Minimal de Bruijn Sequence
in a Language with Forbidden Substrings⋆

Eduardo Moreno[1,3] and Martín Matamala[1,2]

[1] Departamento de Ingeniería Matemática
Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile
Casilla 170-3, Correo 3, Santiago, Chile
`{emoreno,mmatamal}@dim.uchile.cl`
[2] Centro de Modelamiento Matemático, UMR 2071, UCHILE-CNRS
Casilla 170-3, Correo 3, Santiago, Chile
[3] Institut Gaspard Monge, Université de Marne-la-Vallée
Champs-sur-Marne, 77454 Marne-la-Vallée cedex 2, France

Abstract. Let be the following strategy to construct a walk in a labeled digraph: at each vertex, we follow the unvisited arc of minimum label. In this work we study for which languages, applying the previous strategy over the corresponding de Bruijn graph, we finish with an Eulerian cycle, in order to obtain the minimal de Bruijn sequence of the language.

1 Introduction

Given a language, a de Bruijn sequence of span n is a periodic sequence such that every n-tuple in the language (and no other n-tuple) occurs exactly once. Its first known description appears as a Sanskrit word *yamátárájabhánasalagám* which was a memory aid for Indian drummers, where the accented/unaccented syllables represent long/shorts beats, so all possible triplets of short and long beats are included in the word. De Bruijn sequences are also known as "shift register sequences" and was originally studied by N. G. De Bruijn for the binary alphabet [1]. These sequences have many different applications, such as memory wheels in computers and other technological device, network models, DNA algorithms, pseudo-random number generation, modern public-key cryptographic schemes, to mention a few (see [2],[3],[4]). Historically, de Bruijn sequence was studied in an arbitrary alphabet considering the language of all the n-tuples. There is a large number of de Bruijn sequence in this case, but only a few can be generated efficiently, see [5] for a survey about this subject. In 1978, Fredricksen and Maiorana [6] give an algorithm to generate a de Bruijn sequence of span n based in the Lyndon words of the language, which resulted to be the minimal one in the lexicographic order, and this algorithm was proved to be efficient [7]. Recently, the study of these concepts was extended to languages with forbidden

substrings: in [8] it was given efficient algorithms to generate all the words in a language with one forbidden substring, in [9] the concept of de Bruijn sequences was generalized to restricted languages with a finite set of forbidden substrings and it was proved the existence of these sequences and presented an algorithm to generate one of them, however, to find the minimal sequence is a non-trivial problem in this more general case. This problem is closely related to the "shortest common super-string problem" which is a important problem in the areas of DNA sequencing and data compression.

In this work we study the de Bruijn sequence of minimal lexicographical label. In section 2 we present some definitions and previous results on de Bruijn sequences and the BEST Theorem, necessary to understand the main problem, and we prove a result related with the BEST Theorem which will be useful in the following sections. In section 3 we study the main problem, giving some results on the structure of the de Bruijn graph. Finally, in section 4 we present some remarks and extensions of this work.

2   De Bruijn Sequence of Restricted Languages

2.1   Definitions

Let A be a finite set with a linear order $<$. A *word* on the alphabet A is a finite sequence of elements of A, whose length is denoted by $|w|$.

A word p is said to be a *factor* of a word w if there exist words $u, v \in A^*$ such that $w = upv$. If u is the empty word ε then p is called a *prefix* of w, and if v is empty then is called a *suffix* of w. If $p \neq w$ then p is a *proper factor*, *proper prefix* or *proper suffix*, respectively.

The set A^* of all the words on the alphabet A is linearly ordered by the alphabetic order induced by the order $<$ on A. By definition, $x < y$ either if x is a prefix of y or if $x = uav$, $y = ubw$ with $u, v, w \in A^*$, $a, b \in A$ and $a < b$. A basic property of the alphabetic order is the following: if $x < y$ and if x is not a prefix of y, then for any pair of words u, v, $xu < yv$.

