De Bruijn Sequences Revisited

Lila Kari* Zhi Xu†

The University of Western Ontario,
London, Ontario, Canada N6A 5B7
lila@csd.uwo.ca zxu@google.com

Abstract

A (non-circular) de Bruijn sequence w of order n is a word such that every word of length n appears exactly once in w as a factor. In this paper, we generalize the concept to different settings: the multi-shift de Bruijn sequence and the pseudo de Bruijn sequence. An m-shift de Bruijn sequence of order n is a word such that every word of length n appears exactly once in w as a factor that starts at a position $im + 1$ for some integer $i \geq 0$. A pseudo de Bruijn sequence of order n with respect to an antimorphic involution θ is a word such that for every word u of length n the total number of appearances of u and $\theta(u)$ as a factor is one. We show that the number of m-shift de Bruijn sequences of order n is $a^n!a^{(m-n)(a^n-1)}$ for $1 \leq n \leq m$ and is $(a^m!)^{a^{n-m}}$ for $1 \leq m \leq n$, where a is the size of the alphabet. We provide two algorithms for generating a multi-shift de Bruijn sequence. The multi-shift de Bruijn sequence is important for solving the Frobenius problem in a free monoid. We show that the existence of pseudo de Bruijn sequences depends on the given alphabet and antimorphic involution, and obtain formulas for the number of such sequences in some particular settings.

1 Introduction

If a word w can be written as $w = xyz$, then the words x, y, and z are called the *prefix*, *factor*, and *suffix* of w, respectively. A word w over Σ is

*This research was partially supported by a Natural Sciences and Engineering Council of Canada Discovery Grant, and Canada Research Chair Award to L.K.

†Part of the work was done during this author's stay at University of Waterloo, supported by D. R. Cheriton Scholarship. This author's current address is Google Waterloo, 151 Charles St. W #200, Kitchener, Ontario, Canada N2G 1H6

called a *de Bruijn sequence* of order n if each word in Σ^n appears exactly once in w as a factor. For example, 00110 is a binary de Bruijn sequence of order 2 since each binary word of length two appears in it exactly once as a factor: $\texttt{00110} = \texttt{(00)110} = \texttt{0(01)10} = \texttt{00(11)0} = \texttt{001(10)}$. The de Bruijn sequence can be understood by the following game. Suppose there is an infinite supply of balls, each of which is labeled by a letter in Σ, and suppose there is a glass pipe that can hold balls in a vertical line. On the top of that pipe is an opening, through which one can drop balls into that pipe, and on the bottom is a trap-door, which can support the weight of at most n balls. When there are more than n balls in the pipe, the trap-door opens and those balls at the bottom drop off until only n balls remain. If we put balls with letters in the order as appeared exactly in a de Bruijn sequence of order n on the alphabet Σ, then every n ball sequence will appear exactly once in the pipe. It is easy to see that a de-Bruijn sequence of order n, if it exists, is of length $|\,\Sigma\,|^n + n - 1$ and its suffix of length $n-1$ is identical to its prefix of length $n-1$. So, sometimes a de-Bruijn sequence is written in a circular form by omitting the last $n-1$ letters.

The de Bruijn sequence is also called the de Bruijn-Good sequence, named after de Bruijn [2] and Good [10] who independently studied the existence of such words over the binary alphabet; the former also calculated the formula 2^{2^n} for the total number of those words of order n. The study of the de Bruijn sequence, however, dates back at least to 1894, when Flye Sainte-Marie [6] studied the words and provided the same formula 2^{2^n}. For an arbitrary alphabet Σ, van Aardenne-Ehrenfest and de Bruijn [1] provided the formula $(|\,\Sigma\,|!)^{|\,\Sigma\,|^n}$ for the total number of de Bruijn sequences of order n. Besides the total number of de Bruijn sequences, another interesting topic is how to generate a de Bruijn sequence (arbitrary one, lexicographically least one, lexicographically largest one). For generating de Bruijn sequences, see the surveys [7, 17]. The de Bruijn sequence is some times called the full cycle [7], and has connections to the following concepts: feedback shift registers [9], normal words [10], generating random binary sequences [15], primitive polynomials over a Galois field [18], Lyndon words and necklaces [8], Euler tours and spanning trees [1]. There are generalizations of the de Bruijn sequences from various aspects, such as the de Bruijn torus (two-dimensional generalization). Usually, the de Bruijn sequences are represented by their circular counterparts.

