Journal of Combinatorics
Volume 8, Number 3, 439–474, 2017

Multi de Bruijn sequences

Glenn Tesler*

We generalize the notion of a de Bruijn sequence to a "multi de Bruijn sequence": a cyclic or linear sequence that contains every k-mer over an alphabet of size q exactly m times. For example, over the binary alphabet $\{0, 1\}$, the cyclic sequence (00010111) and the linear sequence 000101110 each contain two instances of each 2-mer $00, 01, 10, 11$. We derive formulas for the number of such sequences. The formulas and derivation generalize classical de Bruijn sequences (the case $m = 1$). We also determine the number of multisets of aperiodic cyclic sequences containing every k-mer exactly m times; for example, the pair of cyclic sequences $(00011)(011)$ contains two instances of each 2-mer listed above. This uses an extension of the Burrows-Wheeler Transform due to Mantaci et al., and generalizes a result by Higgins for the case $m = 1$.

AMS 2000 subject classifications: Primary 68R15, 05C30; secondary 05C38, 05C45, 05C81, 68P30.
Keywords and phrases: Multi de Bruijn sequences, de Bruijn graphs, Extended Burrows-Wheeler Transform, Lyndon words, Eulerian graphs.

1. Introduction

We consider sequences over a totally ordered alphabet Ω of size $q \geq 1$. A *linear sequence* is an ordinary sequence $a_1, \ldots, a_n$ of elements of Ω, denoted in string notation as $a_1 \ldots a_n$. Define the *cyclic shift* of a linear sequence by $\rho(a_1 a_2 \ldots a_{n-1} a_n) = a_n a_1 a_2 \ldots a_{n-1}$. In a *cyclic sequence*, we treat all rotations of a given linear sequence as equivalent:

$$(a_1 \ldots a_n) = \{\rho^i(a_1 \ldots a_n) = a_{i+1} \ldots a_n a_1 \ldots a_{i-1} : i = 0, \ldots, n-1\}.$$

Each rotation $\rho^i(a_1 \ldots a_n)$ is called a *linearization* of the cycle $(a_1 \ldots a_n)$.

A *k-mer* is a sequence of length k over Ω. The set of all k-mers over Ω is Ω^k. A *cyclic de Bruijn sequence* is a cyclic sequence over alphabet Ω (of size q) in which all k-mers occur exactly once. The length of such a sequence is $\ell = q^k$, because each of the q^k k-mers accounts for one starting position.

*This work was supported in part by National Science Foundation grant CCF-1115206.

Table 1: Notation or numerical values considered for de Bruijn sequences and related problems in the references

Reference		Multiplicity	Alphabet size	Word size
[7]	de Rivière (1894)	1	2	n
[16]	Sainte-Marie (1894)	1	2	n
[4]	de Bruijn (1946)	1	2	n
[19]	van A.-E. & de Bruijn (1951)	1	σ	n
[5]	de Bruijn (1975)	1	σ	n
[3]	Dawson & Good (1957)	k	t	m
[8]	Fredricksen (1982)	1	k	n
[11]	Kandel et al. (1996)	variable	any (uses 4)	k
[17]	Stanley (1999)	1	d	n
[10]	Higgins (2012)	1	k	n
[15]	Osipov (2016)	$f = 2$	$\ell = 2$	$1 \leq p \leq 4$
	Tesler (2016) [this paper]	m	q	k

In 1894, the problem of counting cyclic de Bruijn sequences over a binary alphabet (the case $q = 2$) was proposed by de Rivière [7] and solved by Sainte-Marie [16]. In 1946, the same problem was solved by de Bruijn [4], unaware of Sainte-Marie's work. In 1951, the solution was extended to any size alphabet ($q \geq 1$) by van Aardenne-Ehrenfest and de Bruijn [19, p. 203]:

$$\text{(1)} \qquad \# \text{ cyclic de Bruijn sequences} = q!^{q^{k-1}}/q^k \ .$$

The sequences were subsequently named de Bruijn sequences, and work continued on them for decades before the 1894 publication by Sainte-Marie was rediscovered in 1975 [5]. Table 1 summarizes the cases considered and notation used in these and other papers.

We introduce a *cyclic multi de Bruijn sequence*: a cyclic sequence over alphabet Ω (of size q) in which all k-mers occur exactly m times, with $m, q, k \geq 1$. Let $\mathcal{C}(m, q, k)$ denote the set of all such sequences. The length of such a sequence is $\ell = mq^k$, because each of the q^k k-mers accounts for m starting positions. For $m = 2$, $q = 2$, $k = 3$, $\Omega = \{0, 1\}$, one such sequence is (1111011000101000). The length is $\ell = mq^k = 2 \cdot 2^3 = 16$. Each 3-mer 000, 001, 010, 011, 100, 101, 110, 111 occurs twice, including overlapping occurrences and occurrences that wrap around the end.

Let $\mathcal{LC}(m, q, k)$ denote the set of linearizations of cyclic sequences in $\mathcal{C}(m, q, k)$. These will be called *linearized* or *linearized cyclic multi de Bruijn sequences*. Let $\mathcal{LC}_y(m, q, k)$ denote the set of linearizations that start with k-mer $y \in \Omega^k$. In the example above, the two linearizations starting with 000 are 0001111011000101 and 0001010001111011.

For a sequence s and $i \geq 0$, let s^i denote concatenating i copies of s.

A cyclic sequence (s) (or a linearization s) of length n has a d^{th} order rotation for some $d|n$ (positive integer divisor d of n) iff $\rho^{n/d}(s) = s$ iff $s = t^d$ for some sequence t of length n/d. The *order* of cycle (s) (or linearization s) is the largest $d|n$ such that $\rho^{n/d}(s) = s$. For example, (11001100) has order 2, while if all n rotations of s are distinct, then (s) has order 1. Since a cyclic multi de Bruijn sequence has exactly m copies of each k-mer, the order must divide into m. Sets $\mathcal{C}^{(d)}(m,q,k)$, $\mathcal{LC}^{(d)}(m,q,k)$, and $\mathcal{LC}_y^{(d)}(m,q,k)$ denote multi de Bruijn sequences of order d that are cyclic, linearized cyclic, and linearized cyclic starting with y.

A *linear multi de Bruijn sequence* is a linear sequence of length $\ell' = mq^k + k - 1$ in which all k-mers occur exactly m times. Let $\mathcal{L}(m,q,k)$ denote the set of such sequences and $\mathcal{L}_y(m,q,k)$ denote those starting with k-mer y. A linear sequence does not wrap around and is not the same as a linearization of a cyclic sequence. For $m = 2$, $q = 2$, $k = 3$, an example is 111101100010100011, with length $\ell' = 18$.

For $m, q, k \geq 1$, we will compute the number of linearized cyclic (Sec. 2), cyclic (Sec. 3), and linear (Sec. 4) multi de Bruijn sequences. We also compute the number of cyclic and linearized cyclic sequences with a d^{th} order rotational symmetry. We give examples in Sec. 5. In Sec. 6, we use brute force to generate all multi de Bruijn sequences for small values of m, q, k, confirming our formulas in certain cases. In Sec. 7, we show how to select a uniform random linear, linearized cyclic, or cyclic multi de Bruijn sequence. A summary of our notation and formulas is given in Table 2.

We also consider another generalization: a *multicyclic de Bruijn sequence* is a multiset of aperiodic cyclic sequences such that every k-mer occurs exactly m times among all the cycles. For example, (00011)(011) has two occurrences of each 2-mer $00, 01, 10, 11$. This generalizes results of Higgins [10] for the case $m = 1$. In Sec. 8, we develop this generalization using the "Extended Burrows-Wheeler Transform" of Mantaci et al. [13, 14]. In Sec. 9, we give another method to count multicyclic de Bruijn sequences by counting the number of ways to partition a balanced graph into aperiodic cycles with prescribed edge multiplicities.

We implemented these formulas and algorithms in software available at http://math.ucsd.edu/~gptesler/multidebruijn.

Related work. The methods van Aardenne-Ehrenfest and de Bruijn [19] developed in 1951 to generalize de Bruijn sequences to alphabets of any size potentially could have been adapted to $\mathcal{C}(m,q,k)$; see Sec. 3. In 1957, Dawson and Good [3] counted "circular arrays" in which each k-mer occurs m times (in our notation). Their formula corresponds to an intermediate step in our solution, but overcounts cyclic multi de Bruijn sequences; see Sec. 2.3.

Table 2: Notation and summary of formulas for the number of multi de Bruijn sequences of different types

Definition	Notation
Alphabet	Ω
Alphabet size	$q = \|\Omega\|$
Word size	k
Multiplicity of each k-mer	m
Rotational order of a sequence	d; must divide into m
Specific k-mer that sequences start with	y
Complete de Bruijn multigraph	$G(m, q, k)$
Cyclic shift	$\rho(a_1 \ldots a_n) = a_n a_1 \ldots a_{n-1}$
Power of a linear sequence	$s^r = s \cdots s$ (r times)
Möbius function	$\mu(n)$
Euler's totient function	$\phi(n)$
Set of permutations of $0^m 1^m \cdots (q-1)^m$	$\mathcal{W}_{m,q}$

Multi de Bruijn Sequence	Set	Size
Linear	$\mathcal{L}(m, q, k)$	$W(m, q, k) = ((mq)!/m!^q)^{q^{k-1}}$
Cyclic	$\mathcal{C}(m, q, k)$	$\frac{1}{mq^k} \sum_{r\|m} \phi(m/r) W(r, q, k)$
Linearized cyclic	$\mathcal{LC}(m, q, k)$	$W(m, q, k)$
Multicyclic	$\mathcal{MC}(m, q, k)$	$W(m, q, k)$
Linear, starts with k-mer y	$\mathcal{L}_y(m, q, k)$	$W(m, q, k)/q^k$
Linearized, starts with y	$\mathcal{LC}_y(m, q, k)$	$W(m, q, k)/q^k$
Cyclic, order d	$\mathcal{C}^{(d)}(m, q, k)$	$\frac{1}{(m/d)q^k} \sum_{r\|(m/d)} \mu(r) W(\frac{m}{rd}, q, k)$
Linearized, starts with y, order d	$\mathcal{LC}_y^{(d)}(m, q, k)$	$\frac{1}{q^k} \sum_{r\|(m/d)} \mu(r) W(\frac{m}{rd}, q, k)$

Very recently, in 2016, Osipov [15] introduced the problem of counting "p-ary f-fold de Bruijn sequences." However, Osipov only gives a partial solution, which appears to be incorrect; see Sec. 6.

