{a}, R^iE)$ *for* $i \geqslant 1$,

$\triangleright$ $(a^i\bar{a}\bar{a}, R^iEE) \to (a^i\bar{a}\bar{a}a, R^iERE)$ *for* $i \geqslant 1$,

and secondly to replace step by step from left to right every factor of the form:

$\triangleright$ $(ab\bar{b}, \vartheta\overline{\vartheta\vartheta}) \to (ab\bar{b}b, \vartheta\overline{\vartheta}\vartheta\overline{\vartheta})$,

where $a, b \in \{0, 1\}$ and $\vartheta \in \{E, R\}$.

3 Periodicity of generalized pseudostandard words over ternary alphabet

Let us first recall a sufficient and necessary condition for periodicity of binary generalized pseudostandard words. Let us underline that such words are either purely periodic or aperiodic since they are recurrent (their language is closed under R or E which guarantees that every factor occurs at least twice).

3.1 Theorem ([1]). *Let $\Lambda = (\Delta, \Theta)$ be a directive bi-sequence of a binary generalized pseudostandard word, where Δ is a sequence over $\{0,1\}$ and Θ is a sequence over $\{E,R\}$. The generalized pseudostandard word $\mathbf{u}(\Delta,\Theta)$ is periodic if and only if the following condition is satisfied:*

$$(\exists \vartheta \in \{E,R\})(\exists n_0 \in \mathbb{N})(\forall n > n_0)(\delta_{n+1} = 0 \Leftrightarrow \vartheta_n = \vartheta)\ , \qquad (3.2)$$

where $\Delta = \delta_1\delta_2\ldots$ and $\Theta = \vartheta_1\vartheta_2\ldots$

Let us remark that using Theorem 2.2, it is not difficult to find the normalized version of the directive bi-sequence satisfying (3.2) of a binary generalized pseudostandard word:

1. If the sequence Θ contains both E and R infinitely many times, then the normalized directive bi-sequence is of the form
$$(\widetilde{\Delta},\widetilde{\Theta}) = (v(a\overline{a})^\omega, \sigma(RE)^\omega)$$
for some $v \in \{0,1\}^*$, $\sigma \in \{E,R\}^*$, $|v| = |\sigma|$ and $a \in \{0,1\}$.

2. If the sequence Θ contains only one antimorphism ϑ infinitely many times, then the normalized directive bi-sequence is also of the form
$$(\widetilde{\Delta},\widetilde{\Theta}) = (va^\omega, \sigma\vartheta^\omega)$$
for some $v \in \{0,1\}^*$, $\sigma \in \{E,R\}^*$, $|v| = |\sigma|$ and $a \in \{0,1\}$.

3.3 Example. Let us show an example of a periodic binary generalized pseudostandard word. Assume $\Lambda = ((011)^\omega, (EER)^\omega)$. The condition (3.2) is met since E is always followed by 1 and R by 0. Let us write down the first few prefixes w_n:

$$\begin{aligned} w_1 &= 01 \\ w_2 &= 011001 \\ w_3 &= 01100110 \\ w_4 &= 0110011001. \end{aligned}$$

It can be easily verified by the reader that $\mathbf{u}((011)^\omega, (EER)^\omega) = (0110)^\omega$.

For ternary generalized pseudostandard words, straightforward analogy of (3.2) does not work.

3.4 Example. Consider the ternary infinite word $\mathbf{u} = \mathbf{u}((01)^\omega, (RE_1)^\omega)$. It is easy to show that any prefix p of $\mathbf{u}$ is left special – both $1p$ and $2p$ are factors of $\mathbf{u}$, thus $\mathbf{u}$ is an aperiodic word.

The condition for periodicity gets more complicated.

3.5 Theorem. *Let $\mathbf{u} = \mathbf{u}(\Delta, \Theta)$ be a ternary generalized pseudostandard word over $\{0,1,2\}$. Then $\mathbf{u}$ is periodic if and only if one of the following conditions is met:*

1. *The sequences Δ and Θ are eventually constant, i.e., $\Delta = va^\omega$ for some $v \in \{0,1,2\}^*$ and $a \in \{0,1,2\}$ and $\Theta = \sigma\vartheta^\omega$ for some $\sigma \in \{E_0, E_1, E_2, R\}^*$ and $\vartheta \in \{E_0, E_1, E_2, R\}$, $|v| = |\sigma|$.*

2. ▻ *Θ contains exactly two antimorphisms $\vartheta \in \{E_0, E_1, E_2\}$ and R infinitely many times;*

   ▻ *Δ contains two (not necessarily distinct) letters a and b infinitely many times such that $\vartheta(a) = b$;*

   ▻ *there exists $n_0 \in \mathbb{N}$ such that for every $n > n_0$ we have either*

   $$\vartheta_n = \vartheta \Rightarrow \delta_{n+1} = a \ \wedge \ \vartheta_n = R \Rightarrow \delta_{n+1} = b,$$

   or

   $$\vartheta_n = \vartheta \Rightarrow \delta_{n+1} = b \ \wedge \ \vartheta_n = R \Rightarrow \delta_{n+1} = a.$$

3. *The normalized directive bi-sequence $(\widetilde{\Delta}, \widetilde{\Theta})$ of $\mathbf{u}$ satisfies:*

   $$(\widetilde{\Delta}, \widetilde{\Theta}) = (v(ijk)^\omega, \sigma(E_kE_jE_i)^\omega),$$

   where $v \in \{0,1,2\}^*$, $\sigma \in \{E_0, E_1, E_2, R\}^*$, $|v| = |\sigma|$, and $i, j, k \in \{0,1,2\}$ are mutually different letters.

3.6 Example. Consider $\Lambda = (0(211)^\omega, (RE_0E_0)^\omega)$. Since $E_0(1) = 2$, the second condition of Theorem 3.5 is satisfied. Let us write down the first few prefixes w_n of $\mathbf{u}$:

$$\begin{aligned} w_1 &= 0 \\ w_2 &= 0210 \\ w_3 &= 0210120210 \\ w_4 &= 0210120210120. \end{aligned}$$

It is left for the reader to show that $\mathbf{u} = (021012)^\omega$.

3.7 Example. Consider $\Lambda = ((102)^\omega, (E_2E_0E_1)^\omega)$. The third condition of Theorem 3.5 is satisfied. Let us write down the first few prefixes w_n of $\mathbf{u}$:

$$\begin{aligned} w_1 &= 10 \\ w_2 &= 1002 \\ w_3 &= 100221 \\ w_4 &= 10022110 \\ w_5 &= 1002211002. \end{aligned} \tag{3.8}$$

It is not difficult to see that $\mathbf{u} = (100221)^\omega$.

We omit the complete proof of Theorem 3.5 here for it is long and technical (we will publish it in a short time on arXiv under the same title: Periodicity of Generalized Pseudostandard Words). However, we will at least provide some basic ideas. Let us underline that in the proof of the binary case, we made use of the fact that a normalization algorithm was provided in Theorem 2.2. Over ternary alphabet, it is still clear that every directive bi-sequence may be normalized, however the algorithm for normalization similar to the one over binary alphabet (Theorem 2.2) has not been found yet. Consequently, we provide in the sequel some partial results on normalization over ternary alphabet that are interesting themselves and that are moreover essential in the proof of Theorem 3.5.

3.1 Normalization of directive bi-sequences over ternary alphabet

In order to prove a sufficient and necessary condition for periodicity of ternary generalized pseudostandard words, even some partial results on normalization over ternary alphabet suffice. These results are needed in both elimination of some aperiodic cases and determining the period of periodic cases listed in Theorem 3.5. Let us introduce these partial results.

3.1.1 Directive bi-sequences with antimorphisms R and E_i

Let us start with bi-sequences that contain infinitely many times exactly two distinct antimorphisms including R.

3.9 Lemma. *Let the directive bi-sequence (Δ, Θ) of a ternary generalized pseudostandard word* $\mathbf{u}$ *contain as its factor $(abc, \vartheta RR)$, resp., $(abc, R\vartheta\vartheta)$, where $\vartheta \in \{E_0, E_1, E_2\}$ and $a, b, c \in \{0, 1, 2\}$ satisfy $\vartheta(b) = c$. Denote $w_n = \vartheta(w_n)$, $w_{n+1} = R(w_{n+1})$ and $w_{n+2} = R(w_{n+2})$, resp., $w_n = R(w_n)$, $w_{n+1} = \vartheta(w_{n+1})$ and $w_{n+2} = \vartheta(w_{n+2})$ the corresponding pseudopalindromic prefixes of* $\mathbf{u}$*. Then between w_{n+1} and w_{n+2} there is a ϑ-palindromic, resp., R-palindromic prefix w of* $\mathbf{u}$ *followed by the letter b.*

3.10 Corollary. *Under the assumptions of Lemma 3.9 we have:*
If the factor $(abc, \vartheta RR)$, resp., $(abc, R\vartheta\vartheta)$ of the directive bi-sequence (Δ, Θ) of the word $\mathbf{u}$ *is replaced with the factor $(abcb, \vartheta R\vartheta R)$, resp., $(abcb, R\vartheta R\vartheta)$, the same generalized pseudostandard word is obtained.*

3.11 Example. Let us illustrate Lemma 3.9 and Corollary 3.10. Assume we have already constructed the prefix $w_k = 012$ of a generalized pseudostandard word. Suppose further that the factor $(120, E_1RR)$ of the directive bi-sequence follows. It is readily seen that the assumptions of Lemma 3.9 are met (in particular we have $E_1(2) = 0$). Let us write down the prefixes w_{k+1}, w_{k+2} and w_{k+3}.

$$\begin{aligned} w_{k+1} &= \quad 0121012, \\ w_{k+2} &= \quad 01210122101210, \\ w_{k+3} &= \quad 0121012210121001210122101210. \end{aligned}$$

It is evident that between the prefixes w_{k+2} and w_{k+3}, there is the E_1-palindrome

$$01210122101210012101 2,$$

followed by 2. Corollary 3.10 moreover states that the generalized pseudostandard word remains the same if we replace the factor $(120, E_1RR)$ of the directive bi-sequence with the factor $(1202, E_1RE_1R)$ – the reader can check it easily.

3.12 Corollary. *Let the directive bi-sequence $\Lambda = (\Delta, \Theta)$ of a ternary generalized pseudostandard word* $\mathbf{u}$ *satisfy: The sequence $\Theta = \vartheta_1\vartheta_2\ldots$ contains infinitely many times exactly two distinct antimorphisms ϑ and R. The sequence $\Delta = \delta_1\delta_2\ldots$ contains infinitely many times two (not necessarily distinct) letters a, b such that $\vartheta(a) = b$. Let further the bi-sequence Λ satisfy: There exists $n_0 \in \mathbb{N}$*

such that for all $n > n_0$ we have either

$$\vartheta_n = \vartheta \Rightarrow \delta_{n+1} = a \text{ and } \vartheta_n = R \Rightarrow \delta_{n+1} = b\,, \tag{3.13}$$

or

$$\vartheta_n = \vartheta \Rightarrow \delta_{n+1} = b \text{ and } \vartheta_n = R \Rightarrow \delta_{n+1} = a\,. \tag{3.14}$$

Then there exists a directive bi-sequence $\widetilde{\Lambda} = (v(ab)^\omega, \sigma(R\vartheta)^\omega)$, where $v \in \{0,1,2\}^*$, $\sigma \in \{E_0, E_1, E_2, R\}^*$ such that $\mathbf{u}(\Lambda) = \mathbf{u}(\widetilde{\Lambda})$.

At this moment, we know that if a directive bi-sequence satisfies the assumptions of Lemma 3.9, it is not normalized. The remaining question is whether the new bi-sequence whose existence is guaranteed by Corollary 3.12 is normalized (at least from a certain moment on). A partial answer to this question provides the following lemma.