In this paper, we consider two generalizations of the de Bruijn sequence, namely the multi-shift de Bruijn sequence and the pseudo de Bruijn sequence. To understand the concept of multi-shift de Bruijn sequence, let us return to the glass pipe game presented at the beginning. Now the trap-door can support more weight. When there are $n+m$ or more balls in the pipe, the trap-door opens and the balls drop off until there are only n balls in the pipe. Is there an arrangement of putting the balls such that every n ball sequence appears exactly once in the pipe? The answer is "Yes" for arbitrary positive integers m, n. The solution represents a multi-shift de Bruijn sequence. We will discuss the existence of the multi-shift de Bruijn sequence, the total number of multi-shift de Bruijn sequences, generating a multi-shift de Bruijn sequence, and the application of the multi-shift de Bruijn sequence to the Frobenius problem in a free monoid, which is the original motivation we study the multi-shift de Bruijn sequence. To understand the concept of pseudo de Bruijn sequence, we first let the mirror image be the chosen antimorphic involution, where the concept of antimorphic involution is of particular interest in the study of bioinformation. Now if every n ball sequence either appears in the normal order or in a reversed order in the pipe and appears exactly once in this way, then the solution represents a pseudo de Bruijn sequence. No pseudo de Bruijn sequence exist for certain alphabets and antimorphic involutions. We will discuss the total number of pseudo de Bruijn sequences in particular settings.

2 Multi-Shift Generalization of the de Bruijn Sequence

Let $\Sigma \subseteq \{ \mathtt{0}, \mathtt{1}, \dots \}$ be the *alphabet* and let $w = a_1 a_2 \cdots a_n$ be a word over Σ. The *length* of w is denoted by $|\, w \,| = n$ and the *factor* $a_i \cdots a_j$ of w is denoted by $w[i \mathinner{..} j]$. If $u = w[im+1 \mathinner{..} im+n]$ for some non-negative integer i, we say the factor u appears in w at *a modulo m position.* The set of all words of length n is denoted by Σ^n and the set of all finite words is denoted by $\Sigma^* = \{ \epsilon \} \cup \Sigma \cup \Sigma^2 \cdots$, where ϵ is the *empty word.* The concatenation of two words u, v is denoted by $u \cdot v$, or simply uv.

Multi-shift de Bruijn sequences are implicitly defined and used in the second author's paper [11] in solving the Frobenius problem in a free monoid. The precise definition of the multi-shift de Bruijn sequence is given below.

Definition 1. *A word w over Σ is called a* multi-shift de Bruijn sequence *of shift m and order n, if each word in Σ^n appears exactly once in w as a factor at a modulo m position.*

For example, one of the 2-shift de Bruijn sequences of order 3 is

$$\texttt{00010011100110110},$$

which can be verified as follows:

$$\begin{aligned}&\texttt{00010011100110110} = \texttt{(000)10011100110110} = \texttt{00(010)011100110110}\\ &= \texttt{0001(001)1100110110} = \texttt{000100(111)00110110} = \texttt{00010011(100)110110}\\ &= \texttt{0001001110(011)0110} = \texttt{000100111001(101)10} = \texttt{00010011100110(110)}.\end{aligned}$$

The multi-shift de Bruijn sequence generalizes the de Bruijn sequence in the sense that de Bruijn sequences are exactly 1-shift de Bruijn sequences of the same order. It is easy to see that the length of each m-shift de Bruijn sequence of order n, if it exists, is equal to $m|\,\Sigma\,|^n + (n - m)$. By the definition of multi-shift de Bruijn sequence, the following proposition holds.

Proposition 2. *Let w be one m-shift de Bruijn sequence w of order n, $n > m$. Then the suffix of length $n - m$ of w is identical to the prefix of length $n - m$ of w.*

From Proposition 2, we know that when $n > m$, every multi-shift de Bruijn sequence can be written as a circular word and the discussion on multi-shift de Bruijn sequences of the two different forms are equivalent. In this paper, we discuss the multi-shift de Bruijn sequence in the form of ordinary words.

A *(non-strict) directed graph*, or *digraph* for short, is a triple $G = (V, A, \psi)$ consisting of a set V of *vertices*, a set A of *arcs*, and an *incidence function* $\psi : A \to V \times V$. Here we do not take the convention $A \subseteq V \times V$, since we allow a digraph to contain self-loops on a single vertex and multiple arcs between the same pair of vertices. When $\psi(a) = (u, v)$, we say the arc a joins u to v, where vertex $u = \text{tail}(a)$ and vertex $v = \text{head}(a)$ are called *tail* and *head*, respectively. The indegree $\delta^-(v)$ (outdegree $\delta^+(v)$, respectively) of a vertex v is the number of arcs with v being the head (the tail, respectively). A *walk* in G is a sequence $a_1, a_2, \ldots, a_k$ such that $\text{head}(a_i) = \text{tail}(a_{i+1})$ for each $1 \le i < k$. The walk is *closed*, if $\text{head}(a_k) = \text{tail}(a_0)$. Two closed walks are regarded as identical if one is the circular shift of the other. An *Euler tour* is a closed walk that traverses each arc exactly once. A *Hamilton cycle* is a closed walk that traverses each vertex exactly once. An *(spanning) arborescence* is a digraph with a particular vertex, called the *root*, such that it contains every vertex of G, its number of arcs is exactly one less than the number of vertices, and there is exactly one walk from the root to any other vertex. We denote the total number of Euler tours, Hamilton cycles, and arborescences of G by $|\,G\,|_E$, $|\,G\,|_H$, and $|\,G\,|_A$, respectively.