Given an alphabet A, a full shift $A^{\mathbb{Z}}$ is the collection of all bi-infinite sequences of symbols from A. Let $\mathcal{F}$ be a set of words over A^*. A *subshift of finite type* (SFT) is the subset of sequences in $A^{\mathbb{Z}}$ which does not contain any factor in $\mathcal{F}$. We will refer to $\mathcal{F}$ as the set of *forbidden blocks* or *forbidden factors*.

Given a set $\mathcal{F}$ of forbidden blocks, in this work we will say that a word w is in the language if the periodical word w^∞, composed by infinite repetitions of w, is in the language of the SFT defined by $\mathcal{F}$. The set of all the words of length n in the language defined by $\mathcal{F}$ will be denoted by $\mathcal{W}^{\mathcal{F}}(n)$.

A SFT is *irreducible* if for every ordered pair of blocks u, v in the language there is a block w in the language so that uwv is a block of the language.

A de Bruijn sequence of span n in a restricted language is a circular string $B^{\mathcal{F},n}$ of length $\left|\mathcal{W}^{\mathcal{F}}(n)\right|$ such that all the words in the language of length n are factors of $B^{\mathcal{F},n}$. In other words,

$$\{(B^{\mathcal{F},n})_i \ldots (B^{\mathcal{F},n})_{i+n-1 \bmod n} | i = 0 \ldots n-1\} = \mathcal{W}^{\mathcal{F}}(n)$$

These concepts are studied in [9], extending the known results on subshifts of finite type to this context. In particular two results are relevant in this work, the first one is a bound in the number of words of length n in the language:

$$\left|\mathcal{W}^{\mathcal{F}}(n)\right| = \Theta\left(\lambda^n\right)$$

where $\log(\lambda)$ is the *entropy* of the system (see [10]). The second result proves the existence of a de Bruijn sequence:

Theorem 1. *For any set of forbidden substrings $\mathcal{F}$ defining an irreducible subshift of finite type, there exists a de Bruijn sequence of span n.*

This last theorem is a direct consequence of the fact that the de Bruijn graph of span n is an Eulerian graph. The *de Bruijn graph* of span n, denoted by $G^{\mathcal{F},n}$, is the largest strongly connected component of the directed graph with $|A|^n$ vertices, labeled by the words in A^n, and the set of arcs

$$E = \left\{(as, sb) | a, b \in A, s \in A^{n-1}, asb \in \mathcal{W}^{\mathcal{F}}(n+1)\right\}$$

where the label of the arc $e = (as, sb)$ is $l(e) = b$. Note that if the SFT is irreducible, this graph has only one strongly connected component of size greater than 1, so there is no ambiguity in the definition.

There are not two vertices with the same label, hence from now we identify a vertex by its label. If $W = e_1 \dots e_k$ is a walk over $G^{\mathcal{F},n}$, we denote the label of W by $l(W) = l(e_1) \dots l(e_k)$, and by $l(W)^j$ the concatenation of of j times $l(W)$.

There exists a bijection between the arcs of $G^{\mathcal{F},n}$ and the words in $\mathcal{W}^{\mathcal{F}}(n+1)$, because to each arc with label $a \in A$ with tail at $w' \in A^n$ we can associate the word $w'a$ which is, by definition, a word in $\mathcal{W}^{\mathcal{F}}(n+1)$. Equally if $w'a$ is a word of $\mathcal{W}^{\mathcal{F}}(n+1)$, with $a \in A$, then there exists a vertex w' and an arc with tail at this vertex with label a.

Furthermore, if a word w is a label of a walk from u to v then v is a suffix of length n of uw. In the same way, if $w \in \mathcal{W}^{\mathcal{F}}(n+1)$ then there is a cycle C in $G^{\mathcal{F},n}$ with label $l(C)$ such that $l(C)^{\frac{n+1}{|C|}} = w$.

With all these properties it is easy to see that a de Bruijn sequence of span $n+1$ is exactly the label of an Eulerian cycle over $G^{\mathcal{F},n}$.