3.15 Lemma. *Let the directive bi-sequence $\Lambda = (\delta_1\delta_2\ldots, \vartheta_1\vartheta_2\ldots)$ of a generalized pseudostandard word $\mathbf{u}$ be of the form $\Lambda = (v(ab)^\omega, \sigma(R\vartheta)^\omega)$, where $v \in \{0,1,2\}^*$, $\sigma \in \{E_0, E_1, E_2, R\}^*$, and $|v| = |\sigma|$, $\vartheta \in \{E_0, E_1, E_2\}$ and $a, b \in \{0,1,2\}$ such that $\vartheta(a) = b$. Then for all $n > n_0 = |v|$ the sequence $(w_n)_{n>n_0}$ contains all ϑ-, resp., R-palindromic prefixes of length larger than $|w_{n_0}|$ of the word $\mathbf{u}$ followed by the letter a, resp., b.*

3.1.2 Directive bi-sequences with antimorphisms E_i and E_j

Let us now treat directive bi-sequences that contain infinitely many times exactly two antimorphisms E_i, E_j with $i, j \in \{0,1,2\}$. An essential difference between such antimorphisms and those ones studied previously is that $RE_i = E_iR$ for all $i \in \{0,1,2\}$, but E_iE_j is distinct from E_jE_i for $i \neq j$.

3.16 Remark. For $i, j, k \in \{0,1,2\}$ mutually different, we have $E_iE_jE_k = E_j$.

3.17 Corollary. *Let for some $i \in \{0,1,2\}$ and $v \in \{0,1,2\}^*$ hold $v = E_i(v)$. Let further $j, k \in \{0,1,2\}$ be such that i, j, k are mutually different. It holds then that $E_j(v) = E_kE_j(v)$, i.e., $E_j(v)$ is an E_k-palindrome.*

3.18 Lemma. *Let the directive bi-sequence (Δ, Θ) of a ternary generalized pseudostandard word $\mathbf{u}$ contain the factor (kj, E_iE_j), where $i, j, k \in \{0,1,2\}$ and $i \neq j$. Denote $w_n = E_i(w_n)$, $w_{n+1} = E_j(w_{n+1})$ the corresponding pseudopalindromic prefixes of $\mathbf{u}$. Then no E_i-palindromic prefix of $\mathbf{u}$ followed by the letter j is skipped between the pseudopalindromic prefixes w_n and w_{n+1}.*

3.19 Lemma. *Let the directive bi-sequence* (Δ, Θ) *of a ternary generalized pseudostandard word* $\mathbf{u}$ *contain the factor* $(kij, E_jE_iE_j)$*, where* $i, j, k \in \{0, 1, 2\}$ *and* $i \neq j$*. Denote* $w_n = E_j(w_n)$*,* $w_{n+1} = E_i(w_{n+1})$ *and* $w_{n+2} = E_j(w_{n+2})$ *the corresponding pseudopalindromic prefixes of* $\mathbf{u}$*. Then either between the pseudopalindromes* w_n *and* w_{n+1}*, or between* w_{n+1} *and* w_{n+2} *there is an* E_l*-palindromic prefix* w *of* $\mathbf{u}$ *followed by* l*, where* l *is distinct from both* i *and* j*.*

3.20 Example. Let us illustrate Lemma 3.19. Assume we have already constructed the prefix $w_k = 012$ of a generalized pseudostandard word. Suppose further that the factor $(110, E_0E_1E_0)$ of the directive bi-sequence follows. It is readily seen that the assumptions of Lemma 3.19 are met. Let us write down the prefixes w_{k+1}, w_{k+2} and w_{k+3}.

$$\begin{aligned} w_{k+1} &= \quad 012120, \\ w_{k+2} &= \quad 0121201012, \\ w_{k+3} &= \quad 012120101202012120. \end{aligned}$$

It is evident that between the prefixes w_{k+2} and w_{k+3}, there is the E_2-palindrome 01212010120201 followed by 2.

4 Open problems

We have provided a necessary and sufficient condition for periodicity of generalized pseudostandard words over ternary alphabet. This condition concerns the directive bi-sequence and in one case the normalized directive bi-sequence of the corresponding generalized pseudostandard word. The problem is that we only know that the normalized form of every directive bi-sequence exists, but in contrast to binary alphabet, we have no algorithm for producing the normalized directive bi-sequence from a given directive bi-sequence over ternary alphabet. Therefore, it is desirable to find such a normalizing algorithm over ternary or even any alphabet. Section 3.1 may serve as a hint in such an effort.

Observing results for binary and ternary alphabet, we have the following conjecture for multiliteral alphabet.

4.1 Conjecture (Periodicity of generalized pseudostandard words)**.** *Let*

$\mathbf{u}(\Delta, \Theta)$ be a d-ary generalized pseudostandard word. Then $\mathbf{u}$ is periodic if and only if the following conditions are met:

1. *The normalized directive bi-sequence is of the form*

$$(\widetilde{\Delta}, \widetilde{\Theta}) = (v\delta_1\delta_2\delta_3 \ldots, \sigma\vartheta_1\vartheta_2\vartheta_3 \ldots),$$

   where $|v| = |\sigma|$ and $\vartheta_i(\delta_{i+1}) = \vartheta_j(\delta_{j+1})$ for all $i, j \in \mathbb{N}$.

2. *For all $i \in \mathbb{N}$, if w is a ϑ_i-palindrome, then $\vartheta_{i+1}(w)$ is a ϑ_{i+2}-palindrome.*

In order to explain that this conjecture is in correspondence with results over binary and ternary alphabet, let us write down the statements for periodicity over binary and ternary alphabet using the normalized directive bi-sequence. Considering remarks following Theorem 3.1, we have the next corollary.

4.2 Corollary. *Let $(\widetilde{\Delta}, \widetilde{\Theta})$ be the normalized directive bi-sequence of a binary generalized pseudostandard word $\mathbf{u} = \mathbf{u}(\widetilde{\Delta}, \widetilde{\Theta})$. Then $\mathbf{u}$ is periodic if and only if one of the following conditions is met:*

1. *$(\widetilde{\Delta}, \widetilde{\Theta}) = (va^\omega, \sigma\vartheta^\omega)$ for some $v \in \{0,1\}^*, \sigma \in \{E, R\}^*, |v| = |\sigma|, a \in \{0,1\}$.*

2. *$(\widetilde{\Delta}, \widetilde{\Theta}) = (v(a\overline{a})^\omega, \sigma(RE)^\omega)$ for some $v \in \{0,1\}^*, \sigma \in \{E, R\}^*, |v| = |\sigma|$, and $a \in \{0,1\}$.*

Using Theorem 3.5, Lemma 3.9 and Corollary 3.12, we get the following corollary.

4.3 Corollary. *Let $(\widetilde{\Delta}, \widetilde{\Theta})$ be the normalized directive bi-sequence of a ternary generalized pseudostandard word $\mathbf{u} = \mathbf{u}(\widetilde{\Delta}, \widetilde{\Theta})$. Then $\mathbf{u}$ is periodic if and only if one of the following conditions is met:*

1. *$(\widetilde{\Delta}, \widetilde{\Theta}) = (va^\omega, \sigma\vartheta^\omega)$ for some $v \in \{0,1,2\}^*, \sigma \in \{E_0, E_1, E_2, R\}^*, |v| = |\sigma|$, $\vartheta \in \{E_0, E_1, E_2, R\}$ and $a \in \{0,1,2\}$.*

2. *$(\widetilde{\Delta}, \widetilde{\Theta}) = (v(ab)^\omega, \sigma(RE_i)^\omega)$ for some $v \in \{0,1,2\}^*, \sigma \in \{E_0, E_1, E_2, R\}^*$, $|v| = |\sigma|$, $i \in \{0,1,2\}$ and $a, b \in \{0,1,2\}$.*

3. *$(\widetilde{\Delta}, \widetilde{\Theta}) = (v(ijk)^\omega, \sigma(E_kE_jE_i)^\omega)$, where $v \in \{0,1,2\}^*, \sigma \in \{E_0, E_1, E_2, R\}^*$, $|v| = |\sigma|$ and $i, j, k \in \{0,1,2\}$ are mutually different letters.*

Acknowledgements

This work was supported by the Czech Science Foundation grant GAČR 13-03538S and by the grant of the Grant Agency of the Czech Technical University in Prague SGS14/205/OHK4/3T/14.

References

[1] Balková, L., Florian, J.: On Periodicity and Complexity of Generalized Pseudostandard Words, arXiv 1408.5210 [math.CO] (2014)

[2] Blondin Massé, A., Paquin, G., Tremblay, H., Vuillon, L.: On Generalized Pseudostandard Words over Binary Alphabet, Journal of Int. Sequences 16, Article 13.2.11 (2013)

[3] Bucci, M., De Luca, A., de Luca, A., Zamboni, L.: On Some Problems Related to Palindrome Closure, RAIRO – Theoretical Informatics and Applications 42, 679–700 (2008)

[4] De Luca, A., de Luca, A.: Pseudopalindrome Closure Operators in Free Monoids, Theoret. Comput. Sci. 362, 282–300 (2006)

[5] Droubay, X., Justin, J., Pirillo, G.: Episturmian Words and Some Constructions of de Luca and Rauzy, Theoret. Comput. Sci. 255, 539–553 (2001)

[6] Jajcayová, T., Pelatová, E., Starosta, Š.: Palindromic Closures Using Multiple Antimorphisms, Theoret. Comput. Sci. 533, 37–45 (2014)

[7] de Luca, A.: Sturmian Words: Structure, Combinatorics, and their Arithmetics, Theoret. Comput. Sci. 183, 45–82 (1997)

[8] Morse, M., Hedlund, G. A.: Symbolic Dynamics II - Sturmian Trajectories, Amer. J. Math. 62, 1–42 (1940)

De Bruijn Entropy and String Similarity

Steve Huntsman[1], Arman Rezaee[2]

[1] BAE Systems, 4301 North Fairfax Drive, Arlington, Virginia 22203, USA
[2]MIT EECS, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA
steve.huntsman@baesystems.com, armanr@mit.edu

Abstract

We introduce the notion of *de Bruijn entropy* of an Eulerian quiver and show how the corresponding relative entropy can be applied to practical string similarity problems. This approach explicitly links the combinatorial and information-theoretical properties of words and its performance is superior to edit distances in many respects and competitive in most others. The computational complexity of our current implementation is parametrically tunable between linear and cubic, and we outline how an optimized linear algebra subroutine can reduce the cubic complexity to approximately linear. A realistic application to molecular phylogenetics is provided.

1 Introduction

String similarity is a fundamental problem touching on computer science, bioinformatics, machine learning, and many other areas [19]. Most fast approaches to string similarity (e.g., bag-of-words or string kernel methods) are heuristic, whereas most theoretically grounded approaches to string similarity (e.g., Kolmogorov complexity methods) are slow. In this paper, we discuss a technique that bridges the gap, offering performance that can be tuned between linear and (in a sufficiently optimized implementation) subquadratic time while offering a clear interpretation in terms of combinatorial and information-theoretical primitives. Our technique is particularly well suited for comparing words based on their local structure and is agnostic to global structure, which is particularly interesting for comparing words encoding paths through digraphs with cycles (e.g., control flow graphs of computer programs) or for streaming data.

The paper is structured as follows. We begin by establishing notation and graph constructions in Section 2 before discussing basic combinatorial properties in Section 3 and our ultimate information-theoretical considerations in Section 4. Finally, we outline an application to molecular phylogenetics as an example where an approximate "ground truth" furnishes a basis for evaluating the performance of our approach and its comparison with conventional techniques.

2 Preliminaries

We begin with some preliminaries to establish basic definitions and notation. Let $n < \infty$ and consider a finite set $\mathcal{A} := \{a_1, \ldots, a_n\}$ which we call an *alphabet*. A *word* or *string* over $\mathcal{A}$ of length ℓ is an element of $\mathcal{A}^\ell$; a *symbol* is a word of length 1. The word $w = (w_1, \ldots, w_\ell)$ will typically be written as $w = w_1 \ldots w_\ell$. With a slight abuse of notation, we write $\ell(w) = \ell$. The *concatenation* of two words $w = w_1 \ldots w_\ell$ and $w' = w'_1 \ldots w'_{\ell'}$ is $ww' := w_1 \ldots w_\ell w'_1 \ldots w'_{\ell'}$.