2. Linearizations of cyclic multi de Bruijn sequences

2.1. Multi de Bruijn graph

We will compute the number of cyclic multi de Bruijn sequences by counting Eulerian cycles in a graph and adjusting for multiple Eulerian cycles corresponding to each cyclic multi de Bruijn sequence. For now, we assume that $k \geq 2$; we will separately consider $k = 1$ in Sec. 2.4. We will also clarify details for the case $q = 1$ in Sec. 2.5.

Define a multigraph $G = G(m, q, k)$ whose vertices are the $(k-1)$-mers over alphabet Ω of size q. There are $n = q^{k-1}$ vertices. For each k-mer $y = c_1 c_2 \ldots c_k \in \Omega^k$, add m directed edges $c_1 c_2 \ldots c_{k-1} \xrightarrow{y} c_2 \ldots c_{k-1} c_k$, each labelled by y. Every vertex has outdegree mq and indegree mq. Further, we will show in Sec. 2.2 that the graph is strongly connected, so G is Eulerian.

Consider a walk through the graph:

$$c_1 c_2 \cdots c_{k-1} \ \rightarrow \ c_2 c_3 \cdots c_k \ \rightarrow \ c_3 c_4 \cdots c_{k+1} \ \rightarrow \ \cdots \ \rightarrow \ c_r c_{r+1} \cdots c_{r+k-2}$$

This walk corresponds to a linear sequence $c_1 c_2 \cdots c_{r+k-2}$, which can be determined either from vertex labels (when $k \geq 2$) or edge labels (when $k \geq 1$). If the walk is a cycle (first vertex equals last vertex), then it also corresponds to a linearization $c_1 c_2 \cdots c_{r-1}$ of cyclic sequence $(c_1 c_2 \cdots c_{r-1})$. Starting at another location in the cycle will result in a different linearization of the same cyclic sequence.

Any Eulerian cycle in this graph gives a cyclic sequence in $\mathcal{C}(m, q, k)$. Conversely, each sequence in $\mathcal{C}(m, q, k)$ corresponds to at least one Eulerian cycle.

Cyclic multi de Bruijn sequences may also be obtained through a generalization of Hamiltonian cycles. Consider cycles in $G(1, q, k+1)$, with repeated edges and vertices allowed, in which every vertex (k-mer) is visited exactly m times (not double-counting the initial vertex when it is used to close the cycle). Each such graph cycle starting at vertex $y \in \Omega^k$ corresponds to the linearization it spells in $\mathcal{LC}_y(m, q, k)$, and vice-versa. However, we will use the Eulerian cycle approach because it leads to an enumeration formula.

2.2. Matrices

Let $n = q^{k-1}$ and form the $n \times n$ adjacency matrix A of directed graph $G = G(m, n, q)$. When $k \geq 2$,

$$A_{c_1 \ldots c_{k-1}, d_1 \ldots d_{k-1}} = \begin{cases} m & \text{if } c_2 \ldots c_{k-1} = d_1 \ldots d_{k-2} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

For every pair of $(k-1)$-mers $v = c_1 \ldots c_{k-1}$ and $w = d_1 \ldots d_{k-1}$, the walks of length $k-1$ from v to w have this form:

$$c_1 \ldots c_{k-1} \ \rightarrow \ c_2 \ldots c_{k-1} d_1 \ \rightarrow \ \cdots \ \rightarrow \ c_{k-1} d_1 \ldots d_{k-2} \ \rightarrow \ d_1 \ldots d_{k-1}$$

For each of the $k-1$ arrows, there are m parallel edges to choose from. Thus, there are m^{k-1} walks of length $k-1$ from v to w, so the graph is strongly connected and $(A^{k-1})_{v,w} = m^{k-1}$ for all pairs of vertices v, w. This gives $A^{k-1} = m^{k-1}J$, where J is the $n \times n$ matrix of all 1's.

J has an eigenvalue n of multiplicity 1 and an eigenvalue 0 of multiplicity $n-1$. So A^{k-1} has an eigenvalue $m^{k-1}n = (mq)^{k-1}$ of multiplicity 1 and an eigenvalue 0 of multiplicity $n-1 = q^{k-1}-1$.

Thus, A has one eigenvalue of the form $mq\omega$, where ω is a $(k-1)^{\text{th}}$ root of unity, and $q^{k-1}-1$ eigenvalues equal to 0. In fact, the first eigenvalue of A is mq, because the all 1's column vector $\vec{1}$ is a right eigenvector with eigenvalue mq (that is, $A\vec{1} = mq\,\vec{1}$): for every vertex $w \in \Omega^{k-1}$, we have

$$(A\vec{1})_w = \sum_{v \in \Omega^{k-1}} (A_{vw})(\vec{1})_w = \sum_{v \in \Omega^{k-1}} A_{vw} \cdot 1 = \textsc{Indeg}(w) = mq\,.$$

Similarly, the all 1's row vector is a left eigenvector with eigenvalue mq.

The degree matrix of an Eulerian graph is an $n \times n$ diagonal matrix with $D_{vv} = \textsc{Indeg}(v) = \textsc{Outdeg}(v)$ on the diagonal and $D_{vw} = 0$ for $v \neq w$. All vertices in G have indegree and outdegree mq, so $D = mqI$. The Laplacian matrix of G is $L = D - A = mqI - A$. It has one eigenvalue equal to 0 and $q^{k-1}-1$ eigenvalues equal to mq.

2.3. Number of Eulerian cycles in the graph

Choose any edge $e = (v, w)$ in G. Let y be the k-mer represented by e. The number of spanning trees of G with root v (with a directed path from each vertex to root v) is given by Tutte's Matrix-Tree Theorem for directed graphs [18, Theorem 3.6] (also see [19, Theorem 7] and [17, Theorem 5.6.4]). For a directed Eulerian graph, the formula can be expressed as follows [17, Cor. 5.6.6]: the number of spanning trees of G rooted at v is

$$(3)\qquad \frac{1}{n} \cdot (\text{product of the } n-1 \text{ nonzero eigenvalues of } L) = \frac{(mq)^{q^{k-1}-1}}{q^{k-1}}\,.$$

By the BEST Theorem [19, Theorem 5b] (also see [17, pp. 56, 68] and [8]), the number of Eulerian cycles with initial edge e is

$$(4)\qquad (\#\text{ spanning trees rooted at } v) \cdot \prod_{x \in V} (\textsc{Outdeg}(x) - 1)!$$

$$(5)\qquad = \frac{1}{q^{k-1}} \cdot (mq)^{q^{k-1}-1} \cdot (mq-1)!^{q^{k-1}} = \frac{1}{m \cdot q^k} \cdot (mq)!^{q^{k-1}}\,.$$

Each Eulerian cycle spells out a linearized multi de Bruijn sequence that starts with k-mer y. However, when $m \geq 2$, there are multiple cycles generating each such sequence. Let C be an Eulerian cycle spelling out linearization s. For k-mer y, the first edge of the cycle, e, was given; the other $m-1$ edges labelled by y are parallel to e and may be permuted in C in any of $(m-1)!$ ways. For the other $n-1$ k-mers, the m edges representing each k-mer can be permuted in C in $m!$ ways. Thus, each linearization s starting in y is generated by $(m-1)! \cdot m!^{n-1} = (m!^n)/m$ Eulerian cycles that start with edge e. Divide Eq. (5) by this factor to obtain the number of linearized multi de Bruijn sequences starting with k-mer y:

$$(6) \qquad |\mathcal{LC}_y(m,q,k)| = \frac{1}{m \cdot q^k} \cdot \frac{(mq)!^{q^{k-1}}}{m!^{q^k}/m} = \frac{1}{q^k} \cdot \frac{(mq)!^{q^{k-1}}}{m!^{q^k}} = \frac{W(m,q,k)}{q^k}$$

where we set

$$(7) \qquad W(m,q,k) = \frac{(mq)!^{q^{k-1}}}{m!^{q^k}} = \left(\frac{(mq)!}{m!^q}\right)^{q^{k-1}} = \binom{mq}{\underbrace{m,\ldots,m}_{q \text{ of these}}}^{q^{k-1}} .$$

Related work. Dawson and Good [3, p. 955] computed the number of "circular arrays" containing each k-mer exactly m times (in our notation; theirs is different), and obtained an answer equivalent to (5). They counted graph cycles in $G(m,q,k)$ starting on a specific edge, but when $m \geq 2$, this does not equal $|\mathcal{C}(m,q,k)|$; their count includes each cyclic multi de Bruijn sequence multiple times. This is because they use the convention that the multiple occurrences of each symbol are distinguishable [3, p. 947]. We do additional steps to count each cyclic multi de Bruijn sequence just once: we adjust for multiple graph cycles corresponding to each linearization (6) and for multiple linearizations of each cyclic sequence (Sec. 3).

2.4. Special case $k = 1$

When $k = 1$, the number of sequences of length $mq^k = mq$ with each symbol in Ω occurring exactly m times is $(mq)!/m!^q$. Divide by q to obtain the number of linearized de Bruijn sequences starting with each 1-mer $y \in \Omega$:

$$|\mathcal{LC}_y(m,q,1)| = \frac{1}{q} \cdot \frac{(mq)!}{m!^q} .$$

This agrees with plugging $k = 1$ into (6), so (6) holds for $k = 1$ as well.

The derivation of (6) in Secs. 2.1–2.3 requires $k \geq 2$, due to technicalities. The above derivation uses a different method, but we can also adapt the first derivation as follows.

$G(m,q,1)$ has a single vertex '$\emptyset$' (a 0-mer) and mq loops on $\emptyset$, because for each $i \in \Omega$, there are m loops $\emptyset \xrightarrow{i} \emptyset$. The sequence spelling out a walk can be determined from edge labels but not vertex labels, since they're null.