2.2   The BEST Theorem

BEST is an acronym of N. G. de Bruijn, T. van Aardenne-Ehrenfest, C. A. B. Smith and W. T. Tutte, the BEST Theorem (see [11]) gives a correspondence between Eulerian cycles in a digraph and its rooted trees converging to the root vertex.

Let r be a vertex of an Eulerian digraph $G = (V, E)$, a spanning tree converging to the root r is a spanning tree such that there exists a directed path from each vertex to the root.

Given an Eulerian cycle starting at the root of an Eulerian digraph, if for every vertex of G we take the last arc with tail at this vertex in the cycle then
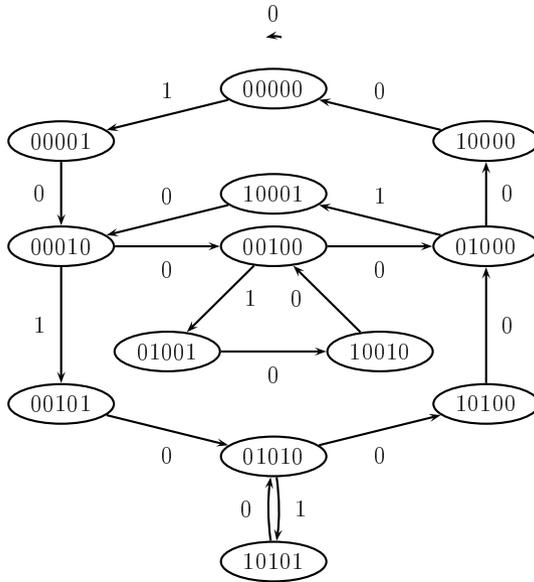
Fig. 1. De Bruijn digraph of span 5 for the Golden Mean ($\mathcal{F} = \{11\}$).

we obtain a spanning tree converging to the root. Conversely, given a spanning tree converging to the root, a walk over G starting at the root and using the arc in the tree only if all the arcs with tail at this vertex has been used, is an Eulerian cycle. A walk over the graph of this kind will be called a walk "avoiding the tree".

The BEST Theorem proves that for every different spanning tree we have a different Eulerian cycle. Therefore it also allows us to calculate the exact number of Eulerian cycles on a digraph, which is given by

$$C_{\mathcal{F}} = M_T \cdot \prod_{i=1}^{|V|} (d^+(v_i) - 1)!$$

where M_T is the number of rooted spanning trees converging to a given vertex. We bound the second term by $((\bar{d}^+ - 1)!)^{|V|}$ where $\bar{d}^+$ is the mean of the outgoing degrees over all the vertices, so we have a lower bound to the number of de Bruijn sequences

$$C_{\mathcal{F}} = \Omega\left(\lfloor \lambda - 1 \rfloor!^{\lambda^{n-1}}\right)$$

in particular, for a system with $\lambda \geq 3$ the number of the Bruijn sequences of span n is exponential in the number of words in the language of length $n - 1$. In the systems with $3 > \lambda > 1$ this bound is generally also true, because the underestimated term M_T is generally exponential, for example, in the system without restrictions of alphabet $\{0, 1\}$, this term is equal to $2^{2^{n-1}}$.

Now, we define formally a walk "avoiding a subgraph". Let r be any vertex. For each vertex $v \neq r$ in $G^{\mathcal{F},n}$ let e_v be any arc starting at v. Let H be the spanning subgraph of $G^{\mathcal{F},n}$ with arc set $\{e_v : v \in V(G^{\mathcal{F},n}) \setminus \{r\}\}$.

Is easy to see that H is composed by cycles, subtrees converging to a cycle, and one subtree converging to r. For a vertex not in a cycle of H, we define H_v as the directed subtree converging to v in H.

We define recursively a walk in $G^{\mathcal{F},n}$ which *avoid* H. It starts at the root vertex r. Let $v_0 e_0 \cdots v_i$ be the current walk. If there is an unvisited arc $e_i = (v_i, v_{i+1})$ not in H we extend the walk by $e_i v_{i+1}$. Otherwise we use the arc e_{v_i} in H.