A *cyclic word* or *necklace* [23] of length ℓ is the set of cyclic shifts of a word. We shall engage in a minor abuse of notation by letting w denote either a word or a cyclic word depending on context. If w is cyclic,

$w_j := w_{((j-1) \bmod \ell)+1}$.

Recall that a *quiver* (also known as a *multidigraph*, *directed multigraph*, etc.) Q is an ordered pair $(V(Q), E(Q)) \equiv (V, E)$ s.t. E is a multiset over $V \times V$ [4]. The *adjacency matrix* $A(Q)$ of Q is defined so that if there are a edges from v_j to v_k, then $A(Q)_{jk} := a$. It is clear that a quiver may be reconstructed from its adjacency matrix and *vice versa*, so that we may write $f(Q) \equiv f(A)$ for a generic function f without any ambiguity so long as either side is defined. Furthermore, we may make the implicit identifications $v_j \equiv j$ and $Q \equiv A(Q)$ for convenience.

For w cyclic and $k < \ell(w)$, the *order k de Bruijn quiver* [1] $Q_k(w)$ is given by

- $V(Q_k(w)) := \mathcal{A}^k$;
- $E(Q_k(w)) := \{(w_{1+j} \dots w_{k+j}, w_{2+j} \dots w_{k+1+j}) : 0 \leqslant j < \ell\}$.

That is, the edges of $Q_k(w)$ correspond to the subwords of length $(k+1)$ (a/k/a $(k+1)$-*grams*) of w, with multiplicities counted. Figure 1 shows an example.

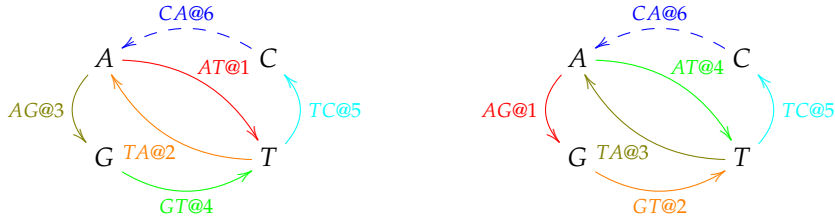


Figure 1. The cyclic words *ATAGTC* (L) and *AGTATC* (R) have identical order 1 de Bruijn quivers and (equivalently) the same 2-grams over the alphabet $\{A, C, G, T\}$. The quiver edges are annotated with $w_j w_{j+1}@j$ and colored according to j. Removing the dashed edges yields quivers for non-cyclic words in the obvious way.

[1] NB. Similar constructions (usually digraphs rather than quivers) appear throughout the literature: see e.g., [16] for a recent example.

Remarks.

- ▷ The order 0 de Bruijn quiver of a word w has one vertex corresponding to the empty word and edges corresponding to each symbol in w, with multiplicity.
- ▷ If w is a n-ary de Bruijn sequence of length n^k, then $Q_{k-1}(w)$ is the n-ary de Bruijn graph with n^k edges.
- ▷ $Q_k(w)$ is *Eulerian* (i.e., [strongly] connected and with indegrees equal to outdegrees, so that we may unambiguously write $\deg(v)$ for either quantity at vertex v) iff w contains every possible k-gram (otherwise, there are isolated vertices, but we may elide this technicality without comment at times). An Euler circuit on $Q_k(w)$ corresponds to a Hamiltonian path on $Q_{k+1}(w)$. These properties are why we deal with cyclic words.

3 Combinatorics

This section is an application of the so-called *transfer matrix* method [23].

Write $w \sim_k w'$ iff $Q_k(w) = Q_k(w')$. It is clear that $\sim_k$ is an equivalence relation; denote the corresponding equivalence class of w by $[w]_k$. Let $W_k(w) := |[w]_k|$ denote the number of cyclic words with the same order k de Bruijn quiver as w. In order to compute $W_k(w)$ it is convenient to consider the adjacency matrix $A_k(w) \equiv A(Q_k(w))$.

If A is a square matrix, write $d(A)$ for the vector with components given by the diagonal entries of A; similarly, write $d(x)$ for the diagonal matrix with diagonal entries given by the components of x. Note that since $d(d(x)) = x$ this is hardly an abuse of notation. If now 1 denotes a vector of ones, then $L(A) := d(A1) - A$ is the *Laplacian* of A. We recall two classical theorems:

Matrix-tree theorem. Let Q be a quiver. The diagonal cofactors of $L(A(Q))$ are all equal to each other and to the number $t(Q)$ of directed spanning trees of Q oriented towards any fixed vertex. □

BEST THEOREM. The number of Euler circuits of an Eulerian quiver Q is

$$c(Q) = t(Q) \cdot \prod_{v \in V(Q)} (\deg(v) - 1)!. \ \square \tag{3.1}$$

These readily yield the following

COROLLARY. [10, 8] Let $A := A_k(w)$ correspond to an Eulerian de Bruijn quiver. Then

$$W_k(w) = W(A) := \sum_{d \mid \gcd(A)} \frac{\phi(d) \cdot c(A/d)}{d \cdot (A/d)!}. \ \square \tag{3.2}$$

Here $\phi(\cdot)$ is the totient function, the gcd is defined elementwise and $M! := \prod_{i,j} M_{ij}!$. The $d = 1$ term dominates, giving the simple and effective approximation $W(A) \approx c(A)/A!$. We note that if Q is an Eulerian (not necessarily de Bruijn) quiver with adjacency matrix A, then we may still write $W(A)$ or $W(Q)$ for the RHS of (3.2). However, it is not necessary to directly interpret W in this more abstract context, since any finite Eulerian quiver can be embedded in some de Bruijn quiver.

SKETCH OF PROOF. The formula is obvious when $\gcd(A) = 1$, as in this case every cyclic word in $[w]_k$ corresponds to $A!$ Euler circuits. More generally, for $d|m| \gcd(A)$, the term $\frac{\phi(d) \cdot c(A/d)}{d \cdot (A/d)!}$ counts the cyclic words in $[w]_k$ of period ℓ/m with multiplicity $\frac{\phi(d)}{d} \cdot \frac{1}{m/d} = \frac{\phi(d)}{m}$. Since $\frac{1}{m} \sum_{d|m} \phi(d) = 1$, the result follows. $\square$

3.1 Example

Consider $\mathcal{A} = \{0, 1\}$, $k = 1$, and fix ℓ. If $g \in \{00, 01, 10, 11\}$, let $x_g(w)$ be the number of times that g occurs in w. Because w is cyclic, we must have $x_{01} = x_{10} = (\ell - x_{00} - x_{11})/2 =: x_*$, and $A_1(w) = \left(\begin{smallmatrix} x_{00} & x_* \\ x_* & x_{11} \end{smallmatrix}\right)$. [2] We have that $L(A_1(w)) = x_* \left(\begin{smallmatrix} 1 & -1 \\ -1 & 1 \end{smallmatrix}\right)$, so $t(Q_1(w)) = x_*$. Furthermore, $\deg(0) = x_{00} + x_*$ and $\deg(1) = x_* + x_{11}$, so $c(Q_1(w)) = x_* \cdot (x_{00} +$

[2] The degenerate case $x_* = 0$ corresponds to the words 0^ℓ and 1^ℓ, and must be treated separately.

$x_* - 1)! \cdot (x_* + x_{11} - 1)!$. It follows after a line or two of algebra that $W_1(w) \equiv W_1(x_{00}, x_*; \ell)$ equals

$$\frac{x_*}{(x_{00} + x_*)(x_* + x_{11})} \cdot \sum_{d \mid \gcd(x_{00}, x_{11}, x_*)} \phi(d) \cdot \binom{(x_{00} + x_*)/d}{x_*/d} \binom{(x_* + x_{11})/d}{x_*/d}.$$

Explicitly, for $\ell = 16$, we have the following table of values (zeros omitted):

		x_{00}															
	W_1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	0	1																1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		
	2	7	12	17	20	23	24	25	24	23	20	17	12	7				
	3	22	55	90	120	140	147	140	120	90	55	22						
x_*	4	43	120	212	280	309	280	212	120	43								
	5	42	126	210	245	210	126	42										
	6	22	56	75	56	22												
	7	4	7	4														
	8	1																

Summing over the table shows that there are 4116 distinct cyclic binary words of length 16, which can be confirmed via the Cauchy-Frobenius lemma.

Figure 2 shows results in the same vein for $\ell = 256$. It is evident that W_1 behaves very much like a Gaussian, with the only significant qualitative difference resulting from the triangular domain. Similar results hold for more general contexts, and this fact might enable analytical estimates for $W(A)$. □

4 Information theory

4.1 De Bruijn entropy

Definition. The *order k de Bruijn entropy* of a cyclic word w is $H_k(w) := \log W_k(w)$.

As in Section 3, we may also consider the entropy of an Eulerian quiver Q or of its adjacency matrix A, written respectively $H(Q)$ and $H(A)$.

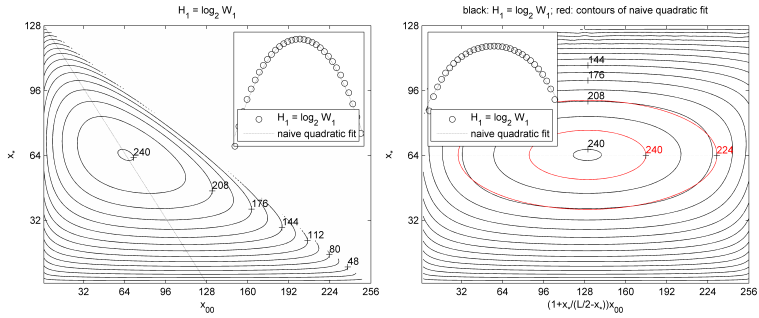


Figure 2. (L) Contour plot of $H_1 := \log_2 W_1$ for $\ell = 256$. Inset: plot of H_1 values intermittently sampled along the dotted line. Samples along horizontal lines behave similarly. (R) Black countours: plot of H_1 after the horizontal transformation $x_{00} \mapsto (1 + \frac{x_*}{\ell/2 - x_*}) \cdot x_{00}$. Red contours: naive quadratic fit. Inset: plot of transformed H_1 values intermittently sampled along the line $x_* = \ell/4$. Note that the naive quadratic fit is a slight overestimate at the peak.

Typically the logarithm will be taken with base $|\mathcal{A}|$ unless otherwise indicated.

This definition evokes Boltzmann's physical interpretation of entropy as the logarithm of the number of microscopic configurations of a system that are consistent with the system's macroscopic characterization. Here, the "macroscopic characterization" of w is just $Q_k(w)$, and "microscopic configurations" are just members of $[w]_k$. Another perspective realizes this definition as an analogue for finite words of the capacity of the discrete noiseless channel *à la* Shannon [20], or equivalently of the topological entropy of a subshift of finite type [11].

4.1.1 Compression arguments

Consider $\mathcal{A} = \{0, 1\}$ and $k = 1$ as in Section 3.1. In order to fully specify a cyclic binary word w of length ℓ, it suffices to specify both

- $A_1(w)$, which requires a total of $2\lceil \log_2 \ell \rceil - 1$ bits (because it requires $\lceil \log_2 \ell \rceil$ bits to specify x_{00} and $\lceil \log_2 \ell \rceil - 1$ bits to specify x_*);
- The appropriate element of $[w]_1$, which requires at most $\lceil H_1(w) \rceil \lessapprox$

$\ell - \log_2 \ell$ bits (because there are roughly $\ell^{-1}2^\ell$ cyclic binary words of length ℓ).