The adjacency matrix is $A = [mq]$ rather than Eq. (2). The Laplacian matrix $L = mqI - A = [0]$ has one eigenvalue: 0. In (3), the product of nonzero eigenvalues is vacuous; on plugging in $k = 1$, the formula correctly gives that there is 1 spanning tree (it is the vertex $\emptyset$ without any edges). The remaining steps work for $k = 1$ as-is.

2.5. Special case $q = 1$

For $q = 1$, the only cyclic sequence of length $\ell = mq^k = m \cdot 1^k = m$ is (0^m). We regard cycle (0^m) and linearization 0^m as having an occurrence of 0^k starting at each position, even if $k > m$, in which case some positions of (0^m) are used multiple times to form 0^k. With this convention, there are exactly m occurrences of 0^k in (0^m). Thus, for $q = 1$ and $m, k \geq 1$,

$$\mathcal{LC}(m,1,k) = \{0^m\}, \qquad \mathcal{LC}_{0^k}(m,1,k) = \{0^m\},$$
$$\mathcal{C}(m,1,k) = \{(0^m)\}, \qquad \mathcal{L}(m,1,k) = \{0^{m+k-1}\}.$$

Using positions of a cycle multiple times to form a k-mer will also arise in multicyclic de Bruijn sequences in Sec. 8.

3. Cyclic multi de Bruijn sequences

In a classical de Bruijn sequence ($m = 1$), each k-mer occurs exactly once, so the number of cyclic de Bruijn sequences equals the number of linearizations that begin with any fixed k-mer y:

$$|\mathcal{C}(1,q,k)| = |\mathcal{LC}_y(1,q,k)| = (q!)^{q^{k-1}}/q^k \; .$$

This agrees with (1). But a multi de Bruijn sequence has m cyclic shifts starting with y, and the number of these that are distinct varies by sequence. Let $d|m$ be the order of the cyclic symmetry of the sequence; then there are m/d distinct linearizations starting with y, so

$$|\mathcal{LC}_y^{(d)}(m,q,k)| = \frac{m}{d} \cdot |\mathcal{C}^{(d)}(m,q,k)| \; . \tag{8}$$

Further, we have the following:

Lemma 3.1. *For $m, q, k \geq 1$ and $d|m$:*

(a) $|\mathcal{LC}^{(d)}(m,q,k)| = |\mathcal{LC}^{(1)}(m/d,q,k)|$

(b) For each k-mer $y \in \Omega^k$, $|\mathcal{LC}_y^{(d)}(m,q,k)| = |\mathcal{LC}_y^{(1)}(m/d,q,k)|$.
(c) $|\mathcal{C}^{(d)}(m,q,k)| = |\mathcal{C}^{(1)}(m/d,q,k)|$

Proof. At $q = 1$, all sets in (a)–(c) have size 1. Below, we consider $q \geq 2$.

(a) Let $s \in \mathcal{LC}^{(d)}(m,q,k)$. Since s has order d, it splits into d equal parts, $s = t^d$. The m occurrences of each k-mer in (s) reduce to m/d occurrences of each k-mer in (t), and the order of t is $d/d = 1$, so $t \in \mathcal{LC}^{(1)}(m/d,q,k)$.

Conversely, if $t \in \mathcal{LC}^{(1)}(m/d,q,k)$ then (t^d) has $d \cdot (m/d) = m$ occurrences of each k-mer and has order d, so $(t^d) \in \mathcal{LC}^{(d)}(m,q,k)$.

(b) Continuing with (a), the length of t is at least k (since $|t| = q^k \geq k$ for $q \geq 2$ and $k \geq 1$), so the initial k-mer of s and t must be the same.

(c) The map in (a) induces a bijection $f : \mathcal{C}^{(d)}(m,q,k) \to \mathcal{C}^{(1)}(m/d,q,k)$ as follows. Let $\sigma \in \mathcal{C}^{(d)}(m,q,k)$. For any linearization s of σ, let $s = t^d$ and set $f(\sigma) = (t)$. This is well-defined: the distinct linearizations of σ are $\rho^i(s)$ where $i = 0, \ldots, (mq^k/d) - 1$. Rotating s by i also rotates t by i and gives the same cycle in $\mathcal{C}^{(1)}(m/d,q,k)$, so all linearizations of σ give the same result.

Conversely, given $(t) \in \mathcal{C}^{(1)}(m/d,q,k)$, the unique inverse is $f^{-1}((t)) = (t^d)$. As above, this is well-defined. □

Partitioning the cyclic multi de Bruijn sequences by order gives the following, where the sums run over positive integers d that divide into m:

$$|\mathcal{C}(m,q,k)| = \sum_{d|m} |\mathcal{C}^{(d)}(m,q,k)| = \sum_{d|m} |\mathcal{C}^{(1)}(m/d,q,k)| \,. \tag{9}$$

Each cyclic sequence of order d has m/d distinct linearizations starting with y, so

$$\begin{aligned}|\mathcal{LC}_y(m,q,k)| &= \sum_{d|m} |\mathcal{LC}_y^{(d)}(m,q,k)| = \sum_{d|m} \frac{m}{d} \cdot |\mathcal{C}^{(d)}(m,q,k)| \\ &= \sum_{d|m} \frac{m}{d} \cdot |\mathcal{C}^{(1)}(m/d,q,k)| = \sum_{d|m} d \cdot |\mathcal{C}^{(1)}(d,q,k)| \,.\end{aligned}$$

By Möbius Inversion,

$$m \cdot |\mathcal{C}^{(1)}(m,q,k)| = \sum_{d|m} \mu(d) \cdot |\mathcal{LC}_y(m/d,q,k)| \,.$$

Solve this for $|\mathcal{C}^{(1)}(m,q,k)|$:

$$|\mathcal{C}^{(1)}(m,q,k)| = \frac{1}{m} \sum_{d|m} \mu(d) \cdot |\mathcal{LC}_y(m/d,q,k)| \,. \tag{10}$$

Combining this with (6) and Lemma 3.1(c) gives

$$|\mathcal{C}^{(d)}(m,q,k)| = \frac{1}{(m/d)q^k} \sum_{r|(m/d)} \mu(r) W(m/(rd),q,k) \ . \tag{11}$$

Plug (10) into (9) to obtain

$$\begin{aligned}
|\mathcal{C}(m,q,k)| &= \sum_{d|m} |\mathcal{C}^{(1)}(m/d,q,k)| \\
&= \sum_{d|m} \frac{1}{m/d} \sum_{d'|(m/d)} \mu(d') \cdot |\mathcal{LC}_y(m/(dd'),q,k)| \ .
\end{aligned}$$

Change variables: set $r = dd'$

$$\begin{aligned}
&= \sum_{r|m} \sum_{d|r} \frac{1}{m/d} \cdot \mu(r/d) \cdot |\mathcal{LC}_y(m/r,q,k)| \\
&= \frac{1}{m} \sum_{r|m} \left(\sum_{d|r} d \cdot \mu(r/d) \right) \cdot |\mathcal{LC}_y(m/r,q,k)| \\
&= \frac{1}{m} \sum_{r|m} \phi(r) \cdot |\mathcal{LC}_y(m/r,q,k)| = \frac{1}{m} \sum_{r|m} \phi(m/r) \cdot |\mathcal{LC}_y(r,q,k)| \ ,
\end{aligned}$$

where $\phi(r)$ is the Euler totient function. Thus,

$$|\mathcal{C}(m,q,k)| = \frac{1}{mq^k} \sum_{r|m} \phi(m/r) \cdot W(r,q,k) \ . \tag{12}$$

Related work. Van Aardenne-Ehrenfest and de Bruijn [19, Sec. 5] take a directed Eulerian multigraph G and replace each edge by a "bundle" of λ parallel edges to obtain a multigraph G^λ. They compute the number of Eulerian cycles in G^λ in terms of the number of Eulerian cycles in G, and obtain a formula related to (12). Their bundle method could have been applied to count cyclic multi de Bruijn sequences in terms of ordinary cyclic de Bruijn sequences by taking $G = G(1,q,k)$, $\lambda = m$, and considering G^m, but they did not do this. Instead, they set $\lambda = q$ and found a correspondence between Eulerian cycles in $G(1,q,k+1)$ vs. G^q with $G = G(1,q,k)$, yielding a recursion in k for $|\mathcal{C}(1,q,k)|$. They derived (1) by this recursion rather than the BEST Theorem, which was also introduced in that paper. Dawson and Good [3, p. 955] subsequently derived (1) by the BEST Theorem.

4. Linear multi de Bruijn sequences

Now we consider $\mathcal{L}(m,q,k)$, the set of *linear sequences* in which every k-mer over Ω occurs exactly m times. A linear sequence is not the same as a linearized representation of a circular sequence. Recall that a linearized sequence has length $\ell = mq^k$. A linear sequence has length $\ell' = mq^k + k - 1$, because there are mq^k positions at which k-mers start, and an additional $k-1$ positions at the end to complete the final k-mers. Below, assume $q \geq 2$.

Lemma 4.1. *For $q \geq 2$, for each sequence in $\mathcal{L}(m,q,k)$, the first and last $(k-1)$-mer are the same.*

Proof. Count the number of times each $(k-1)$-mer $x \in \Omega^{k-1}$ occurs in $s' = a_1 a_2 \ldots a_{\ell'} \in \mathcal{L}(m,q,k)$ as follows:

Each k-mer occurs m times in s' among starting positions $1, \ldots, mq^k$. Since x is the first $k-1$ letters of q different k-mers, there are mq occurrences of x starting in this range. There is also a $(k-1)$-mer x_1 with an additional occurrence at starting position $mq^k + 1$ (the last $k-1$ positions of s'), so x_1 occurs $mq + 1$ times in s' while all other $(k-1)$-mers occur mq times.

By a similar argument, the $(k-1)$-mer x_2 at the start of s' has a total of $mq+1$ occurrences in s', and all other $(k-1)$-mers occur exactly mq times.

So x_1 and x_2 each occur $mq + 1$ times in s'. But we showed that there is only one $(k-1)$-mer occurring $mq + 1$ times, so $x_1 = x_2$. □

For $q \geq 2$, this leads to a bijection $f : \mathcal{L}(m,q,k) \to \mathcal{LC}(m,q,k)$. Given $s' \in \mathcal{L}(m,q,k)$, drop the last $k-1$ characters to obtain sequence $s = f(s')$. To invert, given $s \in \mathcal{LC}(m,q,k)$, form $s' = f^{-1}(s)$ by repeating the first $k-1$ characters of s at the end.