We say that a walk over the graph *exhausts* a vertex if the walk use all the arc having the vertex as head or tail.

The next lemma studies in which order the vertices are exhausted in a walk avoiding H

Lemma 1. *Let W be a walk starting at vertex r avoiding H, let v be a vertex and let Wv the subpath of W starting at vertex r and finishing when it exhausts the vertex v. Then for each vertex u in H_v, u is exhausted in Wv.*

Proof. By induction in the depth of the subtree with root v. If v is a leaf of H then $H_v = \{v\}$. If v is not a leaf and Wv exhaust v, then Wv visit all arc $(v, w) \in E$, and therefore all the arcs $(u, v) \in E$, applying induction hypothesis to all vertices u such that $(u, v) \in E$ we prove the result. $\square$

3   Minimal de Bruijn Sequence

Let $m = m_1, \ldots m_n$ be the vertex of $G^{\mathcal{F},n}$ of maximum label in the lexicographic order. We are interested in to obtain the Eulerian cycle of minimum label starting at m. In order to obtain this cycle, we define the following walk: Starting at m, at each vertex we continue by the arc with the lowest label between the unvisited arcs with tail at this vertex. A walk constructed by this way will be called a *minimal walk*. By definition, there is no walk with a lexicographically lower label, except its subwalks. In this section we characterize when a minimal walk starting at m is an Eulerian cycle, obtaining the minimal de Bruijn sequence.

For each vertex v let $e(v)$ be the arc with tail at the vertex v and with maximum label. Let T be the spanning subgraph of $G^{\mathcal{F},n}$ composed by the set of arcs $e(v)$, for $v \in V(G^{\mathcal{F},n})$, $v \neq m$. The label of $e(v)$ will be denoted by $\gamma(v)$.

Is easy to see that a minimal walk is a walk avoiding T, hence we can study a minimal walk analyzing the structure of T.

Theorem 2. *A minimal walk is an Eulerian cycle if and only if T is a tree.*

Proof. A minimal walk W exhaust m, if T is a tree then by Lemma 1 all vertices of T are exhausted by W, hence W is an Eulerian cycle. Conversely, if W is an Eulerian cycle, by the BEST Theorem the subgraph composed by the last arc visited at each vertex is a tree, but this subgraph is T, concluding that T is a tree. $\square$

In the unrestricted case (when $\mathcal{W}^{\mathcal{F}}(n) = A^n$), the subgraph T is a regular tree of depth n where each non-leaf vertex has $|A|$ sons, therefore the minimal walk is an Eulerian cycle.

In the restricted case, we do not obtain necessarily an Eulerian cycle, because T is not necessarily a spanning tree converging to the root due to the existence of cycles.

We will study the structure of the graph $G^{\mathcal{F},n}$ and the subgraph T, specially the cycles in T. The main theorem of this section characterizes the label of cycles in T, allowing us to characterize the languages where the minimal walk is an Eulerian cycle.

First of all, we will prove some properties of the de Bruijn graph to understand the structure of the arcs and cycles in T.

Lemma 2. *Let $k \geq n + 2$. Let $W = v_0 e_0 v_1 e_1 \cdots e_{k-1} v_k$ be a walk in T. Then $l(e_0) \leq l(e_{n+1})$.*

Proof. Since $v_n = l(e_0) \cdots l(e_{n-1})$ we have that $l(e_1) \cdots l(e_{n-1})l(e_n)l(e_0) \in \mathcal{W}^{\mathcal{F}}(n+1)$. Hence there exists an arc (v_{n+1}, u) with label $l(e_0)$, where $v_{n+1} = l(e_1) \cdots l(e_{n-1})l(e_n)$. By the definition of T, $l(e_0) \leq \gamma(v_{n+1}) = l(e_{n+1})$. $\qquad \square$

Corollary 1. *Let C be a cycle in T. Then $|C|$ divides $n+1$. Moreover for every vertex u in C, $u\gamma(u) = l(C)^{\frac{n+1}{|C|}}$.*