In particular, if $H_1(w) < \ell - 2\log_2 \ell + 1$, then we have the outline of a scheme for losslessly compressing w (the generalizations to both $k > 1$ and the nonbinary case $n > 2$ are not fundamentally different). Note that while most words are too statistically uniform (or more precisely, the adjacency matrices of their de Bruijn quivers have elements that are too similar) to be compressed in this way, in practice one is rarely interested in compressing statistically uniform data. Indeed, we recall that a standard diagonal argument shows that any fixed compression scheme will fail to compress most data [14].

The perspective of algorithmic information theory hinted at here will directly motivate the definition of relative de Bruijn entropy in Section 4.2.

4.1.2 Maximally informative values of k

Although the paper [21] leverages the empirical probability distribution of k-tuples rather than the more detailed notion of de Bruijn quivers, it nevertheless gives strong experimental evidence that the natural heuristic $k = \lfloor \log_n \ell \rfloor$ is a good approximation for the lower limit of a reasonably narrow range of maximally informative values of k in practice. While this paper also discusses upper limits on this range, these depend on the particular word and are of less practical interest for the obvious reason that increasing k requires more storage.

4.1.3 Remarks on computational complexity

A detailed analysis of the complexity of computing the de Bruijn entropy is likely to be both more intricate and less informative than an experimental one, owing to the complex relationship between the local statistical behavior of words and their corresponding quiver connectivity structure as a function of k (this is particularly true for the relative de Bruijn entropy, for which see below). However, we note that the dominant contribution to runtime is a matrix determinant (note that forming the adjacency matrices of quivers of words can be done in linear time), and we briefly discuss its complexity here.

Let ω denote the exponent for the complexity of matrix multiplication (say, 3 or perhaps 2.808 in practice, or 2.373 in theory [24]). Now $O((n^k)^\omega) = O(\ell) \iff k = \omega^{-1} \log_n \ell + O(1)$ determines k s.t. computing the de Bruijn entropy requires linear time with standard techniques of linear algebra (e.g., computing the determinant via LU or QR decomposition as in our current implementation). Meanwhile, as pointed out in Section 4.1.2, a reasonable rule of thumb for the maximally informative value of k is $\lfloor \log_n \ell \rfloor$.

These two observations can be combined by thinking of k as a scale below which where we have complete information about the structure of words and of $\ell^{1/\omega}$ as a scale above which negligible information suitable for comparisons between words is discernable in linear time using standard techniques of linear algebra. That is, the computation of a de Bruijn entropy can be *forced* to run in linear time by choosing $k = \lfloor \omega^{-1} \log_n \ell \rfloor$ (or, for that matter, $k = O(1)$), with the consequence that this amounts to neglecting not only correlations at scales greater than k (as usual), but also the ability to capture statistical fluctuations *of any sort* at scales beyond $\ell^{1/\omega}$. Insisting on $k = \lfloor \log_n \ell \rfloor$ means in practical terms that our technique requires cubic time in the implementation used here.

However, it is possible to do better, though for the sake of keeping this paper reasonably circumscribed we will confine ourselves here to a brief discussion. The reader will probably have noticed the phrase "standard techniques of linear algebra" repeated above, and considered the associated references to matrix decompositions for computing a determinant in (what is in practice) cubic time. In fact the determinant can be evaluated in less than $O((n^k)^\omega)$ time: it can be done in $O((n^k)^2)$ or even $O((n^k) \log^2(n^k))$ time using so-called *black box linear algebra* [25, 27]. The key here is that a diagonal minor $\hat{L}$ of the Laplacian has a predetermined sparse structure, so that the oracle $x \mapsto \hat{L}x$ can be realized in subquadratic time. This faciliates the computation of the characteristic polynomial of $\hat{L}$ using so-called *superfast* Toeplitz solvers in $O((n^k) \log^2(n^k))$ time [3, 1], from which the determinant follows trivially. [3]

[3] The notorious instability of superfast Toeplitz solvers for asymmetric matrices [6] is probably not a critical concern in the present context since the Laplacian matrix has integer entries; in any event, recently developed superfast solvers (see, e.g., [28]) have also addressed this problem.

Thus although our current implementation essentially has cubic time complexity for maximally informative values of k, it can already be regarded as having linear complexity for k independent of ℓ, and a sufficiently optimized linear algebra subroutine would yield complexity that is just the product of a linear and a polylogarithmic term in the general regime of interest, rendering it competitive with bag-of-words or kernel methods [19] that have linear complexity but weaker or more *ad hoc* theoretical justification.

4.2 Entropy of componentwise Eulerian quivers and relative de Bruijn entropy

Write

$$A \boxminus A' := (A - A') \vee 0 + (A' - A)^T \vee 0, \tag{4.1}$$

where the maxima are taken elementwise. It is easy to see that if A and A' both correspond to *componentwise Eulerian* quivers (i.e., the in- and outdegrees coincide, but some are zero, so that the quiver is not connected), then so does $A \boxminus A'$. Indeed, this is the adjacency matrix of the quiver that naturally corresponds to $A - A'$ after reversing edges with negative matrix entries. [4]

With this in mind, let $A^{(j)}$ be adjacency matrices respectively corresponding to Eulerian quivers $Q^{(j)}$, so that $Q := \bigcup_j Q^{(j)}$ is a componentwise Eulerian quiver with corresponding adjacency matrix A. Define

$$W(A) := \prod_j W(A^{(j)}) \tag{4.2}$$

and

$$H(A) := \log W(A) = \sum_j H(A^{(j)}). \tag{4.3}$$

To avoid degeneracies, we define $W(a) \equiv 1$ and $H(a) \equiv 0$, where here

[4] Similarly, define $A \boxplus A' := A \boxminus A'^T$. If A and A' both correspond to Eulerian (resp., componentwise Eulerian) quivers, then so does $A \boxplus A'$. The operations $\boxplus$ and $\boxminus$ therefore induce an abelian semigroup structure on the set of (possibly degenerate) Eulerian quivers and an abelian group structure on the set of (possibly degenerate) componentwise Eulerian quivers.

a is the 1×1 adjacency matrix corresponding to the quiver Γ_a with a single vertex and $a \geqslant 0$ edges (i.e., loops). This definition extends the prior one from Eulerian quivers to componentwise Eulerian quivers. Note that $H(A) = H(A^T)$ and $(A \boxminus A')^T = A' \boxminus A$, so that $H(A \boxminus A') = H(A' \boxminus A)$.

Suppose now that we have two cyclic words w and w' over $\mathcal{A}$. Given w and therefore also $A_k(w)$, all that is needed to determine $A_k(w')$ is the difference $A_k(w') - A_k(w)$, or equivalently the two nonnegative matrices

$$A_k(w|w') := [A_k(w) - A_k(w')] \vee 0; \tag{4.4}$$

$$A_k(w'|w) := [A_k(w') - A_k(w)] \vee 0. \tag{4.5}$$

In order to completely specify w' given w, it therefore suffices to specify $A_k(w|w')$, $A_k(w'|w)$, and a number of roughly $H_k(w')$ bits. It is clear that $H_{k+1}(w') \leqslant H_k(w')$. If w' is far from statistically uniform, then there will be some critical value $k << \ell(w')$ s.t. $H_k(w') = 0$ (note that $H_{\ell(w')-1}(w') \equiv 0$). At this point all the information in w' that is not latent in w is encoded in the matrices $A_k(w|w')$ and $A_k(w'|w)$. In other words, the conditional Kolmogorov complexity $K(w'|w)$ as well as the information distance [2, 13] can be approximated by a function of these matrices. [5]

This motivates the following

Definition. The *order k relative de Bruijn entropy* of w' given w is

$$H_k(w'||w) := H(A_k(w, w')),$$

where $A_k(w, w') := A_k(w) \boxminus A_k(w') = A_k(w|w') + A_k^T(w'|w)$. More generally, the relative entropy of A' given A is defined as

$$H(A'||A) := H(A' \boxminus A).$$

Note that (unlike the Kullback-Leiber incarnation of relative entropy for probability distributions) the relative entropy of componentwise Eulerian quivers is symmetric. Our experiments have shown that it is however not a

[5] We note in passing that the problem of comparing two words of vastly different length is sometimes of interest, and strategies such as those discussed in [21] may be appropriate in certain (but certainly not all) contexts.

pseudometric: i.e., it does not satisfy the triangle inequality. Nevertheless, it is straightforward to use the relative entropy to derive a pseudometric on a fixed set of words using the results of [5].

4.2.1 Example

For $1 < m \in \mathbb{N}$, let $w := 0^{m\ell}1^{m\ell}$ and $w' := (0^\ell 1^\ell)^m$. For $k < \ell$, a straightforward (if somewhat tedious) calculation shows that $H_k(w||w') \lesssim m \log mk$, whereas the (Levenshtein) edit distance between w and w' is $2\lfloor m/2 \rfloor \ell$. That is, for k and m fixed we have that $H_k(w||w') = O(1)$, whereas the corresponding edit distance is $O(\ell)$. □

5 Application to molecular phylogenetics

Molecular phylogenetics–i.e., the analysis of evolutionary relationships based on hereditary molecular characteristics–and biological classification of organisms typically focus on comparing DNA sequences [7]. A particularly convenient form of DNA for this purposes is mitochondrial DNA (mtDNA). mtDNA is an extremely economical repository of information in that nearly every base pair in human mtDNA is known to code for some protein or RNA product, and there is even overlap between coding regions; meanwhile mammalian mtDNA sequences are only on the order of 20k base pairs. Moreover, mtDNA is not highly conserved and mutates rapidly.

In figure 3 below we show that relative de Bruijn entropy produces results comparable if not superior to an edit distance (cf. figure 4) for constructing phylogenetic trees that easily capture most of the evolutionary relationships among primates (cf. figure 2 of [17]) from mtDNA sequences alone. Furthermore, the comparative performance of the relative entropy is likely to improve in other problem domains (e.g., dynamic analysis of computer programs) that can actually leverage the fact that the relative entropy captures local correlations while ignoring global correlations.

It is worth noting that the technique outlined here is *alignment-free* [7], and a suitable implementation optimized for speed (which our current implementation is not, cf. Section 4.1.3) is a promising candidate tool

for bioinformatics. In particular, it is an attractive alternative to current techniques such as those in [15, 26, 21, 22, 9] and the older but perhaps conceptually closer approach of [12]. We note also that de Bruijn quivers have been considered in the context of multiple alignment [18, 29, 30].

6 Acknowledgements

This material is based upon work supported by Defense Advanced Research Projects Agency (DARPA) and SPAWAR Systems Center Pacific (SSC Pacific) under Contract No. N66001-13-C-4047. The views, opinions, and/or findings contained in this article/presentation are those of the author(s)/presenter(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution Statement "A" (Approved for Public Release; Distribution Unlimited).

The authors are grateful to R. Ross for his helpful comments and assistance. Prior conversations between the first author and H. Fredricksen and D. H. Wood also substantively informed the discussion here.

References

[1] Ammar, G. S. and Gragg, W. B. "Superfast solution of real positive definite Toeplitz systems." *SIAM J. Matrix Anal. Appl.* **9**, 61 (1988).

[2] Bennett, C. H., Gács, P., Li, M., Vitányi, P. M. B., and Zurek, W. H. "Information distance." *IEEE Trans. Info. Th.* **44**, 1407 (1998).

[3] Brent, R. P., *et al.* "Fast solution of Toeplitz systems of equations and computation of Padé approximants." *J. Algorithms* **1**, 259 (1980).

[4] Bollobás, B. *Modern Graph Theory.* Springer (1998).

[5] Brickell, J., Dhillon, I. S., Sra, S., and Tropp, J. A. "The metric nearness problem." *SIAM J. Matrix Anal. Appl.* **30**, 375 (2008).

[6] Bunch, J. R. "Stability of methods for solving Toeplitz systems of equations." *SIAM J. Sci. Stat. Comp.* **6**, 349 (1985).

[7] Chan, C. X. and Ragan, M. A. "Next-generation phylogenomics." *Biology Direct* **8**, 1 (2013).