For $i \in \{1, \ldots, \ell\}$, the same k-mer occurs linearly at position i of s' and circularly at position i of s, so every k-mer occurs the same number of times in both linear sequence s' and cycle (s). Thus, f maps $\mathcal{L}(m,q,k)$ to $\mathcal{LC}(m,q,k)$, and f^{-1} does the reverse.

This also gives $|\mathcal{L}_y(m,q,k)| = |\mathcal{LC}_y(m,q,k)|$. Multiply (6) by q^k choices of initial k-mer y to obtain

$$|\mathcal{L}(m,q,k)| = |\mathcal{LC}(m,q,k)| = q^k \cdot |\mathcal{LC}_y(m,q,k)| = W(m,q,k)\ . \tag{13}$$

Above, we assumed $q \geq 2$. Eq. (13) also holds for $q = 1$, since all four parts equal 1 (see Sec. 2.5).

5. Examples

The case $m = 1$, arbitrary $q, k \geq 1$, is the classical de Bruijn sequence. Eq. (12) has just one term, $r = 1$, and agrees with (1):

$$|\mathcal{C}(1,q,k)| = \frac{1}{q^k} \cdot \phi(1) \cdot \binom{q}{1,\dots,1}^{q^{k-1}} = \frac{1}{q^k} \cdot 1 \cdot q!^{q^{k-1}} = \frac{q!^{q^{k-1}}}{q^k} \, .$$

When $m = p$ is prime, the sum in (12) for the number of cyclic multi de Bruijn sequences has two terms (divisors $r = 1, p$):

$$\begin{aligned} |\mathcal{C}(p,q,k)| &= \frac{1}{pq^k}\Big(\phi(1) \cdot W(p,q,k) + \phi(p) \cdot W(1,q,k)\Big) \\ &= \frac{1}{pq^k}\left(\left(\frac{(pq)!}{p!^q}\right)^{q^{k-1}} + (p-1) \cdot (q!)^{q^{k-1}}\right) . \end{aligned} \tag{14}$$

For $(m,q,k) = (2,2,2)$, Table 3 lists all multi de Bruijn sequences of each type. Table 3(A) shows 9 linearizations beginning with $y = 00$. By (6),

$$|\mathcal{LC}_{00}(2,2,2)| = \frac{1}{2^2}\binom{2 \cdot 2}{2,2}^{2^{2-1}} = \frac{1}{4}\binom{4}{2,2}^2 = \frac{1}{4} \cdot 6^2 = 9 \, .$$

Some of these are equivalent upon cyclic rotations, yielding 5 distinct cyclic multi de Bruijn sequences, shown in Table 3(B). By (12) or (14),

$$\begin{aligned} |\mathcal{C}(2,2,2)| &= \frac{1}{2 \cdot 2^2} \sum_{r|2} \phi(2/r) \binom{2r}{r,r}^{2^1} = \frac{1}{8}\left(\phi(1)\binom{2 \cdot 2}{2,2}^2 + \phi(2)\binom{2 \cdot 1}{1,1}^2\right) \\ &= \frac{1}{8}\left(1 \cdot \binom{4}{2,2}^2 + 1 \cdot \binom{2}{1,1}^2\right) = \frac{1}{8}\left(6^2 + 2^2\right) = \frac{40}{8} = 5 \, . \end{aligned}$$

There are 9 linearizations starting with each 2-mer 00, 01, 10, 11. Converting these to linear multi de Bruijn sequences yields 36 linear multi de Bruijn sequences in total, shown in Table 3(C). By (13),

$$|\mathcal{L}(2,2,2)| = \binom{2 \cdot 2}{2,2}^{2^1} = \binom{4}{2,2}^2 = 6^2 = 36.$$

Table 3(D) shows the 36 multicyclic de Bruijn sequences; see Sec. 8.

Table 3: Multi de Bruijn sequences of each type, with $m = 2$ copies of each k-mer, alphabet $\Omega = \{0, 1\}$ of size $q = 2$, and word size $k = 2$

(A) $\mathcal{LC}_{00}(m = 2, q = 2, k = 2)$: Linearizations starting with 00

Order 1:	00010111	00011011	00011101	00100111
	00101110	00110110	00111010	00111001
Order 2:	00110011			

(B) $\mathcal{C}(m = 2, q = 2, k = 2)$: Cyclic multi de Bruijn sequences

Order 1:	(00010111)	(00011011)	(00011101)	(00100111)
Order 2:	(00110011)			

(C) $\mathcal{L}(m = 2, q = 2, k = 2)$: Linear multi de Bruijn sequences

000101110	010001110	100010111	110001011
000110110	010011100	100011011	110001101
000111010	010111000	100011101	110010011
001001110	011000110	100100111	110011001
001011100	011001100	100110011	110100011
001100110	011011000	100111001	110110001
001101100	011100010	101000111	111000101
001110010	011100100	101100011	111001001
001110100	011101000	101110001	111010001

(D) $\mathcal{MC}(m = 2, q = 2, k = 2)$: Multicyclic de Bruijn sequences

(0)(0)(01)(01)(1)(1)	(0)(001011)(1)	(0001)(01)(1)(1)	(00011)(01)(1)
(0)(0)(01)(011)(1)	(0)(0010111)	(0001)(011)(1)	(000111)(01)
(0)(0)(01)(0111)	(0)(0011)(011)	(0001)(0111)	(00011011)
(0)(0)(01011)(1)	(0)(00101)(1)(1)	(0001011)(1)	(001)(001)(1)(1)
(0)(0)(010111)	(0)(001101)(1)	(00010111)	(001)(0011)(1)
(0)(0)(011)(011)	(0)(0011101)	(00011)(011)	(001)(00111)
(0)(001)(01)(1)(1)	(0)(0011)(01)(1)	(000101)(1)(1)	(0010011)(1)
(0)(001)(011)(1)	(0)(00111)(01)	(0001101)(1)	(00100111)
(0)(001)(0111)	(0)(0011011)	(00011101)	(0011)(0011)

6. Generating multi de Bruijn sequences by brute force

Osipov [15] recently defined a "p-ary f-fold de Bruijn sequence"; this is the same as a cyclic multi de Bruijn sequence but in different terminology and notation. The description below is in terms of our notation m, q, k, which respectively correspond to Osipov's f, ℓ, p (see Table 2).

Osipov developed a method [15, Prop. 3] to compute the number of cyclic multi de Bruijn sequences for multiplicity $m = 2$, alphabet size $q = 2$, and word size $k \geq 1$. This method requires performing a detailed construction for each k. The paper states the two sequences for $k = 1$, which agrees with our $\mathcal{C}(2,2,1) = \{(0011),(0101)\}$ of size 2, and then carries out this construction to evaluate cases $k = 2,3,4$. It determines complete answers for those specific cases only [15, pp. 158–161]. For $k = 2$, Osipov's answer agrees with ours, $|\mathcal{C}(2,2,2)| = 5$. But for $k = 3$ and $k = 4$ respectively, Osipov's method gives 72 and 43768, whereas by (12) or (14), we obtain $|\mathcal{C}(2,2,3)| = 82$ and $|\mathcal{C}(2,2,4)| = 52496$.

For these parameters, it is feasible to find all cyclic multi de Bruijn sequences by a brute force search in Ω^ℓ (with $\ell = mq^k$). This agrees with our formula (12), while disagreeing with Osipov's results for $k = 3$ and 4. Table 4 lists the 82 sequences in $\mathcal{C}(2,2,3)$. The 52496 sequences for $\mathcal{C}(2,2,4)$ and an implementation of the algorithm below are on the website listed in the introduction. We also used brute force to confirm (12) for all combinations of parameters $m,q,k \geq 2$ with at most a 32-bit search space ($q^\ell \leq 2^{32}$):

$$\begin{array}{lll}
|\mathcal{C}(2,2,2)| = 5 & |\mathcal{C}(6,2,2)| = 35594 & |\mathcal{C}(2,2,3)| = 82 \\
|\mathcal{C}(3,2,2)| = 34 & |\mathcal{C}(7,2,2)| = 420666 & |\mathcal{C}(3,2,3)| = 6668 \\
|\mathcal{C}(4,2,2)| = 309 & |\mathcal{C}(8,2,2)| = 5176309 & |\mathcal{C}(4,2,3)| = 750354 \\
|\mathcal{C}(5,2,2)| = 3176 & |\mathcal{C}(2,3,2)| = 40512 & |\mathcal{C}(2,2,4)| = 52496.
\end{array}$$

Represent Ω^ℓ by ℓ-digit base q integers (range $N \in [0, q^\ell - 1]$):

$$N = \sum_{j=0}^{\ell-1} a_j q^j \quad (0 \leq a_j < q) \quad \text{represents } s = a_{\ell-1}a_{\ell-2}\ldots a_0\,.$$

Since we represent each element of $\mathcal{C}(m,q,k)$ by a linearization that starts with k-mer 0^k, we restrict to $N \in [0, q^{\ell-k+1}-1]$. To generate $\mathcal{C}(m,q,k)$:

```
(B1) for $N$ from 0 to $q^{\ell-k+1} - 1$ (pruning as described later):
(B2)     let $s = a_{\ell-1}a_{\ell-2}\ldots a_0$ be the $\ell$-digit base $q$ expansion of $N$
(B3)     if (cycle $(s)$ is a cyclic multi de Bruijn sequence
(B4)         and $s$ is smallest among its $\ell$ cyclic shifts)
(B5)     then output cycle $(s)$
```

To generate $\mathcal{LC}_{0^k}(m,q,k)$, skip (B4) and output linearizations s in (B5).