An *(undirected) graph* is defined as a digraph such that for any pair of vertices v_1, v_2, there is an arc a, $\psi(a) = (v_1, v_2)$, if and only if there is a corresponding arc a', $\psi(a') = (v_2, v_1)$. In this case, we write $\delta^-(v) = \delta^+(v) = \delta(v)$ and a spanning arborescence is just a *spanning tree*.

The line-graph $L(G)$ of $G = (V, A, \psi)$ is defined as (A, C, φ) such that for every pair of arcs $a_1, a_2 \in A$, $\mathrm{head}(a_1) = \mathrm{tail}(a_2)$, there is an arc $c \in C$, $\varphi(c) = (a_1, a_2)$ and those arcs are the only arcs in C. Euler tours exist in a graph G if and only if Hamilton cycles exist in the line-graph $L(G)$.

We define the word graph $G(m,n)$ by $(\Sigma^n, \Sigma^{n+m}, \psi)$, where $\psi(w) = (u, v)$ for $u = w[1\,..\,n], v = w[m+1\,..\,m+n]$. Then by definition, the following lemmas are straightforward.

Lemma 3. *The digraph $L(G(m,n))$ is the digraph $G(m, n+m)$.*

Lemma 4. *Suppose $m \le n$. (1) There is a $|\,\Sigma\,|^n$-to-1 mapping from the set of m-shift de Bruijn sequences of order n onto the set of Hamilton cycles in $G(m,n)$. (2) There is a $|\,\Sigma\,|^n$-to-1 mapping from the set of m-shift de Bruijn sequences of order n onto the set of Euler tours in $G(m, n-m)$.*

Theorem 5. *For any alphabet Σ, positive integers m, n, some m-shift de Bruijn sequences of order n over Σ exist.*

Proof. First we assume $m \ge n$. Let $u_1, u_2, \ldots, u_l$ be any permutation of the words in Σ^n for $l = |\,\Sigma\,|^n$. Then the word $u_1 \mathtt{0}^{m-n} u_2 \mathtt{0}^{m-n} \cdots \mathtt{0}^{m-n} u_l$ is one m-shift de Bruijn sequence of order n over Σ.

Now we assume $m < n$ and prove there exists an Euler tour in $G(m, n-m)$. Then by Lemma 4, the existence of m-shift de Bruijn sequences of order n over Σ is ensured. To show the existence of an Euler tour, we only need to verify that $G(m, n-m)$ is connected and that $\delta^-(v) = \delta^+(v)$ for every vertex v, both of which are straightforward: for every vertex v in $G(m, n-m)$, v is connected to the vertex $\mathtt{0}^{n-m}$ in both directions and $\delta^-(v) = \delta^+(v) = |\,\Sigma\,|^m$. $\square$

2.1 Counting the Number of Multi-Shift de Bruijn Sequences

Since m-shift de Bruijn sequences of order n exist, in this section we discuss the total number of different m-shift de Bruijn sequences of order n, and we denote the number by $\#(m,n)$. First, we study the degenerate case.

Lemma 6. *For $1 \leq n \leq m$, $\#(m,n) = a^n! a^{(m-n)(a^n-1)}$, where $a = |\,\Sigma\,|$.*

To study the case $1 \leq m \leq n$, we need a theorem by van Aardenne-Ehrenfest and de Bruijn [1], which describes the relation between the number of Euler tours in a particular type of digraph and the number of Euler tours in its line-graph.

Theorem 7 (van Aardenne-Ehrenfest and de Bruijn)**.** *Let $G = (V, A, \psi)$ be a digraph such that $a = \delta^-(v) = \delta^+(v)$ for every $v \in V$. Then $|\,L(G)\,|_E = a^{-1}(a!)^{|\,V\,|(a-1)}|\,G\,|_E$.*

The digraph $G(m,n)$ satisfies the conditions in Theorem 7 with $a = |\,\Sigma\,|^m$. So, by the relation between the multi-shift de Bruijn sequences and the Euler tours in the word graph $G(m,n)$, we have the following recursive expression on $\#(m,n)$.

Lemma 8. *For $m \geq 1, n \geq 2m$, $\#(m,n) = (a^m!)^{a^{n-m}-a^r} \#(m, m+r)$, where $a = |\,\Sigma\,|$, $r = n \bmod m$.*

To finish the last step of obtaining $\#(m,n)$ for $1 \leq m \leq n$, we again need two theorems, the BEST theorem [19, 1] and Kirchhoff's matrix tree theorem [14], which are often used in the literature to count the number of Euler tours in various types of digraphs.