[8] Fredricksen, H. and Huntsman, S. "A string sampling protocol for de Bruijn sequence generation." *Proc. NSA Conf. SRC Computing* (2005).

[9] Hatje, K. and Kollmar, M. "A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method." *Frontiers in Plant Sci.* **3**, article 192 (2012).

[10] Jonsson, J. Personal communication (2002).

[11] Kitchens, B. P. *Symbolic Dynamics: One-sided, Two-sided, and Countable State Markov Shifts.* Springer (1998).

[12] Li, M., *et al.* "An information-based sequence distance and its application to whole mitochondrial genome phylogeny." *Bioinfo.* **17**, 149 (2001).

[13] Li, M., *et al.* "The similarity metric." *IEEE Trans. Info. Th.* **50**, 3250 (2004).

[14] Li, M. and Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications.* 3rd ed. Springer (2008).

[15] Li, Q., Xu, Z., and Hao, B. "Composition vector approach to whole-genome-based prokaryotic phylogeny: success and foundations." *J. Biotech.* **149**, 115 (2010).

[16] Moreno, E. "De Bruijn sequences and de Bruijn graphs for a general language." *Info. Proc. Lett.* **96**, 214 (2005).

[17] Perelman, P., *et al.* "A molecular phylogeny of living primates." *PLoS Genetics* **7**, e1001342 (2011).

[18] Raphael, B., Zhi, D., Tang, H., and Pevzner, P. "A novel method for multiple alignment of sequences with repeated and shuffled elements." *Genome Res.* **14**, 2336 (2004).

[19] Rieck, K. "Similarity measures for sequential data." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**, 296 (2011).

[20] Shannon, C. E. "A mathematical theory of communication." *Bell Sys. Tech. J.* **27**, 379 (1948).

[21] Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. "Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions." *Proc. Nat. Acad. Sci. USA* **106**, 2677 (2009).

[22] Sims, G. E. and Kim, S.-H. "Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs)." *Proc. Nat. Acad. Sci. USA* **108**, 8329 (2011).

[23] Stanley, R. P. *Enumerative Combinatorics, vols. 1-2.* Cambridge (1999).

[24] Vassilevska Williams, V. "Multiplying matrices faster than Coppersmith-Winograd." STOC (2012).

[25] von zur Gathen, J. and Gerhard, J. *Modern Computer Algebra*. Cambridge (1999).

[26] Wang, H., Xu, Z., Gao, L., and Hao, B. "A fungal phylogeny based on 82 complete genomes using the composition vector method." *BMC Evol. Bio.* **9**, 195 (2009).

[27] Wiedemann, D. H. "Solving sparse linear equations over finite fields." *IEEE Trans. Info. Th.* **32**, 54 (1986).

[28] Xia, J., Xi, Y., and Gu, M. "A superfast structured solver for Toeplitz linear systems via randomized sampling." *SIAM J. Matrix Anal. Appl.* **33**, 837 (2012).

[29] Zhang, Y. and Waterman, M. "An Eulerian path approach to global multiple alignment for DNA sequences." *J. Comp. Bio.* **10**, 803 (2003).

[30] Zhang, Y. and Waterman, M. "An Eulerian path approach to local multiple alignment for DNA sequences." *Proc. Nat. Acad. Sci* **102**, 1285(2005).

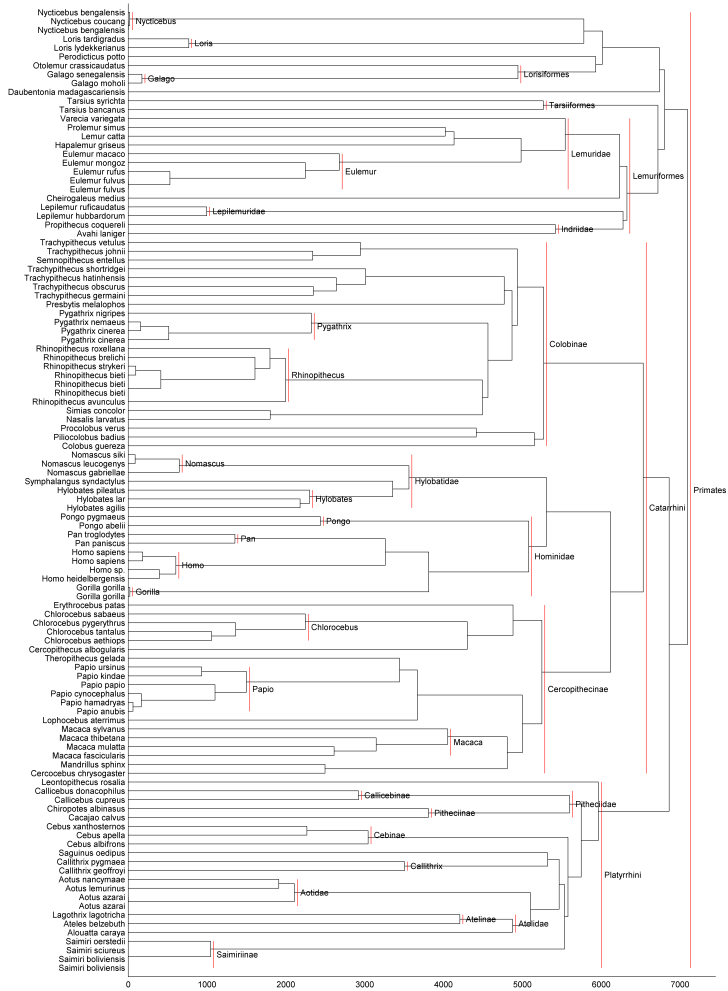


Figure 3. Automatically generated phylogenetic tree using average linkage for $k = 7$ relative de Bruijn entropy (unnormalized). Mismatches w/r/t the tree in figure 2 of [17] appear for the suborders *Haplorrhini* and *Strepsirrhini*, for the families *Cebidae* and *Lorisidae*, and for the subfamily *Callitrichinae*. Note that *Lorisiformes* here should be labeled as *Galagidae*, but this merely reflects an ambiguity in the input data annotation. Not explicitly shown, but also matched, are the tribes *Cercopithecini*, *Colobini*, *Papionini*, and *Presbytini*, the subfamilies *Homininae* and *Lorisnae*, the superfamily *Hominoidea*, and the infraorders *Lorisformes* (cf. previous comment) and *Simiformes*.

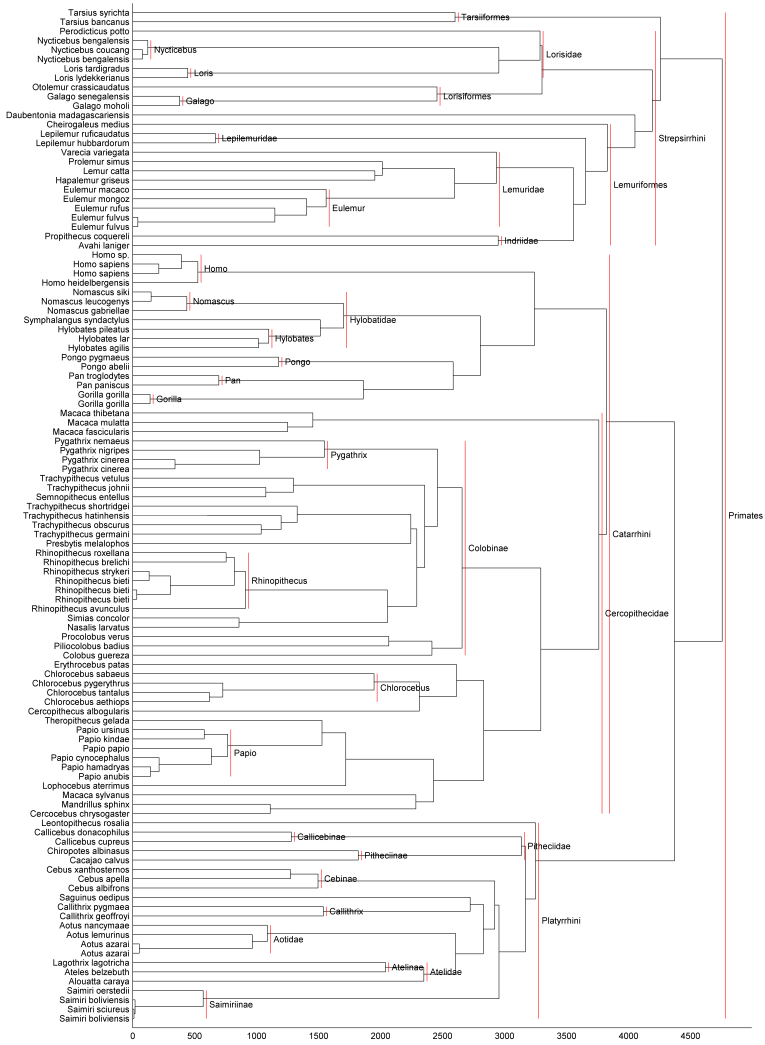


Figure 4. Average linkage for (Levenshtein) edit distance (unnormalized). Note that while the edit distance results match *Strepsirrhini* and *Lorisidae*, they fail to match the more fine-grained taxa *Macaca* and *Homindae*, the latter of which is particularly important to members of, e.g., *Homo sapiens*. (NB. *Cercopithecinae* and *Cercopithecidae* are ambiguously labeled here and in figure 3 due to the input data annotation and as such are not remarked on for comparative evaluation.)

Factor Complexity of Letter-to-Letter Images of Arnoux–Rauzy Words

Štěpán Starosta[1], Vojtěch Veselý[2]

[1] Faculty of Information Technology
Czech Technical University in Prague
Thákurova 9, 160 00 Praha 6, Czech Republic
[2]Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague
Trojanova 13, 120 00 Praha 2, Czech Republic

Abstract

We prove that if $\mathbf{u}$ is a k-ary Arnoux–Rauzy word and π a non-trivial letter-to-letter homomorphism, then the word $\pi(\mathbf{u})$ has its factor complexity equal to $(k-1)n+q$ for all sufficiently large n and some integer q.

1 Introduction

In combinatorics on words, the factor complexity $\mathcal{C}_{\mathbf{u}}(n)$ of an infinite word $\mathbf{u}$ is an important property. It can be seen as a measure of disorder as it counts the number of distinct factors of length n of $\mathbf{u}$. The famous Sturmian words can be characterized by their factor complexity: they are infinite aperiodic words with the least factor complexity possible $\mathcal{C}_{\mathbf{u}}(n) = n + 1$. This class is extensively studied, see for instance [14] for an overview.

The study of factor complexity is usually limited to some specific classes of words. An important class of infinite words is formed by purely morphic words, i.e., words that are fixed points of a morphism which is prolongable on the first letter of the infinite word. For a purely morphic word, due to [15], general asymptotic behaviour of its factor complexity is known. General techniques to determine its factor complexity are also known, see [6, 7], and for a specific class of morphisms, techniques described in [13] can be used to calculate factor complexity more efficiently.

A word is morphic if it is an image under a morphism of a purely morphic word. A theorem of Cobham [8] says that a word $\mathbf{u}$ is morphic if and only if it can be written as $\tau(\sigma^{\omega}(a))$ where $a \in \mathcal{A}$, σ is a non-erasing endomorphism of $\mathcal{A}^*$ and $\tau : \mathcal{A}^* \to \mathcal{B}^*$ is a letter-to-letter morphism. Thus, when studying morphic images of purely morphic words, we can restrict ourselves to letter-to-letter morphisms.