To determine if (s) is a cyclic multi de Bruijn sequence, examine its k-mers $a_{i+k-1}a_{i+k-2}\ldots a_i$ (subscripts taken modulo ℓ) in the order $i = \ell-k, \ell-k-1, \ldots, 1-k$, counting how many times each k-mer is encountered. Terminate the test early (and potentially prune as described below) if any

Table 4: $\mathcal{C}(2,2,3)$: For multiplicity $m = 2$, alphabet size $q = 2$, and word size $k = 3$, there are 82 cyclic multi de Bruijn sequences. We show the lexicographically least rotation of each

(0000100101101111), (0000100101110111), (0000100101111011),
(0000100110101111), (0000100110111101), (0000100111010111),
(0000100111011101), (0000100111101011), (0000100111101101),
(0000101001101111), (0000101001110111), (0000101001111011),
(0000101011001111), (0000101011100111), (0000101011110011),
(0000101100101111), (0000101100111101), (0000101101001111),
(0000101101111001), (0000101110010111), (0000101110011101),
(0000101110100111), (0000101110111001), (0000101111001011),
(0000101111001101), (0000101111010011), (0000101111011001),
(0000110010101111), (0000110010111101), (0000110011110101),
(0000110100101111), (0000110100111101), (0000110101001111),
(0000110101111001), (0000110111100101), (0000110111101001),
(0000111001010111), (0000111001011101), (0000111001110101),
(0000111010010111), (0000111010011101), (0000111010100111),
(0000111010111001), (0000111011100101), (0000111011101001),
(0000111100101011), (0000111100101101), (0000111100110101),
(0000111101001011), (0000111101001101), (0000111101010011),
(0000111101011001), (0000111101100101), (0000111101101001),
(0001000101101111), (0001000101110111), (0001000101111011),
(0001000110101111), (0001000110111101), (0001000111010111),
(0001000111011101), (0001000111101011), (0001000111101101),
(0001010001101111), (0001010001110111), (0001010001111011),
(0001010110001111), (0001010111000111), (0001010111100011),
(0001011000101111), (0001011000111101), (0001011010001111),
(0001011100010111), (0001011100011101), (0001011101000111),
(0001011110001101), (0001011110100011), (0001100011110101),
(0001101000111101), (0001101010001111), (0001110001110101),
(0001110100011101)

k-mer count exceeds m. If the test does not terminate early, then all k-mer counts must equal m, and it is a multi de Bruijn sequence.

If $i > 0$ when k-mer counting terminates, then some k-mer count exceeded m without examining $a_{i-1} \ldots a_0$. Advance N to $\left(\lfloor N/q^i \rfloor + 1\right) q^i$ instead of $N+1$, skipping all remaining sequences starting with the same $\ell - i$ characters. This is equivalent to incrementing the $\ell - i$ digit prefix $a_{\ell-1} \ldots a_i$ as a base q number and then concatenating i 0's to the end. However, if $i \leq 0$, then all characters have been examined, so advance N to $N+1$.

We may do additional pruning by counting k'-mers for $k' \in \{1, \ldots, k-1\}$. Each k'-mer x is a prefix of $q^{k-k'}$ k-mers and thus should occur $mq^{k-k'}$ times in (s). If a k'-mer has more than $mq^{k-k'}$ occurrences in $a_{\ell-1}\cdots a_i$ with $i > 0$, advance N as described above.

We may also restrict the search space to length ℓ sequences with exactly mq^{k-1} occurrences of each of the q characters. We did not need to implement this for the parameters considered.

7. Generating a uniform random multi de Bruijn sequence

We present algorithms to select a uniform random multi de Bruijn sequence in the linear, linearized, and cyclic cases.

Kandel et al. [11] present two algorithms to generate random sequences with a specified number of occurrences of each k-mer: (i) a "shuffling" algorithm that permutes the characters of a sequence in a manner that preserves the number of times each k-mer occurs, and (ii) a method to choose a uniform random Eulerian cycle in an Eulerian graph and spell out a cyclic sequence from the cycle. Both methods may be applied to select linear or linearized multi de Bruijn sequences uniformly. However, additional steps are required to choose a uniform random cyclic de Bruijn sequence; see Sec. 7.2.

Example 7.1. For $(m, q, k) = (2, 2, 2)$, Table 3(A) lists the nine linearizations starting with 00. The random Eulerian cycle algorithm selects each with probability $1/9$. Table 3(B) lists the five cyclic de Bruijn sequences. The four cyclic sequences of order 1 are each selected with probability $2/9$ since they have two linearizations starting with 00, while (00110011) has one such linearization and thus is selected with probability $1/9$. Thus, the cyclic sequences are not chosen uniformly (which is probability $1/5$ for each).

The shuffling algorithm of Kandel et al. performs a large number of random swaps of intervals in a sequence and approximates the probabilities listed above. Below, we focus on the random Eulerian cycle algorithm.

7.1. Random linear(ized) multi de Bruijn sequences

Kandel et al. [11] present a general algorithm for finding random cycles in a directed Eulerian graph, and apply it to the de Bruijn graph of a string: the vertices are the $(k-1)$-mers of the string and the edges are the k-mers repeated with their multiplicities in the string. As a special case, the de Bruijn graph of any cyclic sequence in $\mathcal{C}(m, q, k)$ is the same as our $G(m, q, k)$. We explain their algorithm and how it specializes to $G(m, q, k)$.

Let G be a directed Eulerian multigraph and $e = (v, w)$ be an edge in G. The proof of the BEST Theorem [19, Theorem 5b] gives a bijection between

(a) Directed Eulerian cycles of G whose first edge is e, and

(b) ordered pairs (T, f) consisting of

- a spanning tree T of G rooted at v, and
- a function f assigning each vertex x of G a permutation of its outgoing edges except e (for $x = v$) or the unique outgoing edge of x in T (for $x \neq v$).

In Eq. (4), the left side counts (a) while the right side counts (b). Note that for each T, there are $\prod_{x \in V}(\textsc{Outdeg}(x) - 1)!$ choices of f.

We will use the BEST Theorem map from (b) to (a). Given (T, f) and $e = (v, w)$ as above, construct an Eulerian cycle C with successive vertices $v_0, v_1, \ldots, v_\ell$ and successive edges $e_1, e_2, \ldots, e_\ell$ as follows:

- Set $e_1 = e$, $v_0 = v$, $v_1 = w$, and $i = 1$.
- While v_i has at least one outgoing edge not yet used in $e_1, \ldots, e_i$:
  - Let j be the number of times vertex v_i appears among $v_0, \ldots, v_i$.
  - If $j < \textsc{Outdeg}(v_i)$, set e_{i+1} to the j^{th} edge given by $f(v_i)$. If $j = \textsc{Outdeg}(v_i)$, set e_{i+1} to the outgoing edge of v_i in T.
  - Let $v_{i+1} = \textsc{Head}(e_{i+1})$.
  - Increment i.

This reduces the problem of selecting a uniform random Eulerian cycle to selecting T uniformly and f uniformly. Kandel et al. [11] give the following algorithm to select a uniform random spanning tree T with root v in G.

Initialize $T = \{v\}$. Form a random walk in G starting at $v_0 = v$ and going backwards on edges (using incoming edges instead of outgoing edges):

$$v_0 \xleftarrow{e_0} v_1 \xleftarrow{e_1} v_2 \xleftarrow{e_2} v_3 \ \cdots$$

At each vertex v_i, choose edge e_i uniformly from its incoming edges, and set $v_{i+1} = \textsc{Tail}(e_i)$. (Kandel et al. gives the option of ignoring loops in selecting an incoming edge, but we allow them.) If this is the first encounter with vertex v_{i+1} on the walk, add vertex v_{i+1} and edge e_i to T. Repeat until all vertices of G have been visited at least once. Kandel et al. show that T is selected uniformly from spanning trees of G directed towards root v.

We apply this to random spanning trees in $G(m, q, k)$. Let $v \in \Omega^{k-1}$. Write $v = b_{k-2}b_{k-3} \cdots b_1 b_0$. Select $b_i \in \Omega$ (for $i \geq k-1$) uniformly and form a random walk with $v_i = b_{i+k-2} \cdots b_i$ and e_i labelled by $b_{i+k-1} \cdots b_i$:

$$b_{k-2} \cdots b_1 b_0 \xleftarrow{b_{k-1} \cdots b_0} b_{k-1} \cdots b_2 b_1 \xleftarrow{b_k \cdots b_1} b_k \cdots b_3 b_2 \xleftarrow{b_{k+1} \cdots b_2} \cdots$$

For each $a \in \Omega$, the fraction of incoming edges of v_i with form $ab_{i+k-2} \cdots b_i$ is $m/(mq) = 1/q$, so b_{i+k-1} is selected uniformly from Ω. The first time vertex $v_{i+1} = b_{i+k-1} \dots b_{i+1}$ is encountered, add edge $e_i = b_{i+k-1} \dots b_{i+1} b_i$ to T. Terminate as soon as all vertices have been encountered at least once.

Another way to generate a spanning tree T of $G(m, q, k)$ is to generate a spanning tree T' of $G(1, q, k)$ by the above algorithm, and then, for each edge in T', uniformly choose one of the corresponding m edges in $G(m, q, k)$. The probability of each spanning tree of $G(m, q, k)$ is the same as by the previous algorithm. A similar observation is made in [19, Eq. (6.2)]. Since we do not need to distinguish between the m edges of $G(m, q, k)$ with the same k-mer, it suffices to find a spanning tree of $G(1, q, k)$ rather than $G(m, q, k)$.

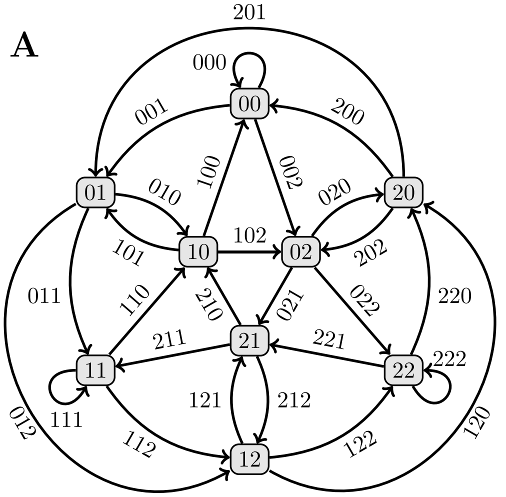
Example 7.2. Fig. 1(A–C) illustrates finding a random spanning tree of $G(1, 3, 3)$ with root $v = 02$. Graph $G(1, 3, 3)$ is shown in Fig. 1(A). A fair 3-sided die gives rolls $0, 0, 2, 0, 1, 2, 2, 1, 0, 1, 1, \dots$, shown in the first column of Fig. 1(B), leading to a random walk with edges and vertices indicated in the next columns. On the first visit to each vertex (besides the root), its outgoing edge is boldfaced in panel (B) and added to the tree in panel (C).