Theorem 9 (BEST theorem)**.** *In a digraph $G = (V, A, \psi)$, the number of Euler tours and arborescences satisfy $|\,G\,|_E = \prod_{v \in V}(\delta^+(v) - 1)!\,|\,G\,|_A$.*

Theorem 10 (Kirchhoff's matrix tree theorem)**.** *In a graph $G = (V, A, \psi)$, the number of spanning trees is equal to any cofactor of the Laplacian matrix of G, which is the diagonal matrix of degrees minus the adjacency matrix.*

Lemma 11. *For $1 \leq m \leq n \leq 2m$, $\#(m,n) = (a^m!)^{a^{n-m}}$, where $a = |\,\Sigma\,|$.*

Proof. Let $r = n - m$ and $a = |\,\Sigma\,|$. Then $0 \leq r \leq m$. By definition, $G = G(m, n-m) = (\Sigma^r, \Sigma^m, \psi)$. So from any vertex to any vertex, there are a^{m-r}-many arcs in G. We convert G into a undirected graph G' by omitting all self-loops; there are a^{m-r}-many of them for each vertex. Since for every pair of vertices v_1, v_2 there are a^{m-r}-many arcs joins v_1 to v_2 and correspondingly there are a^{m-r}-many arcs joins v_2 to v_1, the graph G' is indeed an undirected graph by our definition. Each vertex in G' is of degree $a^m - a^{m-r}$. Then the Laplacian matrix of G' is

$$L = \begin{pmatrix} a^m - a^{m-r} & -a^{m-r} & \cdots & -a^{m-r} \\ -a^{m-r} & a^m - a^{m-r} & \cdots & -a^{m-r} \\ \vdots & \vdots & \ddots & \vdots \\ -a^{m-r} & -a^{m-r} & \cdots & a^m - a^{m-r} \end{pmatrix}.$$

By Theorem 10, the number of arborescences $|\,G\,|_A = |\,G'\,|_A$ is equal to the cofactor of L, which is $(a^m)^{a^r-2}a^{m-r} = (a^m)^{a^r}/a^n$. Then by Theorem 9, the number of Euler tours in digraph G is $|\,G\,|_E = ((a^m-1)!)^{a^r}|\,G\,|_A = ((a^m-1)!)^{a^r}(a^m)^{a^r}/a^n = (a^m!)^{a^r}/a^n$. Finally, by Lemma 4, the number of m-shift de Bruijn sequences of order n is $\#(m,n) = a^n|\,G\,|_E = (a^m!)^{a^r}$. □

Theorem 12. *For $1 \le n \le m$, $\#(m,n) = a^n!a^{(m-n)(a^n-1)}$, and for $1 \le m \le n$, $\#(m,n) = (a^m!)^{a^{n-m}}$, where $a = |\,\Sigma\,|$.*

Proof. For $1 \le n \le m$, the equality $\#(m,n) = a^n!a^{(m-n)(a^n-1)}$ is shown in Lemma 6. Now we assume $1 \le m \le n$. Let $r = n \bmod m$. Following Lemmas 8,11, we have $\#(m,n) = (a^m!)^{a^{n-m}-a^r}\#(m,m+r) = (a^m!)^{a^{n-m}-a^r}(a^m!)^{a^r} = (a^m!)^{a^{n-m}}$. □

2.2 Generating Multi-Shift de Bruijn Sequences

In this section, we study the problem of generating one m-shift de Bruijn sequence of order n for arbitrary alphabet and positive integers m, n. When $1 \le n \le m$, a m-shift de Bruijn sequence of order n is easy to construct as given in Theorem 5. Now we consider the case $1 \le m < n$. We will present two algorithms for generating a m-shift de Bruijn sequence of order n.

We claim that m-shift de Bruijn sequences of order km can be generated using the ordinary de Bruijn sequence generating algorithm, such as described by Fredricksen [7]. To do this, we first generate a de Bruijn sequence w of order k over the alphabet $\Gamma = \Sigma^m$. Then we replace each letter of w in Γ by the corresponding word of length m over Σ. It is easy to see that the new word is a m-shift de Bruijn sequence of order km.

The first algorithm of generating multi-shift de Bruijn sequence is to generate m_i-shift de Bruijn sequences of order $k_i m_i$ for some $k_i, m_i, i = 1, 2$ before rearranging the words to obtain an arbitrary m-shift de Bruijn sequence of order n. Let $1 \le m < n$ be two integers, and $n = km + r$, where $r = n \bmod m$. The case $r = 0$ is already discussed and the case $|\,\Sigma\,| = 1$ is trivial. So we assume $r \ne 0$ and $|\,\Sigma\,| \ge 2$. We define $m_1 = r$, $n_1 = (k+1)r$ and generate $w_1 = \tau(m_1, n_1)\mathtt{0}^{m_1}$ such that $\tau(m_1, n_1)$ is a m_1-shift de Bruijn sequence of order n_1 and $w_1[1\,..\,n_1] = \mathtt{0}^{n_1}$; and define $m_2 = m - r$, $n_2 = k(m-r)$ and generate $w_2 = \tau(m_2, n_2)\mathtt{0}^{m_2}$ such that $\tau(m_2, n_2)$ is a m_2-shift de Bruijn sequence of order n_2 and $w_2[1\,..\,n_2] = \mathtt{0}^{n_2}$. Let $a = |\,\Sigma\,|$, $N_1 = a^{n_1}$, $N_2 = a^{n_2}$. We define $u_i = w_1[n_1 + (i-1)m_1 + 1\,..\,n_1 + im_1]$, $u'_i = u_{1+(i \bmod (N_1-1))}$, $v_i = w_2[n_2 + (i-1)m_2 + 1\,..\,n_2 + im_2]$, $v'_i = v_{1+(i-1 \bmod N_2)}$.