A further generalization of the notion of morphic word is an S-adic word. Let $(\mathcal{A}_i)$ be a sequence of alphabets and $\sigma_i : \mathcal{A}_{i+1}^* \to \mathcal{A}_i^*$ a sequence of morphisms. If the limit

$$\mathbf{u} = \lim_{i \to +\infty} \sigma_0 \sigma_1 \cdots \sigma_{i-1}(a_i),$$

where $a_i \in \mathcal{A}_i$, exists, then $\mathbf{u}$ *admits an S-adic representation* and its *S-adic expansion* is the sequence $((\sigma_i, a_i))_{i \in \mathbb{N}}$. A purely morphic word has a periodic S-adic expansion. For more general information about this concept, see [3, 7, 9] where one can find also results on factor complexity in general and for some specific cases. In [4], the authors investigate factor complexity of S-adic words generated by Arnoux–Rauzy–Poincaré algorithm.

In this paper, we enlarge the classes of words with known factor

complexity by non-trivial letter-to-letter projections of k-ary Arnoux–Rauzy words. Arnoux–Rauzy words were studied in [16] and later in [1]. Since one can associate with every Arnoux–Rauzy word a so-called standard word that admits an S-adic representation and has the same set of factors, the sets of factors of Arnoux–Rauzy words belong to the class that can be generated using S-adic rules. Thus, letter-to-letter projections of k-ary Arnoux–Rauzy words can also be perceived as S-adic words.

In Section 3, we prove the following theorem saying that the factor complexity of a non-trivial letter-to-letter image of an k-ary Arnoux-Rauzy word $\mathbf{u}$ reaches its maximal order of growth which is given by the factor complexity of $\mathbf{u}$, i.e., we prove that the complexity is $(k-1)n + \mathcal{O}(1)$:

1.1 Theorem. *Let $\mathbf{u}$ be an Arnoux–Rauzy word over $\mathcal{A}$ with $\#\mathcal{A} \geqslant 3$. Let π be a non-trivial letter-to-letter morphism of $\mathcal{A}$. There exist integers N and q such that*

$$\mathcal{C}_{\pi(\mathbf{u})}(n) = (\#\mathcal{A} - 1)\, n + q$$

for all n greater than N.

In [17], the first author shows that another property of Arnoux–Rauzy words is preserved when applying a specific morphism. Namely, it is shown that letter-to-letter projections of ternary Arnoux–Rauzy words are rich in palindromes, i.e., are saturated by palindromic factors to the highest possible level as Arnoux–Rauzy words are. For another result on morphic images of Arnoux–Rauzy words one may refer to [5].

2 Preliminaries

We recall needed notions from combinatorics on words and results on Arnoux–Rauzy words.

2.1 Combinatorics on words

An *alphabet* $\mathcal{A}$ is a finite set of symbols, called *letters*. A finite word $w = w_0 w_1 \cdots w_{n-1}$ is a finite sequence of letters $w_i \in \mathcal{A}$. The integer n is the length of w and is denoted by $|w|$. The unique word of length 0 is the *empty word* and is denoted by ε. The set of all finite words over $\mathcal{A}$ is denoted by

$\mathcal{A}^*$. An infinite word $\mathbf{u} = (u_i)_{i=0}^{+\infty}$ is an infinite sequence of letters $u_i \in \mathcal{A}$. The set of all infinite words over $\mathcal{A}$ is denoted by $\mathcal{A}^{\mathbb{N}}$. Let p, v, s and w be words such that $w = pvs$ with $p, v \in \mathcal{A}^*$ and $w, s \in \mathcal{A}^* \cup \mathcal{A}^{\mathbb{N}}$. The word v is said to be a *factor* of the word w, the word p is a *prefix* of w and s is its *suffix*. Given an infinite word, the set of all its factors is denoted by $\mathcal{L}(\mathbf{u})$. Thus, using this notation, the *factor complexity* of $\mathbf{u}$ can be expressed as

$$\mathcal{C}_{\mathbf{u}}(n) = \#\left(\mathcal{L}(\mathbf{u}) \cap \mathcal{A}^n\right).$$

Given a finite or infinite word $u = u_0u_1u_2\ldots$ with $u_i \in \mathcal{A}$, the letter having occurrence i in u (while indexing from 0) is denoted by $u[i]$, i.e., $u[i] = u_i$. The prefix of u of length n is denoted by $u[:n] = u_0u_1\ldots u_{n-1}$. If w is a factor of u and $i \in \mathbb{N}$ such that $w = u_i \cdots u_{i+|w|-1}$, then i is an *occurrence* of w in u. An infinite word $\mathbf{u}$ is *recurrent* if every its factor has infinitely many occurrences in $\mathbf{u}$.

If $w = w_0 \cdots w_{n-1}$ is a finite word, its *reversal* is the word

$$\overline{w} = w_{n-1} \cdots w_0,$$

i.e., the word w read from the last letter to the first. A word w is a *palindrome* if $w = \overline{w}$.

If v, w, z are finite words such that $vw = z$, we also write $w = v^{-1}z$ and $v = zw^{-1}$.

A mapping $\sigma : \mathcal{A}^* \to \mathcal{A}^*$ is a *morphism* if for all $w, v \in \mathcal{A}^*$ we have $\sigma(wv) = \sigma(w)\sigma(v)$. This allows us to naturally extend the domain of σ to infinite words. We say that σ is *prolongable* on $a \in \mathcal{A}$ if $\sigma(a) = au$ for some non-empty word u. As $\sigma^n(a)$ is a prefix of $\sigma^{n+1}(a)$ for all n, we can set $\mathbf{u} = \lim_{n\to+\infty} \sigma^n(a) = \sigma^\omega(a)$. Such a word $\mathbf{u}$ is a fixed point of σ and is said to be *purely morphic* (or purely substitutive). If $\mathbf{u}$ is a purely morphic word and $\tau : \mathcal{A}^* \to \mathcal{B}^*$ a morphism, then the word $\tau(\mathbf{u})$ is said to be *morphic*.

Let $\mathbf{u}$ be an infinite word and $w \in \mathcal{L}(\mathbf{u})$. We define the set of *right extensions* of w as $\mathrm{Rext}(w) = \{a \in \mathcal{A} : wa \in \mathcal{L}(\mathbf{u})\}$. If $\#\mathrm{Rext}(w) > 1$, then w is *right special*. The notion of *left special* is defined analogously. If w is both left special and right special, it is *bispecial*. Counting the number of right

extensions of all words of fixed length n leads to the following formula:

$$\mathcal{C}_{\mathbf{u}}(n+1) - \mathcal{C}_{\mathbf{u}}(n) = \Delta(\mathcal{C}_{\mathbf{u}}(n)) = \sum_{\substack{w \in \mathcal{L}(\mathbf{u}) \\ |w|=n}} (\#\mathrm{Rext}(w) - 1)\,. \tag{2.1}$$

Since if w is not right special, we have $\#\mathrm{Rext}(w) = 1$, the summand is non-zero only if w is right special.

2.2 Arnoux–Rauzy words

In this subsection we recall the definition and needed properties of Arnoux–Rauzy words.

Instead of giving the definition of an Arnoux–Rauzy word using S-adic expansion (one can refer to [1] for $k = 3$ and to [3, 12] for larger alphabets), we give a combinatorial definition. An infinite word $\mathbf{u} \in \mathcal{A}^{\mathbb{N}}$ is a k-ary Arnoux–Rauzy word if $\#\mathcal{A} = k$, the word $\mathbf{u}$ is recurrent, and for each n there is exactly one right special factor and one left special factor, both of length n and having exactly k right respectively left extensions in $\mathbf{u}$. In fact, Arnoux–Rauzy words belong to the family of so-called episturmian words as they are a generalization of Sturmian words. The reader may refer to [10, 12] or to [2], where other generalizations of Sturmian words are also investigated.

We list some properties of Arnoux–Rauzy words. Let $\mathbf{u}$ be a k-ary Arnoux–Rauzy word over $\mathcal{A}$. It follows from the definition that there are no two distinct bispecial factors of the same length. Thus, we may let $(w_i)_{i=0}^{+\infty}$ be the sequence of all bispecial factors of $\mathbf{u}$ ordered by their increasing length. This sequence has the following properties:

1. There exists a sequence $\Delta(\mathbf{u}) = (\delta_i)_{i=0}^{+\infty} \in \mathcal{A}^{\mathbb{N}}$ such that for each $i > 0$ we have
$$w_i = (w_{i-1}\delta_{i-1})^{(+)} \tag{2.2}$$
where $v^{(+)}$ is the shortest palindrome that has v as its prefix (the so-called *palindromic closure* of the finite word v). The sequence $\Delta(\mathbf{u})$ is the *directive sequence* of $\mathbf{u}$.

2. For all i, the word w_i is palindrome.

3. For all i, the word w_j is a suffix and a prefix of the word w_i for all $j < i$. In particular, every suffix of w_i is a right special factor.

4. Every letter of $\mathcal{A}$ occurs in $\Delta(\mathbf{u})$ infinitely many times.

The following lemma due to [11] states how one can evaluate the palindromic closure in (2.2) while constructing the sequence (w_i).

2.3 Lemma (Justin's formula)**.** *If the letter δ_i occurs in $\delta_0 \cdots \delta_{i-1}$, we set j such that $j < i$, $\delta_j = \delta_i$, and δ_i does not occur in $\delta_{j+1} \cdots \delta_{i-1}$.*

$$w_{i+1} = \begin{cases} w_i \delta_i w_i & \text{if } \delta_i \text{ does not occur in } \delta_0 \cdots \delta_{i-1}; \\ w_i w_j^{-1} w_i & \text{otherwise.} \end{cases}$$

2.4 Example. Let $\mathcal{A} = \{0, 1, 2\}$ and let Ψ be a morphism over $\mathcal{A}$ determined by

$$0 \mapsto 01, \quad 1 \mapsto 02, \text{ and } 2 \mapsto 0.$$

Let $\mathbf{u}_T$ be the fixed point of Ψ starting with 0:

$$\mathbf{u} = 0102010010201010201001020102010010201010201001020102010\ldots$$

The word $\mathbf{u}_T$ is probably the most famous Arnoux–Rauzy word — the Tribonacci word. The sequence of its bispecial factors starts as follows:

$$\begin{aligned} w_0 &= \varepsilon, \\ w_1 &= 0, \\ w_2 &= 010, \\ w_3 &= 0102010, \\ w_4 &= 01020100102010, \\ &\vdots \end{aligned}$$

It is known that the directive sequence of $\mathbf{u}_T$ is $\Delta(\mathbf{u}_T) = 012012012\ldots$.

3 Proofs

In this section, $\mathbf{u}$ stands for an Arnoux–Rauzy word over $\mathcal{A}$ with $\#\mathcal{A} \geqslant 3$, the sequence $(w_i)_{i=0}^{+\infty}$ the sequence of all bispecial factors of $\mathbf{u}$ ordered by

their increasing length and $\Delta(\mathbf{u}) = (\delta_i)_{i=0}^{+\infty}$ is the directive word of $\mathbf{u}$. The morphism π is a non-trivial letter-to-letter morphism from $\mathcal{A}$ to $\mathcal{B}$ with $\#\mathcal{B} > 1$.

3.1 Definition. Let $x \in \mathcal{A}$. We define $p_x : \mathbb{N} \to \mathbb{N}$ such that $(p_x(n))_{n=0}^{\infty}$ is a strictly increasing sequence of all indices i such that $\delta_i = x$. Furthermore, for all $i \in \mathbb{N}$ we define v_i to be the word such that

$$w_{i+1} = w_i v_i.$$

Let $x \in \mathcal{A}$. In accordance with Lemma 2.3, we have $v_{p_x(0)} = x w_{p_x(0)-1}$ (we set $w_{-1} = \varepsilon$). Let $i > 0$. We have

$$w_{p_x(i)} = w_{p_x(i)-1} v_{p_x(i)-1} = \ldots = w_{p_x(i-1)} v_{p_x(i-1)} v_{p_x(i-1)+1} \cdots v_{p_x(i)-1}.$$

This equality together with

$$w_{p_x(i)+1} = w_{p_x(i)} w_{p_x(i-1)}^{-1} w_{p_x(i)},$$

which follows from Lemma 2.3, leads to

$$v_{p_x(i)} = v_{p_x(i-1)} v_{p_x(i-1)+1} \cdots v_{p_x(i)-1}. \tag{3.2}$$

The last equation ensures that the following definition is correct.