Let $\mathcal{W}_{m,q}$ denote the set of all permutations of $0^m 1^m \dots (q-1)^m$; that is, sequences of length mq with each symbol $0, \dots, q-1$ appearing m times.

For the graph $G(m, q, k)$, we will replace (T, f) in the BEST Theorem bijection by a function $g : \Omega^{k-1} \to \mathcal{W}_{m,q}$. As we traverse a cycle C starting on edge e, the successive visits to vertex $x \in \Omega^{k-1}$ have outgoing edges with k-mers $xc_1, xc_2, \dots, xc_{mq}$ (with $c_i \in \Omega$). Set $g(x) = c_1 c_2 \dots c_{mq}$. For each $b \in \Omega$, there are m edges xb, so exactly m of the c_i's equal b. Thus, $g(x) \in \mathcal{W}_{m,q}$. This encoding does not distinguish the m edges xb since permuting their order in the cycle does not change the sequence the cycle spells. The first edge $e = (v, w)$ is encoded in the first position of $g(v)$; the outgoing edge of x in T is encoded in the last position of $g(x)$ for $x \neq v$; and f is encoded in the other $mq - 1$ positions.

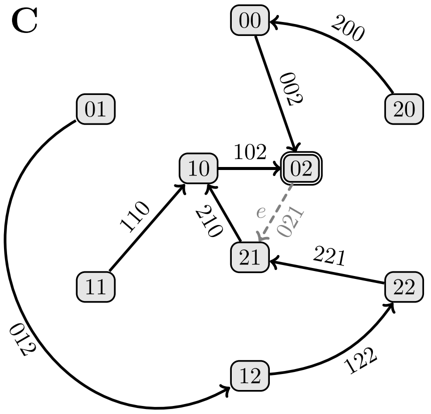
To pick a uniform random element of $\mathcal{LC}_y(m, q, k)$ or $\mathcal{L}(m, q, k)$:

- Input: m, q, k (and y for the linearized case).
- In the linear case, select an initial k-mer y uniformly from Ω^k.
- Decompose $y = va$ ($v \in \Omega^{k-1}$, $a \in \Omega$).
- Select a uniform random spanning tree T of $G(1, q, k)$ with root v.
- Each vertex $x \in \Omega^{k-1} \setminus \{v\}$ has a unique outgoing edge in T, say xb with $b \in \Omega$. Set $g(x)$ to a random element of $\mathcal{W}_{m,q}$ ending in b.
- Set $g(v)$ to a random element of $\mathcal{W}_{m,q}$ starting in a.
- Traverse the cycle starting at v, following the edge order encoded in g, to form the sequence represented by this cycle.
- In the linearized case, delete the last $k - 1$ characters of the sequence.



B

Roll	Edge	Vertex	First visit?
n/a	n/a	02	Yes
0	**002**	00	Yes
0	000	00	No
2	**200**	20	Yes
0	020	02	No
1	**102**	10	Yes
2	**210**	21	Yes
2	**221**	22	Yes
1	**122**	12	Yes
0	**012**	01	Yes
1	101	10	No
1	**110**	11	Yes



D

Vertex	Exit order
00	21010**2**
01	12010**2**
02	**1**20120
10	12100**2**
11	22101**0**
12	00121**2**
20	21210**0**
21	20211**0**
22	12020**1**

E

Linear, s' 02120221011201212220201021121110111002220010001221020002
Linearized, s 021202210112012122202010211211101110022200100012210200

Figure 1: Selecting a uniform random linear(ized) multi de Bruijn sequence with multiplicity $m = 2$, alphabet $\Omega = \{0, 1, 2\}$ of size $q = 3$, word size $k = 3$, initial k-mer 021. (A) De Bruijn graph $G(1, q, k)$. Edges are $k = 3$-mers, vertices are $(k - 1) = 2$-mers. (B) Random process generates edges of a spanning tree with root $v = 02$. (C) Spanning tree (solid edges) with root $v = 02$ (double border), plus selected initial edge 021 (dashed). (D) Order to use outgoing edges at each vertex. Each of the q symbols occurs $m = 2$ times. Boldface in the last column encodes the tree; boldface in the first column encodes initial edge; others are random. (E) Linear(ized) multi de Bruijn sequences starting at $v = 02$ and following edge orders (D).

Example 7.3. Fig. 1 illustrates this algorithm for $(m, q, k) = (2, 3, 3)$. For a random element of $\mathcal{L}(m, q, k)$, the initial k-mer is selected uniformly from Ω^k; say $y = 021$. For a random element of $\mathcal{LC}_y(m, q, k)$, the k-mer y is specified; we specify $y = 021$. Panel (A) shows the de Bruijn graph $G(1, 3, 3)$. Panels (B–C) select a random tree T as described in Example 7.2. Since $y = 021$ (dashed edge in (C)), the root vertex is $v = 02$; the 1 in the last position of y is not used in selecting T.

Panel (D) shows the order of exits from each vertex: the function $g(x) = c_1 \ldots c_{mq}$ gives that the i^{th} exit from x is on edge xc_i. Sequence $c_1 \ldots c_{mq}$ is a permutation of 001122, with a constraint (boldfaced in (D)) on either the first or last position and a random permutation on the other positions. Since the linearization starts with 021, the first exit from vertex $v = 02$ is 1. Tree T is encoded in (D) as the last exit from each vertex except v; e.g., since the tree has edge 002, the last exit from vertex 00 is 2.

To generate a sequence from (D), start at $v = 02$ with first exit 1 (edge 021). At vertex 21, the first exit is 2 (edge 212). At vertex 12, the first exit is 0 (edge 120). At vertex 20, the first exit is 2 (edge 202). Now we visit vertex 02 a second time. Its second exit is 2 (edge 022). The sequence so far is 0212022. Continue until reaching a vertex with no remaining exits; this happens on the $mq+1 = 7^{\text{th}}$ visit to initial vertex $v = 02$, and yields a linear sequence (E). The final $(k-1)$-mer, 02, duplicates the initial $(k-1)$-mer. Remove it from the end to obtain a linearization, also in (E).

7.2. Random cyclic multi de Bruijn sequence

Consider generating random cyclic sequences by choosing a uniform random element of $\mathcal{L}_y(m, q, k)$ and circularizing it. Each element of $\mathcal{C}^{(d)}(m, q, k)$ is represented by m/d linearizations and thus is generated with probability $(m/d)/|\mathcal{LC}_y(m, q, k)|$, which depends on d. So this procedure does not select elements of $\mathcal{C}(m, q, k)$ uniformly. We will show how to adjust for d.

Partition $\mathcal{LC}_y(m, q, k)$ by rotation order, d:

$$\mathcal{LC}_y(m, q, k) = \bigcup_{d|m} \mathcal{LC}_y^{(d)}(m, q, k) \ .$$

We now construct certain subsets of this. For each divisor r of m, define

$$\mathcal{LC}_y(m, r; q, k) = \{t^{m/r} : t \in \mathcal{LC}_y(r, q, k)\} \ . \tag{15}$$

This set has size $|\mathcal{LC}_y(m, r; q, k)| = |\mathcal{LC}_y(r, q, k)|$ and this decomposition:

Lemma 7.4.

$$\mathcal{LC}_y(m,r;q,k) = \bigcup_{d: d|m \textit{ and } (m/r)|d} \mathcal{LC}_y^{(d)}(m,q,k)\ . \tag{16}$$

Proof. Given $t \in \mathcal{LC}_y(r,q,k)$ of order d', then $t^{m/r}$ is in $\mathcal{LC}_y(m,q,k)$ and has order $d = (m/r)d'$. Thus, the left side of (16) is a subset of the right side. Conversely, consider $s \in \mathcal{LC}_y^{(d)}(m,q,k)$ and suppose $(m/r)|d$. Then s splits into m/r equal segments: $s = t^{m/r}$. Each t is in $\mathcal{LC}_y(r,q,k)$ and has order $d' = d/(m/r)$. So the right side of (16) is a subset of the left side. □

Now consider the following random process. Fix $y \in \Omega^k$; it is sufficient to use $y = 0^k$. Let $p(r)$ be a probability distribution on the positive integer divisors of m. Generate a random element σ of $\mathcal{C}(m,q,k)$ as follows:

- Select a random divisor r of m with probability $p(r)$.
- Select a uniform random sequence t from $\mathcal{LC}_y(r,q,k)$.
- Output $\sigma = (t^{m/r})$.

Consider all ways each $\sigma \in \mathcal{C}(m,q,k)$ can be generated by this algorithm. Let d be the order of σ. Each $r|m$ is selected with probability $p(r)$. If $(m/r)|d$, then by Lemma 7.4, the m/d linearizations of σ beginning with y are contained in $\mathcal{LC}_y^{(d)}(m,r;q,k)$, and one of them may be generated as $t^{m/r}$ with probability $(m/d)/|\mathcal{LC}_y(r,q,k)|$. But if m/r does not divide into d, these linearizations will not be generated. In total,

$$P(\sigma) = \sum_{\substack{r:\ r|m \\ \text{and } (m/d)|r}} p(r) \cdot \frac{m/d}{|\mathcal{LC}_y(r,q,k)|}. \tag{17}$$

The following gives $P(\sigma) = 1/|\mathcal{C}(m,q,k)|$ for all $\sigma \in \mathcal{C}(m,q,k)$. For $r|m$, set

$$p(r) = \frac{\phi(m/r)}{m} \cdot \frac{|\mathcal{LC}_y(r,q,k)|}{|\mathcal{C}(m,q,k)|} = \frac{\phi(m/r)\cdot (rq)!^{q^{k-1}}/r!^{q^k}}{\sum_{d|m} \phi(m/d)\cdot (dq)!^{q^{k-1}}/d!^{q^k}} \tag{18}$$

We used (6) and (12) to evaluate this. To verify that σ is selected uniformly, plug the middle expression in (18) into (17), with d equal to the order of σ:

$$P(\sigma) = \sum_{\substack{r:\ r|m \\ \text{and } (m/r)|d}} \frac{\phi(m/r)}{m} \cdot \frac{|\mathcal{LC}_y(r,q,k)|}{|\mathcal{C}(m,q,k)|} \cdot \frac{m/d}{|\mathcal{LC}_y(r,q,k)|} \tag{19}$$

$$= \frac{1}{d \cdot |\mathcal{C}(m,q,k)|} \sum_{\substack{r\colon r|m \\ \text{and } (m/r)|d}} \phi(m/r) \tag{20}$$

$$= \frac{1}{d \cdot |\mathcal{C}(m,q,k)|} \cdot d = \frac{1}{|\mathcal{C}(m,q,k)|}. \tag{21}$$

To evaluate the summation in (20), substitute $u = m/r$. Variable u runs over divisors of m that also divide d. Since $d|m$, this is equivalent to $u|d$, so

$$\sum_{\substack{r\colon r|m \\ \text{and } (m/r)|d}} \phi(m/r) = \sum_{u|d} \phi(u) = d\,.$$

8. Multicyclic de Bruijn sequences and the Extended Burrows-Wheeler Transformation

8.1. Multicyclic sequences

Higgins [10] defined a generalization of de Bruijn sequences called *de Bruijn sets* and showed how to generate them using an extension [13, 14] of the Burrows-Wheeler Transformation (BWT) [1]. We generalize this to incorporate our multiplicity parameter m.