Proof. Let consider the walk $W = v_0 e_0 \cdots e_{|C|-1} v_{|C|} = v_0 e_0 \cdots e_{(n+1)|C|-1} v_0 e_0 v_1$ as $n+1$ repetitions of the cycle C. From Lemma 2 we have $l(e_0) \leq l(e_{n+1}) \leq l(e_{2(n+1)}) \leq l(e_{(n+1)|C|}) = l(e_0)$. Since we can start the cycle in any vertex we conclude that $l(e_i) = l(e_{(n+1)+i})$ for every $i = 0, \ldots, |C| - 1$. Hence $|C|$ divides $n + 1$. The second conclusion comes from the fact that the label of any walk of length at most n ending in a vertex u is a suffix of u. $\qquad \square$

Let $u \neq m$ be a vertex. Among all the words which are prefix of m and suffix of u, let $g(u)$ be the longest one (notice that $g(u)$ could be the empty word ε and $|g(u)| < n$). Let $\alpha(u) = m_{|g(u)|+1}$ be the letter following the end of $g(u)$ in m.

Notice that in the unrestricted case, $|g(u)|$ is the distance over the graph from the vertex u to m. This function will be essential in the study of T. The next lemma give us a bound over the label of the arcs in terms of the function $g(\cdot)$.

Lemma 3. *For all pairs of adjacent vertices u and v, $l(uv) \leq \alpha(u)$. Moreover, if $l(uv) < \alpha(u)$ then $g(v) = \varepsilon$ and if $l(uv) = \alpha(u)$ then $g(v) = g(u)l(uv)$.*

Proof. $g(u)$ is a suffix of u, and $ul(uv) \in \mathcal{W}^{\mathcal{F}}(n+1)$, so $g(u)l(uv)$ is a prefix of a word in $\mathcal{W}^{\mathcal{F}}(n+1)$. Since m is the maximal word and $g(u)$ is a prefix of m we get $l(uv) \leq \alpha(u)$.

If $l(uv) = \alpha(u)$ then $g(u)l(uv)$ is a prefix of m and a suffix of v. Hence $g(u)l(uv)$ is a suffix of $g(v)$. Since by removing the last letter of a suffix of v we obtain a suffix of u we conclude $g(v) = g(u)l(uv)$.

We show that if $g(v) \neq \varepsilon$ then $\alpha(u) \geq l(uv)$. Let $g(v) = g'(v)l(uv)$, then $g'(v)$ is a suffix of u and a prefix of m. Hence $g'(v)$ is a suffix of $g(u)$. Therefore $g'(v)\alpha(u)$ is a factor of m. By the definition of $g(v)$ and the maximality of m $g(v)$ is greater or equal (lexicographically) than $g'(v)\alpha(u)$. We conclude that $\alpha(u) \geq l(uv)$.   □

In the unrestricted case, where T is a tree of depth n, all the arcs not in T go to a leaf. In the general case we can define an analog to the leaves.

We say that a vertex u is a *floor* vertex if $g(u) = \varepsilon$. Notice that in the unrestricted case the leaves of T are the floor vertices. We say that a vertex u is a *restricted* vertex if $\gamma(u) < \alpha(u)$.

Corollary 2. *If a cycle in T contains l restricted vertices, then it has exactly l floor vertices.*

Proof. From Lemma 3 we know that if a vertex u is restricted then for every arc (u, v) the vertex v is a floor vertex. To conclude it is enough to see that in T an arc (u, v) with u unrestricted has label $\alpha(u)$. Then v is not a floor vertex.   □

Corollary 3. *Let P be a path in T starting in a floor vertex, ending in a vertex v and with unrestricted inner vertices. Then $l(P) = g(v)$.*

Proof. We apply induction on the length of P. The case where the length of P is zero is direct since v is a floor vertex. Let us consider the case where P has length at least 1. Since v is not a restricted vertex, from Lemma 3 we know that $g(v) = g(u)l(uv)$, where u is its neighbor in P. By the induction assumption $g(u) = l(P')$ where P' is the path obtained from P removing the arc (u, v). Hence $g(v) = l(P')l(uv) = l(P)$.   □

We will use these results to characterize the label of cycles in T, specially we will characterize the restricted vertices of a cycle.