Input: two integers m, n with $1 \leq m < n$ and alphabet size a.
Output: an m-shift de Bruijn sequence of order n over $\{\mathtt{0}, \ldots, a-1\}$.

Let $n = km + r$, where $r = n \bmod m$;
if $r = 0$ **then return** *an m-shift de Bruijn sequence of order n* ;
generate an r-shift de Bruijn sequence of order $(k+1)r$;
generate an $(m-r)$-shift de Bruijn sequence of order $k(m-r)$;
return *a word as constructed by Eq. (1)*

Figure 1: Generating a multi-shift de Bruijn sequence, method one.

Then the following word

$$\mathtt{0}^n\, v_1 \mathtt{0}^{m_1}\, v_2\, \cdots\, v_{N_2-1} \mathtt{0}^{m_1}\, v_{N_2} u'_{(N_1-1)N_2}\, v'_1 u'_1\, v'_2 u'_2\, \cdots\, v'_{(N_1-1)N_2-1} u'_{(N_1-1)N_2-1} \tag{1}$$

is one m-shift de Bruijn sequence of order n, where $v_{N_2} = \mathtt{0}^{km}$ and $u'_{(N_1-1)N_2} = u_1$. The algorithm is illustrated in Fig. 1.

Theorem 13. *The algorithm in Fig. 1 correctly generates an m-shift de Bruijn sequence of order n.*

Now, we will see an example. Consider generating a 2-shift de Bruijn sequence of order 5. Then $m_1 = 1, n_1 = 3, m_2 = 1, n_2 = 2$ and we can obtain two words $w_1 = \mathtt{00011101000}$, which is $\tau(1,3)\mathtt{0}$, and $w_2 = \mathtt{001100}$, which is $\tau(1,2)\mathtt{0}$. So one 2-shift de Bruijn sequence of order 5 is as follows

$$\mathtt{00000}1_2\mathtt{0}1_2\mathtt{0}0_2\mathtt{0}0_2$$
$$1_1 1_2 1_1 1_2 1_1 0_2 0_1 0_2 1_1 1_2 0_1 1_2 0_1 0_2 1_1 0_2 1_1 1_2 1_1 1_2 0_1 0_2 1_1 0_2 0_1 1_2 0_1 1_2$$
$$1_1 0_2 1_1 0_2 1_1 1_2 0_1 1_2 1_1 0_2 0_1 0_2 0_1 1_2 1_1 1_2 1_1 0_2 1_1 0_2 0_1 1_2 1_1 1_2 0_1 0_2 0_1,$$

where the subscripts 1 and 2 denote whether the letter is from the word w_1 (words u_i, u'_i) or from the word w_2 (words v_i, v'_i).

Now we present the second algorithm, which uses the same idea of "prefer one" algorithm [16] for generating ordinary de Bruijn sequences. Let m, n be two positive integers. To generate a m-shift de Bruijn sequence w of order n, we start the sequence w with n zeros. Then we append to the end of current sequence w the lexicographically largest word of length m such that the suffix of length n of new sequence has not yet appeared as factor at a modulo m position. We repeat this step until no word can be appended to w. The algorithm is illustrated in Fig. 2.

Input: two integers m, n with $1 \le m < n$ and alphabet size a.
Output: an m-shift de Bruijn sequence of order n over $\{0, \ldots, a-1\}$.

Let $w := 0^n$;
Mark all word of length n except w as unvisited ;
repeat
  Find the lexicographically largest u of length m such that $w[|w| - n + m + 1 \,..\, |w|]u$ is unvisited ;
  Then let $w := wu$ and mark word $w[|w| - n + m + 1 \,..\, |w|]u$ visited ;
until *no such word can be found*;
return w

Figure 2: Generating a multi-shift de Bruijn sequence, method two.

Theorem 14. *The algorithm in Fig. 2 correctly generates an m-shift de Bruijn sequence of order n.*

Now, we use the algorithm to generate one 2-shift de Bruijn sequence of order 5. Starting from 00000, since 00011 does not appear as a factor at a modulo 2 position, we append 11 to the current sequence 00000. Repeating this procedure and appending words 11, 11, 10, 11, …, finally we obtain the word:

0000011111110111010110111011001110011001
010011000100001010100010000

If we circularly move the prefix 0^n to the end, the sequence generated by the second algorithm is the lexicographically largest m-shift de Bruijn sequence of order n.