The length of a sequence s is denoted $|s|$. For $i \geq 0$, let s^i denote concatenating i copies of s. A sequence s is *primitive* if its length is positive and s is not a power of a shorter word. Equivalently, a nonempty sequence s is primitive iff the cycle (s) is *aperiodic*, that is, $|s| > 0$ and the $|s|$ cyclic rotations of s are distinct. The *root* of s is the shortest prefix t of s such that $s = t^d$ for some $d \geq 1$; note that the root is primitive and $d = |s|/|t|$ is the rotation order of s. See [13, Sec. 3.1], [14, Sec. 2.1], and [10, p. 129].

For example, $abab$ is not primitive, but its root ab is, so (ab) is an aperiodic cycle while $(abab)$ is not. A *multicyclic sequence* is a multiset of aperiodic cycles; let $\mathcal{M}$ denote the set of all multicyclic sequences. In $\mathcal{M}$, instead of multiset notation $\{(s_1), (s_2), \ldots, (s_r)\}$, we will use cycle notation $(s_1)(s_2)\ldots(s_r)$, where different orderings of the cycles and different rotations are considered to be equivalent; this resembles permutation cycle notation, but symbols may appear multiple times and cycles are not composed as permutations. For example, $\sigma = (a)(a)(ababbb)$ denotes a multiset $\{(a), (a), (ababbb)\}$ with two distinct cycles (a) and one of $(ababbb)$. This representation is not unique; e.g., σ could also be written as $(a)(bbabab)(a)$.

The *length* of $\sigma = (s_1)\ldots(s_r)$ is $|\sigma| = \sum_{i=1}^{r} |s_i|$; e.g., $|(a)(a)(ababbb)| = 1 + 1 + 6 = 8$. For $n \geq 0$, let $\mathcal{M}_n = \{\sigma \in \mathcal{M} : |\sigma| = n\}$.

For $r \geq 0$, the notation $(s)^r$ denotes r copies of cycle (s), while $\langle s\rangle^r$ denotes concatenating r copies of s. E.g., $(ab)^3 = (ab)(ab)(ab)$ while $\langle ab\rangle^3 = ababab$. For $\sigma \in \mathcal{M}$, σ^r denotes multiplying each cycle's multiplicity by r.

Let s, w be nonempty linear sequences with s primitive. The number of occurrences of w in (s), denoted $\#\textsc{Occurrences}((s), w)$, is

$$\left|\{j = 0, 1, \ldots, |s| - 1 : w \text{ is a prefix of some power of } \rho^j(s)\}\right| .$$

For each j, it suffices to check if s is a prefix of $\langle\rho^j(s)\rangle^{\lceil |s|/|w| \rceil}$. For example, (ab) has one occurrence of $babab$, because $babab$ is a prefix of $\langle ba\rangle^{\lceil 5/2 \rceil} = \langle ba\rangle^3 = bababa$. The number of occurrences of w in $\sigma \in \mathcal{M}$ is

$$\#\textsc{Occurrences}(\sigma, w) = \sum_{(s) \in \sigma} \#\textsc{Occurrences}((s), w) \tag{22}$$

where each cycle (s) is repeated with its multiplicity in σ. For example, bab occurs five times in $(ab)(ab)(ab)(baababa)$: once in each (ab) as a prefix of $\langle ba\rangle^2 = baba$, and twice in $(baababa)$ as prefixes of $bababaa$ and $babaaba$.

A *multicyclic de Bruijn sequence* with parameters (m, q, k) is a multicyclic sequence $\sigma \in \mathcal{M}$ over an alphabet Ω of size q, such that every k-mer in Ω^k occurs exactly m times in σ. Let $\mathcal{MC}(m, q, k)$ denote the set of such σ. Higgins [10] introduced the case $m = 1$ and called it a *de Bruijn set*.

Example 8.1. Each of aa, ab, ba, bb occurs twice in $(a)(a)(ababbb)$. aa occurs once in the "first" (a) as a prefix of a^2, and once in the "second" (a). ab, ba, bb each occur twice in $(ababbb)$, including the occurrence of ba that wraps around and the two overlapping occurrences of bb. This is a multicyclic de Bruijn sequence with multiplicity $m = 2$, alphabet size $q = 2$, and word size $k = 2$. Table 3(D) lists all multicyclic sequences with these parameters.

8.2. Extended Burrows-Wheeler Transformation (EBWT)

The Burrows-Wheeler Transformation [1] is a certain map $\mathrm{BWT} : \Omega^n \to \Omega^n$ (for $n \geq 0$) that preserves the number of times each character occurs. For $n, q \geq 2$, it is neither injective nor surjective [14, p. 300], but it can be inverted in practical applications via a "terminator character" (see Sec. 8.5).

Mantaci et al. [13, 14] introduced the Extended Burrows-Wheeler Transformation, $\mathrm{EBWT} : \mathcal{M}_n \to \Omega^n$, which modifies the BWT to make it bijective. As with the BWT, the EBWT also preserves the number of times each character occurs. EBWT is equivalent to a bijection $\Omega^n \to \mathcal{M}_n$ of Gessel

Table 5: Notation for Extended Burrows-Wheeler Transformation

Definition	Notation
Linear sequence	abc or $\langle abc\rangle$
Cyclic sequence	(abc)
Power of a linear sequence	$\langle abc\rangle^r = abcabc\cdots$, r times
Repeated cycle	$(abc)^r = (abc)(abc)\cdots$, r times
Set of all multisets of aperiodic cycles (a.k.a. multicyclic sequences)	$\mathcal{M}$
Set of multicyclic sequences of length n	$\mathcal{M}_n$
Set of permutations of $0^m 1^m \cdots (q-1)^m$	$\mathcal{W}_{m,q}$
Burrows-Wheeler Transform	BWT
Extended Burrows-Wheeler Transform	EBWT
BWT table from forward transform	$T_B(s)$; $n \times n$ with $s \in \Omega^n$
BWT table from inverse transform	$T_{IB}(w)$; $n \times n$ with $w \in \Omega^n$
EBWT table from forward transform	$T_E(\sigma)$; $n \times c$ with $\sigma \in \mathcal{M}_n$
EBWT table from inverse transform	$T_{IE}(w)$; $n \times c$ with $w \in \Omega^n$

and Reutenauer [9, Lemma 3.4], but computed by an algorithm similar to the BWT rather than the method in [9]. There is also an earlier bijection $\Omega^n \to \mathcal{M}_n$ by de Bruijn and Klarner [6, Sec. 4] based on factorization into Lyndon words [2, Lemma 1.6], but it is not equivalent to the bijection in [9].

Higgins [10] provided further analysis of EBWT and used it to construct de Bruijn sets. We further generalize this to multicyclic de Bruijn sequences, as defined above. We will give a brief description of the EBWT algorithm and its inverse; see [13, 14, 10] for proofs.

Let $\sigma = (s_1)\cdots(s_r) \in \mathcal{M}$. Let c be the least common multiple of $|s_1|, |s_2|, \ldots, |s_r|$, and let $n = |\sigma| = \sum_{i=1}^{r} |s_i|$.

Consider a rotation $t = \rho^j(s_i)$ of a cycle of σ. Then $u = t^{c/|s_i|}$ has length c, and t can be recovered from u as the root of u.

Construct a table $T_E(\sigma)$ with n rows, c columns, and entries in Ω:

- Form a table of the powers of the rotations of each cycle of σ: the i^{th} cycle ($i = 1, \ldots, r$) generates rows $\langle \rho^j(s_i)\rangle^{c/|s_i|}$ for $j = 0, \ldots, |s_i - 1|$.
- Sort the rows into lexicographic order to form $T_E(\sigma)$.

The *Extended Burrows-Wheeler Transform of* σ, denoted EBWT(σ), is the linear sequence given by the last column of $T_E(\sigma)$ read from top to bottom.

Example 8.2. Let $\sigma = (0001)(011)(1)$. The number of columns is $c = \operatorname{lcm}(4, 3, 1) = 12$. The rotations of each cycle are shown on the left, and the table $T_E(\sigma)$ obtained by sorting these is shown on the right:

s_i	Powers of rotations of s_i	$T_E(\sigma)$
0001	0 0 0 1 0 0 0 1 0 0 0 1	0 0 0 1 0 0 0 1 0 0 0 1
	1 0 0 0 1 0 0 0 1 0 0 0	0 0 1 0 0 0 1 0 0 0 1 0
	0 1 0 0 0 1 0 0 0 1 0 0	0 1 0 0 0 1 0 0 0 1 0 0
	0 0 1 0 0 0 1 0 0 0 1 0	0 1 1 0 1 1 0 1 1 0 1 1
011	0 1 1 0 1 1 0 1 1 0 1 1	1 0 0 0 1 0 0 0 1 0 0 0
	1 0 1 1 0 1 1 0 1 1 0 1	1 0 1 1 0 1 1 0 1 1 0 1
	1 1 0 1 1 0 1 1 0 1 1 0	1 1 0 1 1 0 1 1 0 1 1 0
1	1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1 1 1 1 1 1 1 1 1

The last column of $T_E(\sigma)$ gives $\mathrm{EBWT}((0001)(011)(1)) = 10010101$.