Theorem 3. *Let C be a cycle in T, let $u^0, \ldots, u^{k-1}$ be the restricted vertices in C ordered according to the order of C. Then $u^i = g(u^{i+1})\gamma(u^{i+1}) \cdots \gamma(u^{i-1})g(u^i)$ for $i = 0, \ldots, k-1$, where $i+1, \ldots, i-1$ are computed $\mod k$.*

Proof. From Corollary 3 the label of C is $g(u^0)\gamma(u^0) \cdots g(u^{k-1}) \gamma(u^{k-1})$, and by definition of $G^{\mathcal{F},n}$, u^i is the label of any walk over $G^{\mathcal{F},n}$ of length n finishing in u^i, so by Corollary 1 we can take the walk C^k composed by $k = (n+1)/|C|$ repetitions of C finishing in u^i, concluding that $u^i = g(u^{i+1})\gamma(u^{i+1}) \cdots \gamma(u^l)$ $l(C^{k-1})g(u_1) \cdots \gamma(u^{i-1}) g(u^i)$.   □

Now we are able to give a characterization of the languages where a minimal walk produces an Eulerian cycle.

Let $\mathcal{H}$ be the subset of $\mathcal{W}^{\mathcal{F}}(n+1)$ where $w \in \mathcal{H}$ if and only if w can be decomposed by $w = h^0\beta_1 \ldots h^{k-1}\beta_{k-1}$ where each $h^i \in A^*$ and $\beta_i \in A$ satisfy the following conditions:
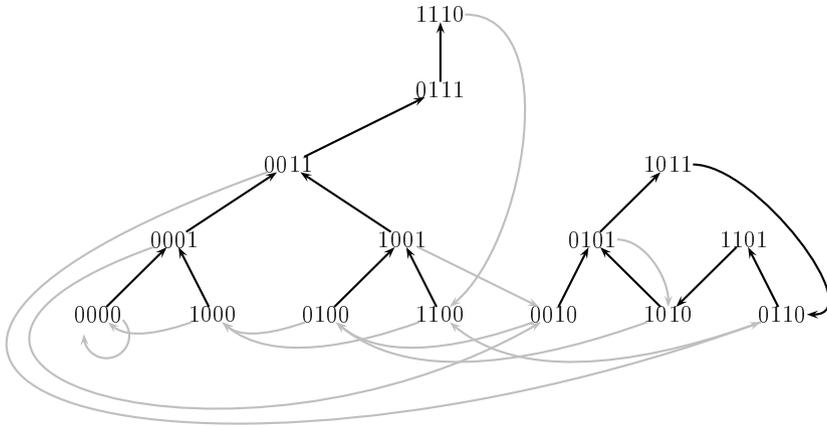
Fig. 2. Example of the subgraph T for $n = 4$ and $\mathcal{F} = \{01111\}$ in a binary alphabet.

1. $h^i = m_1 \ldots m_{|h^i|}$ (a prefix of m)
2. $\beta_i < m_{|h^i|+1}$
3. $\forall \beta' > \beta_i, \ h^{i+1} \beta_{i+1} \ldots \beta_{i-1} h^i \beta' \notin \mathcal{W}^{\mathcal{F}}(n+1)$

Now, we are able to characterize the languages where a minimal walk is an Eulerian cycle.

Theorem 4. *A minimal walk is an Eulerian cycle if and only if $\mathcal{H} = \emptyset$.*

Proof. From Theorem 2, we have to prove that T is a tree if and only if $\mathcal{H} = \emptyset$.