2.3 Application to the Frobenius Problem in a Free Monoid

The study of multi-shift de Bruijn sequences is inspired by a problems of words, called the Frobenius problem in a free monoid. Given k integers $x_1, \ldots, x_k$, such that $\gcd(x_1, \ldots, x_k) = 1$, then there are only finitely many positive integers that *cannot* be written as a non-negative integer linear combination of $x_1, \ldots, x_k$. The integer *Frobenius problem* is to find the largest such integer, which is denoted by $g(x_1, \ldots, x_k)$. For example, $g(3, 5) = 7$.

If words $x_1, \dots, x_k$, instead of integers, are given such that there are only finitely many words that *cannot* be written as concatenation of words from the set $\{ x_1, \dots, x_k \}$, the *Frobenius problem in a free monoid* [11] is to find the longest such words. If all $x_1, \dots, x_k$ are of length either m or n, $0 < m < n$, there is an upper bound: the length of the longest word that cannot be written as concatenation of words from the set $\{ x_1, \dots, x_k \}$ is less than or equal to $g(m, l) = ml - m - l$, where $l = m\Sigma^{n-m} + n - m$. [11] Furthermore, the upper bound is tight and the construction is based on the multi-shift de Bruijn sequences. We denote the set of all words that can be written as the concatenation of words in S, including the empty word, by S^*.

Theorem 15. *[11] There exists $S \subseteq \Sigma^m \cup \Sigma^n$, $0 < m < n$, such that $\Sigma^* \setminus S^*$ is finite and the longest words in $\Sigma^* \setminus S^*$ constitute exactly the language $(\tau\Sigma^m)^{m-2}\tau$, where τ is a m-shift de Bruijn sequence of order $n - m$.*

For example, for any set of words $S \subseteq U = \{ \mathtt{0}, \mathtt{1} \}^3 \cup \{ \mathtt{0}, \mathtt{1} \}^7$ such that $\{ \mathtt{0}, \mathtt{1} \}^* \setminus S^*$ is finite, the longest words in $\{ \mathtt{0}, \mathtt{1} \}^* \setminus S^*$ are of length less than or equal to $g(3, 3 \cdot 2^4 + 4) = g(3, 52) = 101$. To construct S to reach the upper bound, we first choose an anbitrary 3-shift de Bruijn sequence of order 4 as $\tau =$ `0000111111110110101101100100011011010010001001000`. Then based on τ, we construct the set $S = U \setminus \{$ `0000111`, `0111111`, `1111110`, `1110110`, `0110101`, `0101101`, `1101100`, `1100100`, `0100011`, `0011011`, `1011010`, `1010010`, `0010001`, `0001001`, `1001000` $\}$. We have $L = \{ \mathtt{0}, \mathtt{1} \}^* \setminus S^* = \tau\{ \mathtt{0}, \mathtt{1} \}^3\tau$ and one of the longest words in L of length exactly 101 is given below:

```
0000111111110110101101100100011011010010001001000
          1110000111111110110101101100100011011010010001001000.
```

3 Pseudo de Bruijn Sequence Defined by Antimorphic Involutions

Here we discuss another generalization of the de Bruijn sequence. Let $\Sigma \subseteq \{ \mathtt{0}, \mathtt{1}, \mathtt{2}, \dots \}$ be the alphabet. A function $\theta : \Sigma^* \to \Sigma^*$ is called an *involution* if $\theta(\theta(w)) = w$ for $w \in \Sigma^*$ and called an *antimorphism* if $\theta(uv) = \theta(v)\theta(u)$ for $u, v \in \Sigma^*$. We call θ an *antimorphic involution* if θ is both an involution and an antimorphism. For example, the classic

Watson-Crick complementarity of DNA strands in biology is an antimorphic involution over the four-letter alphabet of DNA nucleotides { A, T, C, G }, where θ(A) =T, θ(C) =G, and θ(ACG) =CGT. The *mirror image*, or *reverse*, $\theta(a_1a_2\cdots a_n) = a_n\cdots a_2a_1$ is another antimorphic involution. Let θ be an antimorphic involution. We write $tr(\theta) = \{\, a : a \in \Sigma, \theta(a) \neq a \,\}$ and thus θ can be written as composition of $tr(\theta)$ transpositions with a mirror image. The antimorphic involution is motivated by the particularities of DNA-encoded information for the purpose of DNA computing. Several concepts in combinatorics on words have natural counterparts in this setting, e.g., pseudo-palindromes [5], involutively bordered words [13], Watson-Crick conjugate words, Watson-Crick commutativity [12], pseudo-primitive words [4], and pseudo-powers of words [3]. In the following, we define and discuss the pseudo de Bruijn sequence.