3.3 Definition. Let $x \in \mathcal{A}$. Let $\mathbf{v}_x$ denote the infinite word given by

$$\mathbf{v}_x = \lim_{n \to \infty} v_{p_x(n)}.$$

3.4 Example (Example 2.4 continued). Since the directive word of the Tribonacci word $\mathbf{u}_T$ is purely periodic, we have

$$p_x(n) = 3n + x$$

for all $x \in \{0, 1, 2\}$. We have

$$\begin{aligned} \mathbf{v}_0 &= 0102\ldots, \\ \mathbf{v}_1 &= 1020\ldots, \\ \mathbf{v}_2 &= 2010\ldots. \end{aligned}$$

3.5 Lemma. *Let $x, y \in \mathcal{A}$ with $x \neq y$. There exists an integer N such that*

$$\pi(\mathbf{v}_x[N]) \neq \pi(\mathbf{v}_y[N]).$$

Proof. If $\pi(x) \neq \pi(y)$, we set $N = 0$. Suppose $\pi(x) = \pi(y)$.

Let i, j and k be integers such that

- $p_x(i) < k < p_x(i+1)$;
- $p_y(j) < k < p_y(j+1)$;
- $\pi(\delta_k) \neq \pi(x)$.

Since every letter occurs infinitely many times in $\Delta(\mathbf{u})$, such a choice is always possible.

Suppose without loss of generality that $p_x(i) < p_y(j)$. Using (3.2), we obtain

$$v_{p_y(j+1)} = v_{p_y(j)} \cdots v_k \cdots v_{p_y(j+1)-1}.$$

Set $s = v_{p_y(j)} \cdots v_{k-1}$, i.e., $v_{p_y(j+1)} = s v_k \cdots v_{p_y(j+1)-1}$.

Similarly, we obtain

$$v_{p_x(i+1)} = v_{p_x(i)} \cdots v_{p_x(i+1)-1} = v_{p_x(i)} \cdots v_{p_y(j)-1} s v_k \cdots v_{p_x(i+1)-1}. \quad (3.6)$$

As

$$w_k = w_{p_x(i)} v_{p_x(i)} v_{p_x(i)+1} \cdots v_{k-1} = w_{p_x(i)} v_{p_x(i)} \cdots v_{p_y(j)-1} s \quad (3.7)$$

we obtain

$$w_k = \overline{w_k} = \bar{s}\ \overline{v_{p_y(j)-1}} \cdots \overline{v_{p_x(i)}} w_{p_x(i)} = \bar{s} w_{p_x(i)} v_{p_x(i)} \cdots v_{p_y(j)-1} \quad (3.8)$$

since $w_{p_x(i)} v_{p_x(i)} \cdots v_{p_y(j)-1} = w_{p_y(j)}$ is a palindrome.

Let us now compare $v_{p_y(j+1)}[|s|]$ and $v_{p_x(i+1)}[|s|]$. It follows from the definition of s that $v_{p_y(j+1)}[|s|] = \delta_k$. Comparing (3.7) and (3.6), we obtain

$$v_{p_x(i+1)}[|s|] = w_k[|s| + |w_{p_x(i)}|] = x$$

where the last equality follows from (3.8).

Thus,

$$\pi(v_{p_y(j+1)}[|s|]) = \pi(\delta_k) \neq \pi(x) = \pi(v_{p_x(i+1)}[|s|])$$

and the proof is finished. □

An immediate corollary of the last lemma is that if we have two distinct letters x and y, then the longest common prefix of $\pi(\mathbf{v}_x)$ and $\pi(\mathbf{v}_y)$ is a finite word. This fact allows us to introduce the following notation.

3.9 Definition. Let $\mathcal{A}' \subset A$ with $\#\mathcal{A}' \geqslant 2$. We define the word $v_{\mathcal{A}'}$ to be the longest common prefix of all the words of $\{\pi(\mathbf{v}_x)\colon x \in \mathcal{A}'\}$. Let $n_{\mathcal{A}'}$ denote its length, i.e., $n_{\mathcal{A}'} = |v_{\mathcal{A}'}|$.

3.10 Example (Example 3.4 continued). Let $\mathcal{B} = \{a, b\}$ and $\xi : \mathcal{A}^* \to \mathcal{B}^*$ be determined by

$$\xi(x) = \begin{cases} a & \text{for } x \in \{0, 1\} \\ b & \text{for } x = 2. \end{cases}$$

We have

$$\begin{aligned} \xi(\mathbf{v}_0) &= aaa\ldots, \\ \xi(\mathbf{v}_1) &= aab\ldots, \\ \xi(\mathbf{v}_2) &= baa\ldots. \end{aligned}$$

Thus, $v_{\mathcal{A}} = v_{\{0,2\}} = v_{\{1,2\}} = \varepsilon$ and $v_{\{0,1\}} = aa$.

3.11 Lemma. *Let $\mathcal{A}' \subset \mathcal{A}$ with $\#\mathcal{A}' \geqslant 2$. If $u \in \mathcal{L}(\mathbf{u})$ is right special, then $\pi(u)v_{\mathcal{A}'} \in \mathcal{L}(\pi(\mathbf{u}))$ is right special.*

Proof. There exists an integer n such that the right special factor u is a suffix of w_i for all $i > n$. Thus, since for all $i > n$ the word $w_i v_i$ is a factor of $\mathbf{u}$, we conclude that up is a factor of $\mathbf{u}$ for all prefixes p of $\mathbf{v}_x$ for any $x \in \mathcal{A}'$. In particular, we obtain that $\pi(u)v_{\mathcal{A}'}$ is a factor of $\pi(\mathbf{u})$.

The definition of $v_{\mathcal{A}'}$ also implies that there exist letters $x, y \in \mathcal{A}'$ such that $\pi(\mathbf{v}_x[n_{\mathcal{A}'}]) \neq \pi(\mathbf{v}_y[n_{\mathcal{A}'}])$. It implies that $\pi(u)v_{\mathcal{A}'}$ is right special as $\pi(\mathbf{v}_x[n_{\mathcal{A}'}])$ and $\pi(\mathbf{v}_y[n_{\mathcal{A}'}])$ are its right extensions in $\pi(\mathbf{u})$. □

3.12 Lemma. *Let $\mathcal{A}_1$ and $\mathcal{A}_2$ be subsets of $\mathcal{A}$ of cardinality at least 2 such that $v_{\mathcal{A}_1} \neq v_{\mathcal{A}_2}$. There exist right special factors $u_1, u_2 \in \mathcal{L}(\mathbf{u})$ such that $|\pi(u_1)v_{\mathcal{A}_1}| = |\pi(u_2)v_{\mathcal{A}_2}|$ and $\pi(u_1)v_{\mathcal{A}_1} \neq \pi(u_2)v_{\mathcal{A}_2}$.*

Proof. If $n_{\mathcal{A}_1} = n_{\mathcal{A}_2}$, then we may set $u_1 = u_2 = \varepsilon$ and the claim follows from the assumption $v_{\mathcal{A}_1} \neq v_{\mathcal{A}_2}$.

Suppose without loss of generality that $n_{\mathcal{A}_2} > n_{\mathcal{A}_1}$ and let u be the right special factor of $\mathbf{u}$ of length $n_{\mathcal{A}_2} - n_{\mathcal{A}_1}$. Let $a = \pi(u[0])$ and find $x \in \mathcal{A}$ such that $\pi(x) \neq a$. As $\mathbf{u}$ is an Arnoux–Rauzy word, we have $ux \in \mathcal{L}(\mathbf{u})$. Since the directive word of $\mathbf{u}$ contains every letter infinitely many times,

the factor ux is a factor of w_j for some j. Thus, we may let u_2 be a right special factor of $\mathbf{u}$ such that ux is its prefix. Let u_1 be the right special factor of $\mathbf{u}$ such that $|u_1| + n_{\mathcal{A}_1} = |u_2| + n_{\mathcal{A}_2}$, i.e., $|\pi(u_1)v_{\mathcal{A}_1}| = |\pi(u_2)v_{\mathcal{A}_2}|$. Clearly, $|u_1| > |u_2|$ and thus u_2 is a suffix of u_1.

We have
$$\left(\pi(u_1)v_{\mathcal{A}_1}\right)[|u|] = \pi(u[0]) = a$$
and
$$\left(\pi(u_2)v_{\mathcal{A}_2}\right)[|u|] = \pi(x) \neq a$$
which implies $\pi(u_1)v_{\mathcal{A}_1} \neq \pi(u_2)v_{\mathcal{A}_2}$. □

3.13 Definition. Let $\mathcal{A}' \subset \mathcal{A}$ with $\#\mathcal{A}' \geqslant 2$. We define
$$\gamma(\mathcal{A}') = \#\{\pi(\mathbf{v}_x[n_{\mathcal{A}'}])\colon x \in \mathcal{A}'\} - 1.$$

The reason for the definition of the mapping γ is the following fact that is used later together with (2.1) to evaluate a bound on the factor complexity of $\pi(\mathbf{u})$: for a right special factor $u \in \mathcal{L}(\mathbf{u})$ we have
$$\#\mathrm{Rext}(\pi(u)v_{\mathcal{A}'}) - 1 \geqslant \gamma(\mathcal{A}'). \tag{3.14}$$

We will be interested in right special factors of the form $\pi(u)v_{\mathcal{A}'}$. As there may be distinct subsets $\mathcal{A}'$ and $\mathcal{A}''$ of $\mathcal{A}$ such that $v_{\mathcal{A}'} = v_{\mathcal{A}''}$, we use the following definition to evaluate correctly the number of right extensions of such right special factors of $\pi(\mathbf{u})$.

3.15 Definition. Let $\widetilde{\mathcal{A}}$ be the maximal subset of $2^{\mathcal{A}}$ such that

1. $\forall \mathcal{A}' \in \widetilde{\mathcal{A}}$, $\#\mathcal{A}' \geqslant 2$;

2. $\forall \mathcal{A}' \in \widetilde{\mathcal{A}}$, $\forall x \in \mathcal{A}\backslash\mathcal{A}'$, $v_{\mathcal{A}'\cup\{x\}} \neq v_{\mathcal{A}'}$.

3.16 Example (Example 3.10 continued). We have $\widetilde{\mathcal{A}} = \{\mathcal{A}, \{0,1\}\}$.

3.17 Lemma.
$$\sum_{\mathcal{A}'\in\widetilde{\mathcal{A}}} \gamma(\mathcal{A}') = \#\mathcal{A} - 1$$

Proof. Let us first construct a rooted tree T. The vertices of T are given by $\widetilde{\mathcal{A}}$. The root is $\mathcal{A}$. The set $\mathcal{A}_1 \in \widetilde{\mathcal{A}}$ is a descendant of $\mathcal{A}_2 \in \widetilde{\mathcal{A}}$ if and only if $\mathcal{A}_1 \subsetneq \mathcal{A}_2$.

We first give some properties of T. If $\mathcal{A}_1$ is a child of $\mathcal{A}_2$, then there is no set $\mathcal{A}_3 \in \widetilde{\mathcal{A}}$ such that $\mathcal{A}_1 \subsetneq \mathcal{A}_3 \subsetneq \mathcal{A}_2$. Also, $\mathcal{A}_1$ is a descendant of $\mathcal{A}_2$ if and only if $v_{\mathcal{A}_2}$ is a prefix of $v_{\mathcal{A}_1}$.