If T is not a tree then T has a cycle C. Let $u^0 \ldots u^{k-1}$ be the restricted vertices of the cycle. By Theorem 3 $l(C) = g(u^0)\gamma(u^0) \ldots g(u^{k-1})\gamma(u^{k-1})$ and by Corollary 1 $|C|$ divides $n + 1$. Therefore there exists a word w in $\mathcal{W}^{\mathcal{F}}(n+1)$ composed by $(n+1)/|C|$ repetitions of C. By definition of $\mathcal{H}$ we conclude that $w \in \mathcal{H}$.

Conversely, let us assume that T has no cycles and $\mathcal{H} \neq \emptyset$. Let w be a word in $\mathcal{H}$. By definition of $G^{\mathcal{F},n}$, there is a cycle C in $G^{\mathcal{F},n}$ of length dividing $n + 1$ such that C (or repetitions of C) has label w. We shall prove that C is also a cycle in T.

Let v be a vertex of C, with $v = \ldots \beta_{i-1}(h^i)_1 \ldots (h^i)_j$ where $j = 0 \ldots |h^i|$. If $0 < j < |h^i|$, then $m_1 \ldots m_j$ is a suffix of v, so $\alpha(v) = m_{j+1} = (h^i)_{j+1}$ hence the arc of C with tail at v is in T. If $j = 0$ then $\gamma(v) = m_1$ therefore the arc in C is in T. Finally, let consider the case $j = |h^i|$. If (v, v') is the arc in C then $l(vv') = \beta_i$. Since $w \in \mathcal{H}$, no arc in $G^{\mathcal{F},n}$ with tail at v has a label greater than β_i. Then $(v, v') \in T$. We conclude that C is a cycle in T which leads to a contradiction. $\qquad\square$

4   Some Remarks

The previous analysis considers only the minimal walk starting at the root vertex. This case does not necessarily produce the minimal label over all Eulerian cycles, because there can be Eulerian cycles starting at a non root vertex with a lexicographically lower label.

It is also possible to construct an algorithm which modifies T in order to destroy cycles in T, and obtain the minimal de Bruijn sequence for any irreducible subshift of finite type. However further research in this subject allow us to construct an algorithm to obtain the minimal Eulerian cycle for any edge-labeled digraph (see [12]), but this result escapes to the scope of this work.

References

1. de Bruijn, N.G.: A combinatorial problem. Nederl. Akad. Wetensch., Proc. **49** (1946) 758–764
2. Stein, S.K.: The mathematician as an explorer. Sci. Amer. **204** (1961) 148–158
3. Bermond, J.C., Dawes, R.W., Ergincan, F.Ö.: De Bruijn and Kautz bus networks. Networks **30** (1997) 205–218
4. Chung, F., Diaconis, P., Graham, R.: Universal cycles for combinatorial structures. Discrete Math. **110** (1992) 43–59
5. Fredricksen, H.: A survey of full length nonlinear shift register cycle algorithms. SIAM Rev. **24** (1982) 195–221
6. Fredricksen, H., Maiorana, J.: Necklaces of beads in k colors and k-ary de Bruijn sequences. Discrete Math. **23** (1978) 207–210
7. Ruskey, F., Savage, C., Wang, T.M.: Generating necklaces. J. Algorithms **13** (1992) 414–430
8. Ruskey, F., Sawada, J.: Generating necklaces and strings with forbidden substrings. Lect. Notes Comput. Sci. **1858** (2000) 330–339
9. Moreno, E.: Lyndon words and de bruijn sequences in a subshift of finite type. In Harju, T., Karhumäki, J., eds.: Proceedings of WORDS'03. Number 27 in TUCS General Publications, Turku, Finland, Turku Centre for Computer Science (2003) 400–410
10. Lind, D., Marcus, B.: Symbolic Dynamics and Codings. Cambridge University Press (1995)
11. Tutte, W.T.: Graph theory. Volume 21 of Encyclopedia of Mathematics and its Applications. Addison-Wesley Publishing Company Advanced Book Program, Reading, MA (1984)
12. Matamala, M., Moreno, E.: Minimal Eulerian cycle in a labeled digraph. Technical Report CMM-B-04/08-108, DIM-CMM, Universidad de Chile (2004)