Definition 16. *A word w over Σ is called a* pseudo de Bruijn sequence *of order n if for every word $x \in \Sigma^n$, either x or $\theta(x)$ appears in w as a factor and the total number of those appearances is exactly one.*

For example, `0011` is a pseudo de Bruijn sequence of order 2 with respect to the mirror image (word reverse), by the following observation:

$$\mathtt{0011} = (\mathtt{00})\mathtt{11} = \mathtt{0}(\mathtt{01})\mathtt{1} = \mathtt{0}\theta(\mathtt{10})\mathtt{1} = \mathtt{00}(\mathtt{11}).$$

As we saw in Section 2, most properties of the multi-shift de Bruijn sequence are analogous to those of the usual de Bruijn sequence. This is not true for the pseudo de Bruijn sequence.

3.1 Contrast Between the Usual de Bruijn Sequence and the Pseudo de Bruijn Sequence

The length of a de Bruijn sequence of order n over Σ is $a^n + n - 1$ (or a^n in the circular form), where $a = |\,\Sigma\,|$. By contrast, the length of a pseudo de Bruijn sequence of order n over Σ is $N + n - 1$, where $N = |\,\Sigma\,|^n - |\,\{\, u : u \in \Sigma^n, \theta(u) \neq u \,\}\,|/2$. More precisely:

Proposition 17. *A pseudo de Bruijn sequence of order n over Σ with respect to θ is of length $\left(a^n + (a - 2\cdot tr(\theta))^{n \bmod 2} a^{\lfloor n/2 \rfloor}\right)/2 + (n-1)$, where $a = |\,\Sigma\,|$.*

Obviously, for a unary alphabet, we can always write a pseudo de Bruijn sequence in a circular form, since the last n letters are identical to the first n letters. In general, however, not all pseudo de Bruijn sequences can be written in a circular form.

Proposition 18. *Let $\Sigma = \{ \mathtt{0}, \mathtt{1} \}$, let θ be the mirror image, and let w be a binary de Bruijn sequence of order n. Then either $\mathtt{1}^n$ is a prefix of w and $\mathtt{0}^n$ is a suffix of w; or $\mathtt{0}^n$ is a prefix of w and $\mathtt{0}^n$ is a suffix of w.*

As a direct result, none of the binary de Bruijn sequence can be written in a circular form.

3.2 Counting the Number of Pseudo de Bruijn Sequences for Special Cases

For a pseudo de Bruijn sequence of order 1, say w, the word w is just a permutation of letters in Γ, where $\Gamma \subseteq \Sigma$ consists exactly of letters a with $\theta(a) = a$ and one of the letters b, c with $\theta(b) = c \neq b$. We have the follow proposition.

Proposition 19. *Let Σ be an alphabet and let θ be an antimorphic involution. Then the pseudo de Bruijn sequences of order 1 exist and the total number of them is $2^t(a-t)!$, where $a = |\Sigma|$ and $t = tr(\theta)$.*

Now we assume θ is the mirror image. There are two binary pseudo de Bruijn sequences, $\mathtt{0011}$ and $\mathtt{1100}$, of order 2. To discuss de Bruijn sequence over a more general alphabet, we need the following lemma.

Lemma 20. *Let Σ be an alphabet with $a = |\Sigma| \geq 3$ and let θ be the mirror image. Then every pseudo de Bruijn sequence of order 2 can be written in a circular form and there is an $\frac{a(a+1)}{2}$ to 1 mapping from the pseudo de Bruijn sequences of order 2 onto the Euler tours in K_a^o, where K_a^o is the complete graph K_a where a self-loop is added on each vertex.*

In contrast to the existence of ordinary de Bruijn sequence, not all pseudo de Bruijn sequences exist. In other words, the number of such sequences can be 0.

Proposition 21. *Let Σ be an alphabet with even $a = |\Sigma| \geq 4$ and let θ be the mirror image. Then there is no pseudo de Bruijn sequence of order 2.*

Proof. Since there is no Euler tour in K_a^o for a being even and $a \geq 4$, by Lemma 20, the number of pseudo de Bruijn sequences in this setting is 0. □

Discussion of the total number of Euler tours (also called Euler circuits) in a complete graph dates back at least to the year 1859 by Reiss, about 100 years after Euler's work on Königsberg Bridges Problem. The following proposition discloses the relation between the number of pseudo de Bruijn sequences of order 2 over an odd alphabet with respect to the mirror image and the number of Euler tours in a complete graph.

Proposition 22. *Let Σ be an alphabet with odd $a = |\Sigma| \geq 3$ and let θ be the mirror image. Then the pseudo de Bruijn sequences of order 2 exist and their total number is $\frac{(a-1)^a a(a+1)}{2^{a+1}} E_a$, where E_a is the total number of Euler tours in K_a.*

The precise formula for E_a is complicated and so far there is no closed form for E_a. We know that the formulae for the number of pseudo de Bruijn sequences is at least as hard as that for E_a and any formula for the latter leads to a formula of the former.

4 Conclusion

In this paper, we generalized the classic de Bruijn sequence to a new multi-shift setting and to a bioinformation inspired setting.