Let $\mathcal{A}' \in \widetilde{\mathcal{A}}$ and let $\mathcal{A}_1$ and $\mathcal{A}_2$ be its children with $\mathcal{A}_1 \neq \mathcal{A}_2$. The definition of $\widetilde{\mathcal{A}}$ and $\mathcal{A}_1 \neq \mathcal{A}_2$ imply that $v_{\mathcal{A}_1} \neq v_{\mathcal{A}_2}$. Suppose that there exists a letter $x \in \mathcal{A}_1 \cap \mathcal{A}_2$. It implies that both $v_{\mathcal{A}_1}$ and $v_{\mathcal{A}_2}$ are prefixes of $\pi(\mathbf{v}_x)$. Thus, $v_{\mathcal{A}_1}$ is a prefix of $v_{\mathcal{A}_2}$ or vice versa. Therefore either $\mathcal{A}_1$ or $\mathcal{A}_2$ is not a child of $\mathcal{A}'$ which is a contradiction. We conclude that $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$.

Let $\mathcal{A}'_0$ denote the following subset of $\mathcal{A}'$:

$$\mathcal{A}'_0 = \{x \in \mathcal{A}' \colon \forall y \in \mathcal{A}' \backslash \{x\}, \pi(\mathbf{v}_x[n_{\mathcal{A}'}]) \neq \pi(\mathbf{v}_y[n_{\mathcal{A}'}])\}.$$

In other words, the set $\mathcal{A}'_0$ consists of all letters of $\mathcal{A}'$ that are not elements of any of the children of $\mathcal{A}'$. Therefore, we can write

$$\mathcal{A}' = \mathcal{A}'_0 \quad \cup \bigcup_{\mathcal{A}'' \text{ is a child of } \mathcal{A}'} \mathcal{A}''$$

where all the sets are disjoint, i.e.,

$$\#\mathcal{A}_0 \quad + \sum_{\mathcal{A}'' \text{ is a child of } \mathcal{A}'} \#\mathcal{A}'' = \#\mathcal{A}'. \tag{3.18}$$

Let $h(\mathcal{A}')$ denote the height of the vertex $\mathcal{A}'$, i.e., the length of a longest path to a descendant leaf from $\mathcal{A}'$. Using the following notation

$$S(\mathcal{A}') = \sum_{\substack{\mathcal{A}'' \in \widetilde{\mathcal{A}} \\ \mathcal{A}'' \subset \mathcal{A}'}} \gamma(\mathcal{A}'')$$

we will show by induction on $h(\mathcal{A}')$ that $S(\mathcal{A}') = \#\mathcal{A}' - 1$.

Let $\mathcal{A}'$ be a vertex of height 0, i.e., a leaf of in T. We have

$$S(\mathcal{A}') = \gamma(\mathcal{A}').$$

The fact that $\mathcal{A}'$ is a leaf implies that for all distinct x and y of $\mathcal{A}'$ we have $\pi(\mathbf{v}_x[n_{\mathcal{A}'}]) \neq \pi(\mathbf{v}_y[n_{\mathcal{A}'}])$ as otherwise $\{x, y\}$ would be a subset of some descendant of $\mathcal{A}'$. Thus,

$$\#\{\pi(\mathbf{v}_x[n_{\mathcal{A}'}]) \colon x \in \mathcal{A}'\} = \#\mathcal{A}'$$

and the definition of $\gamma(\mathcal{A}')$ gives $\gamma(\mathcal{A}') = \#\mathcal{A}' - 1$.

Let $n \in \mathbb{N}$ and suppose the claim holds for all $i < n$. Let $h(\mathcal{A}') = n$. We obtain

$$S(\mathcal{A}') = \gamma(\mathcal{A}') \quad + \sum_{\mathcal{A}'' \text{ is a child of } \mathcal{A}'} S(\mathcal{A}'').$$

Since the height of a child of $\mathcal{A}'$ is strictly less than $h(\mathcal{A}')$, we use the induction hypothesis and obtain

$$S(\mathcal{A}') = \gamma(\mathcal{A}') \quad + \sum_{\mathcal{A}'' \text{ is a child of } \mathcal{A}'} \left(\#\mathcal{A}'' - 1\right).$$

It remains to evaluate $\gamma(\mathcal{A}')$. We obtain

$$\gamma(\mathcal{A}') = \#\{\pi(\mathbf{v}_x[n_{\mathcal{A}'}])\colon x \in \mathcal{A}'\} - 1 = \#\mathcal{A}'_0 + C - 1$$

where C is the number of children of $\mathcal{A}'$. Thus,

$$S(\mathcal{A}') = \#\mathcal{A}'_0 + C - 1 \quad + \sum_{\mathcal{A}'' \text{ is a child of } \mathcal{A}'} \left(\#\mathcal{A}''\right) - C.$$

Using (3.18) we conclude that

$$S(\mathcal{A}') = \#\mathcal{A}' - 1.$$

To finish the proof of the lemma, it suffices to evaluate $S(\mathcal{A})$. □

3.19 Lemma. *There exists an integer n_0 such that for all $n \geqslant n_0$ we have*

$$\sum_{\substack{w \in \mathcal{L}(\pi(\mathbf{u})) \\ w \text{ is right special} \\ |w| = n}} (\#\mathrm{Rext}(w) - 1) \geqslant \#A - 1.$$

Proof. As $\mathbf{u}$ is an Arnoux–Rauzy word, there is exactly one right special factor of length n for each $n \in \mathbb{N}$. Thus we may set $\mathbf{z}$ to be the left-infinite word such that all its suffixes are right special factors of $\mathbf{u}$. We associate with each $\mathcal{A}' \in \widetilde{\mathcal{A}}$ the left-infinite word $\zeta(\mathcal{A}') = \pi(\mathbf{z})v_{\mathcal{A}'}$. By Lemma 3.11, each suffix of $\zeta(\mathcal{A}')$ is right special in $\pi(\mathbf{u})$. Let $s_{n,\mathcal{A}'}$ be the suffix of length n of $\zeta(\mathcal{A}')$. Set

$$n_0 = \min\{n\colon \forall \mathcal{A}', \mathcal{A}'' \in \widetilde{\mathcal{A}}, \mathcal{A}' \neq \mathcal{A}'', s_{n,\mathcal{A}'} \neq s_{n,\mathcal{A}''}\}.$$

The existence of the integer n_0 is guaranteed by Lemma 3.12.

Since for all $n \geqslant n_0$, the factor $s_{n,\mathcal{A}'}$ is distinct from $s_{n,\mathcal{A}''}$ if $\mathcal{A}' \neq \mathcal{A}'' \in \widetilde{\mathcal{A}}$,

we conclude, using (3.14), that if $n \geqslant n_0$, then

$$\sum_{\substack{w \in \mathcal{L}(\pi(\mathbf{u})) \\ w \text{ is right special} \\ |w|=n}} (\#\mathrm{Rext}(w) - 1) \geqslant \sum_{\mathcal{A}' \in \widetilde{\mathcal{A}}} \big(\#\mathrm{Rext}(s_{n,\mathcal{A}'}) - 1\big) \geqslant \sum_{\mathcal{A}' \in \widetilde{\mathcal{A}}} \gamma(\mathcal{A}') = \#A - 1$$

where the last equality follows from Lemma 3.17. □

Proof of Theorem 1.1. Let us restate the claim of Lemma 3.19 using (2.1): there exists an integer n_0 such that for all $n \geqslant n_0$ we have

$$\Delta\mathcal{C}_{\pi(\mathbf{u})}(n) \geqslant \#\mathcal{A} - 1.$$

As π is a letter-to-letter morphism, we have

$$\mathcal{C}_{\pi(\mathbf{u})}(n) \leqslant \mathcal{C}_{\mathbf{u}}(n) = (\#\mathcal{A} - 1)n + 1$$

for all n.

Therefore,

$$\Delta\mathcal{C}_{\pi(\mathbf{u})}(n) > \#\mathcal{A} - 1$$

can happen only for finitely many n. This implies that there exists an integer N such that

$$\Delta\mathcal{C}_{\pi(\mathbf{u})}(n) = \#\mathcal{A} - 1$$

for all $n \geqslant N$, i.e.,

$$\mathcal{C}_{\pi(\mathbf{u})}(n) = (\#A - 1)\, n + q$$

for all $n \geqslant N$ for some integer q. □

3.20 Example (Example 3.16 continued). One can show that there are exactly two right special factors in $\xi(\mathbf{u}_T)$ of fixed length at least 4 and they are of the form

$$\xi(u_1)v_{\mathcal{A}} \quad \text{or} \quad \xi(u_2)v_{\{0,1\}}$$

where u_1 and u_2 are right special factors of $\mathbf{u}_T$. Examining shorter factors, one can verify that

$$\mathcal{C}_{\xi(\mathbf{u}_T)}(n) = \begin{cases} n+1 & \text{if } n < 4 \\ 2n-3 & \text{if } n \geqslant 4 \end{cases}.$$

Acknowledgements

The first author acknowledges financial support from the Czech Science Foundation grant 13-35273P. The work of the second author was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS14/205/OHK4/3T/14.

References

[1] P. Arnoux and G. Rauzy. Représentation géométrique de suites de complexité $2n + 1$. *Bull. Soc. Math. France*, 119:199–215, 1991.

[2] L. Balková, E. Pelantová, and Š. Starosta. Sturmian jungle (or garden?) on multiliteral alphabets. *RAIRO-Theoret. Inf. Appl.*, 44:443–470, 2010.

[3] V. Berthé and V. Delecroix. Beyond substitutive dynamical systems: S-adic expansions. *RIMS Lecture note 'Kokyuroku Bessatu'*, B46:81–123, 2014.

[4] V. Berthé and S. Labbé. Factor complexity of S-adic sequences generated by the Arnoux-Rauzy-Poincaré algorithm. *CoRR*, abs/1404.4189, 2014.

[5] M. Bucci and A. De Luca. On a family of morphic images of Arnoux-Rauzy words. In *LATA '09: Proceedings of the 3rd International Conference on Language and Automata Theory and Applications*, pages 259–266. Springer-Verlag, 2009. ISBN 978-3-642-00981-5.

[6] J. Cassaigne. Complexity and special factors. *Bull. Belg. Math. Soc. Simon Stevin 4*, 1:67–88, 1997.

[7] J. Cassaigne and F. Nicolas. Factor complexity. In V. Berthé and M. Rigo, editors, *Combinatorics, automata, and number theory.*, volume 135 of *Encyclopedia of Mathematics and its Applications*, pages 163–247. Cambridge University Press, 2010.

[8] A. Cobham. On the Hartmanis-Stearns problem for a class of TAG machines. In *9th Annual Symposium on Switching and Automata Theory, Schenectady, New York, USA, October 15-18, 1968*, pages 51–60, 1968.

[9] N. P. Fogg. *Substitutions in Arithmetics, Dynamics and Combinatorics*, volume 1794 of *Lecture notes in mathematics*. Springer, 1st edition, 2002. ISBN 9783540441410.

[10] A. Glen and J. Justin. Episturmian words: a survey. *Theoret. Inf. Appl.*, 43(3):403–442, 2009.

[11] J. Justin. Episturmian morphisms and a Galois theorem on continued fractions. *RAIRO-Theoret. Inf. Appl.*, 39:207–215, 2005.

[12] J. Justin and G. Pirillo. Episturmian words and episturmian morphisms. *Theoret. Comput. Sci.*, 276(1-2):281–313, 2002. ISSN 0304-3975.

[13] K. Klouda. Bispecial factors in circular non-pushy D0L languages. *Theoret. Comput. Sci.*, 445(0):63–74, 2012. ISSN 0304-3975. doi: 10.1016/j.tcs.2012.05.007.

[14] M. Lothaire. *Algebraic combinatorics on words*. Number 90 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2002.

[15] J.-J. Pansiot. Complexité des facteurs des mots infinis engendrés par morphismes itérés. In J. Paredaens, editor, *11th ICALP, Antwerpen*, volume 172 of *LNCS*, pages 380–389. Springer, Jul 1984.

[16] G. Rauzy. Suites à termes dans un alphabet fini. *Séminaire de Théorie des Nombres de Bordeaux*, Anné 1982–1983(exposé 25), 1983.

[17] Š. Starosta. Morphic images of episturmian words having finite palindromic defect. *Eur. J. Comb.*, 51:359–371, 2016.

Authors