A word w is an m-shift de Bruijn sequence $\tau(m,n)$ of order n, if each word of length n appears exactly once as a factor at a modulo m position. An ordinary de Bruijn sequence is a 1-shift de Bruijn sequence.

We showed that the total number of distinct m-shift de Bruijn sequences of order n is $\#(m,n) = (a^n)! a^{(m-n)(a^n-1)}$ for $1 \leq n \leq m$ and is $\#(m,n) = (a^m!)^{a^{n-m}}$ for $1 \leq m \leq n$, where $a = |\Sigma|$. This result generalizes the formula $(a!)^{a^{n-1}}$ for the number of ordinary de Bruijn sequences [1]. Here we use an ordinary word form; if counting the sequences in a circular form, then the number is to be divided by a^n.

We provided two algorithms for generating a m-shift de Bruijn sequence of order n. The first algorithm is to rearrange factors from two simpler multi-shift de Bruijn sequences, where the order is a multiple of the shift. The second is the analogue of the "prefer one" algorithm (for example, see [7]) for generating ordinary de Bruijn sequence.

The multi-shift de Bruijn sequence has applications to the Frobenius problem in a free monoid by providing constructions of examples. It will be interesting to see whether this generalized concept of the de Bruijn sequence has an impact in other fields of theoretical computer science and discrete mathematics.

A word w is a pseudo de Bruijn sequence with respect to an antimorphic involution θ if for each word u of length n, either u or $\theta(u)$ appears as a factor and it appears exactly once in this way.

We showed that a binary pseudo de Bruijn sequence with respect to the mirror image does not have a circular form. We showed that a pseudo de Bruijn sequence of order 2 with respect to the mirror image over alphabet of even size ≥ 4 does not exist.

We showed that the number of pseudo de Bruijn sequence of order 2 with respect to the mirror image over an alphabet of odd size ≥ 3 is $(a-1)^a a(a+1)E_a/2^{a+1}$, where E_a is the total number of Euler tours in the complete graph K_a.

With respect to antimorphic involution other than the mirror image, no non-trivial property on the pseudo de Bruijn sequences is known.

Acknowledgements

The authors thank Prof. Jeffrey Shallit for comments on early drafts of the paper.

References

[1] T. van Aardenne-Ehrenfest and N. G. de Bruijn. Circuits and trees in oriented linear graphs. *Simon Stevin*, 28:203–217, 1951.

[2] N. G. de Bruijn. A combinatorial problem. *Indag. Math.*, 8(4):461–467, 1946.

[3] E. Chiniforooshan, L. Kari, and Z. Xu. Pseudo-power avoidance. In *DLT2010*, pages 432–433, 2010.

[4] E. Czeizler, L. Kari, and S. Seki. On a special class of primitive words. *Theoret. Comput. Sci.*, 411(3):617–630, 2010.

[5] A. de Luca and A. De Luca. Pseudopalindrome closure operators in free monoids. *Theoret. Comput. Sci.*, 362:282–300, 2006.

[6] C. Flye Sainte-Marie. Solution to question nr. 48. *L'Intermédiaire Math.*, 1:107–110, 1894.

[7] H. Fredricksen. A survey of full length nonlinear shift register cycle algorithms. *SIAM Review*, 24(2):195–221, 1982.

[8] H. Fredricksen and I. J. Kessler. Lexicographic compositions and de bruijn sequences. *J. Combin. Theory Ser. A*, 22:17–30, 1977.

[9] S. W. Golomb. *Shift Register Sequences.* Holden-Day, 1967.

[10] I. J. Good. Normal recurring decimals. *J. London Math. Soc.*, 21(3):167–169, 1946.

[11] J.-Y. Kao, J. Shallit, and Z. Xu. The Frobenius problem in a free monoid. In *STACS 2008*, pages 421–432, 2008.

[12] L. Kari and K. Mahalingam. Watson-Crick conjugate and commutative words. In *DNA13*, pages 273–283, 2007.

[13] L. Kari and S. Seki. On pseudoknot-bordered words and their properties. *J. Comput. System Sci.*, 75(2):113–121, 2009.

[14] G. Kirchhoff. Über die Auflösung der Gleichungen, auf welche man bei der untersuchung der linearen verteilung galvanischer Ströme geführt wird. *Ann. Phys. Chem.*, 72:497–508, 1847.

[15] D. E. Knuth. *The Art of Computer Programming.* Addison-Wesley, 1969.

[16] M. H. Martin. A problem in arrangements. *Bull. Amer. Math. Soc.*, 40:859–864, 1934.

[17] A. Ralston. De Bruijn sequences — a model example of the interaction of discrete mathematics and computer science. *Math. Mag.*, 55(3):131–143, 1982.

[18] D. Rees. Note on a paper by I. J. Good. *J. London Math. Soc.*, 21(3):169–172, 1946.

[19] W. T. Tutte and C. A. B. Smith. On unicursal paths in a network of degree 4. *Amer. Math. Monthly*, 48(4):233–237, 1941.