Example 8.3. Let $\sigma = (0011)(0011)$. Then $c = \mathrm{lcm}(4, 4) = 4$. Since (0011) has multiplicity 2, each rotation of 0011 generates two equal rows. The last column of $T_E(\sigma)$ gives $\mathrm{EBWT}((0011)(0011)) = 11001100$:

s_i		$T_E(\sigma)$
0011	0 0 1 1	0 0 1 1
	1 0 0 1	0 0 1 1
	1 1 0 0	0 1 1 0
	0 1 1 0	0 1 1 0
0011	0 0 1 1	1 0 0 1
	1 0 0 1	1 0 0 1
	1 1 0 0	1 1 0 0
	0 1 1 0	1 1 0 0

8.3. Inverse of Extended Burrows-Wheeler Transformation

Without loss of generality, take $\Omega = \{0, 1, \ldots, q-1\}$. Let $w \in \Omega^n$. Mantaci et al. [13, 14] give the following method to construct $\sigma \in \mathcal{M}_n$ with $\mathrm{EBWT}(\sigma) = w$. Also see [9, Lemma 3.4] (which gives the same map but without relating it to the Burrows-Wheeler Transform) and [10, Theorem 1.2.11].

Compute the *standard permutation* of $w = a_0 a_1 \ldots a_{n-1}$ as follows.

- For each $i \in \Omega$, let h_i be the number of times i occurs in w, and let the positions of the i's be $p_{i0} < p_{i1} < \cdots < p_{i,h_i-1}$.
- Form partial sums $H_i = \sum_{j<i} h_j$ (for $i \in \Omega$) and $H_q = \sum_{j\in\Omega} h_j = n$.
- The *standard permutation* of w is the permutation π on $\{0, 1, \ldots, n-1\}$ defined by $\pi(H_i + j) = p_{i,j}$ for $i \in \Omega$ and $j = 0, \ldots, h_i - 1$.

Compute $\mathrm{EBWT}^{-1} : \Omega^* \to \mathcal{M}$, the *inverse Extended Burrows-Wheeler Transform*, as follows:

- Compute $\pi(w)$ and express it in permutation cycle form.
- In each cycle of $\pi(w)$, replace entries in the range $[H_i, H_{i+1}-1]$ by i.
- This forms an element $\sigma \in \mathcal{M}$. Output σ.

Mantaci et al. [13], [14, Theorem 20] prove that these algorithms for EBWT and EBWT^{-1} satisfy EBWT$(\sigma) = w$ iff EBWT$^{-1}(w) = \sigma$. Thus,

Theorem 8.4 (Mantaci et al.)**.** *EBWT* : $\mathcal{M}_n \to \Omega^n$ *is a bijection.*

Example 8.5. Let $w = a_0 \dots a_7 = 10010101$. The four 0's have positions 1, 2, 4, 6, so $\pi(0) = 1$, $\pi(1) = 2$, $\pi(2) = 4$, $\pi(3) = 6$. The four 1's have positions 0, 3, 5, 7, so $\pi(4) = 0$, $\pi(5) = 3$, $\pi(6) = 5$, $\pi(7) = 7$. In cycle form, this permutation is $\pi = (0, 1, 2, 4)(3, 6, 5)(7)$. In the cycle form, replace $0, 1, 2, 3$ by 0 and $4, 5, 6, 7$ by 1 to obtain $\sigma = (0, 0, 0, 1)(0, 1, 1)(1)$. Thus, EBWT$^{-1}(10010101) = (0001)(011)(1)$.

The inverse EBWT may also be computed by constructing the EBWT table $T_{IE}(w)$ of a linear sequence w. This is analogous to the procedure used with the original BWT [1, Algorithm D], and was made explicit for the EBWT in [10, pp. 130–131]; also see [13, pp. 182-183] and [14, Theorem 20]. Computing the EBWT by this procedure is useful for theoretical analysis, but the standard permutation is more efficient for computational purposes.

- Input: $w \in \Omega^n$.
- Form a table with n empty rows.
- Let $c = 0$. This is the number of columns constructed so far.
- Repeat the following until the last column equals w:
  - Extend the table from $n \times c$ to $n \times (c+1)$ by shifting all c columns one to the right and filling in the first column with w.
  - Sort the rows lexicographically.
  - Increment c.
- Output the $n \times c$ table $T_{IE}(w)$.

A *Lyndon word* [12] is a linear sequence that is primitive and is lexicographically smaller than its nontrivial cyclic rotations. To compute the inverse EBWT, form $T_{IE}(w)$ and take the primitive root of each row. Output the multiset of cycles generated by the primitive roots that are Lyndon words. Note that if $w =$ EBWT(σ), then the tables match, $T_E(\sigma) = T_{IE}(w)$, and the inverse constructed this way satisfies EBWT$^{-1}(w) = \sigma$. Mantaci et al. [13, 14] also consider selecting other linearizations of the cycles, rather than the Lyndon words or the cycles they generate.

Table 6: $\mathcal{MC}(2,2,2)$: Multicyclic de Bruijn sequences σ for $m = q = k = 2$, and their Extended Burrows-Wheeler Transforms $w = \mathrm{EBWT}(\sigma)$. Entries corresponding to cyclic multi de Bruijn sequences $\mathcal{C}(2,2,2)$ are marked '*'

w	σ
0011 0011	(0)(0)(01)(01)(1)(1)
0011 0101	(0)(0)(01)(011)(1)
0011 0110	(0)(0)(01)(0111)
0011 1001	(0)(0)(01011)(1)
0011 1010	(0)(0)(010111)
0011 1100	(0)(0)(011)(011)
0101 0011	(0)(001)(01)(1)(1)
0101 0101	(0)(001)(011)(1)
0101 0110	(0)(001)(0111)
0101 1001	(0)(001011)(1)
0101 1010	(0)(0010111)
0101 1100	(0)(0011)(011)
0110 0011	(0)(00101)(1)(1)
0110 0101	(0)(001101)(1)
0110 0110	(0)(0011101)
0110 1001	(0)(0011)(01)(1)
0110 1010	(0)(00111)(01)
0110 1100	(0)(0011011)
1001 0011	(0001)(01)(1)(1)
1001 0101	(0001)(011)(1)
1001 0110	(0001)(0111)
1001 1001	(0001011)(1)
1001 1010*	(00010111)*
1001 1100	(00011)(011)
1010 0011	(000101)(1)(1)
1010 0101	(0001101)(1)
1010 0110*	(00011101)*
1010 1001	(00011)(01)(1)
1010 1010	(000111)(01)
1010 1100*	(00011011)*
1100 0011	(001)(001)(1)(1)
1100 0101	(001)(0011)(1)
1100 0110	(001)(00111)
1100 1001	(0010011)(1)
1100 1010*	(00100111)*
1100 1100*	(0011)(0011)*

Example 8.6. For $w = 10010101$, the table $T_{IE}(10010101)$ is identical to the righthand table in Example 8.2. The primitive roots of the rows are 0001, 0010, 0100, 011, 1000, 101, 110, 1. The Lyndon words among these are 0001, 011, and 1, giving $\mathrm{EBWT}^{-1}(10010101) = (0001)(011)(1)$.

Example 8.7. For $w = 11001100$, the table $T_{IE}(11001100)$ is identical to the righthand table in Example 8.3. The primitive roots of the rows are 0011, 0011, 0110, 0110, 1001, 1001, 1100, 1100. The Lyndon word among these is 0011 with multiplicity 2, giving $\mathrm{EBWT}^{-1}(11001100) = (0011)(0011)$.

8.4. Applying the EBWT to multicyclic de Bruijn sequences

We have the following generalization of Higgins [10, Theorem 3.4], which was for the case $m = 1$. Our proof is similar to the one in [10] but introduces the multiplicity m. Table 6 illustrates the theorem for $(m, q, k) = (2, 2, 2)$.

Theorem 8.8. *For $m, q, k \geq 1$,*

(a) The image of $\mathcal{MC}(m,q,k)$ under EBWT *is $(\mathcal{W}_{m,q})^{q^{k-1}}$.*

(b) $|\mathcal{MC}(m,q,k)| = W(m,q,k) = \frac{(mq)!^{q^{k-1}}}{m!^{q^k}}$.

Proof. When $q = 1$, the theorem holds:

$$\begin{aligned} \mathcal{MC}(m,1,k) &= \{(0)^m\} \text{ has size } 1, \\ (\mathcal{W}_{m,q})^{q^{k-1}} = \mathcal{W}_{m,1} &= \{0^m\} \quad \text{has size } W(m,1,k) = 1, \\ \mathrm{EBWT}((0)^m) &= 0^m. \end{aligned}$$

Below, we prove (a) for $q \geq 2$. Since EBWT is invertible, (b) follows from $|\mathcal{MC}(m,q,k)| = |\,\mathrm{EBWT}(\mathcal{MC}(m,q,k))| = |(\mathcal{W}_{m,q})^{q^{k-1}}| = W(m,q,k)$.

In both the forwards and inverse EBWT table, the rows starting with nonempty sequence x (meaning x is a prefix of a suitable power of the row) are consecutive (since the table is sorted), yielding a block B_x of consecutive rows. In the forwards direction, in $T_E(\sigma)$, the number of rows in B_x equals $\#\textsc{Occurrences}(\sigma, x)$, since for each occurrence of x in σ, a row is generated by rotating the occurrence to the beginning and taking a suitable power.

Forwards: Given $\sigma \in \mathcal{MC}(m,q,k)$, we show that $\mathrm{EBWT}(\sigma) \in (\mathcal{W}_{m,q})^{q^{k-1}}$.

Every k-mer occurs exactly m times in σ, so the rows of $T_E(\sigma)$ are partitioned by their initial k-mer into q^k blocks of m consecutive rows. When $q > 1$, the table has at least k columns since 01^{k-1} occurs in σ and generates a row of size at least k.

For each $(k-1)$-mer $x \in \Omega^{k-1}$, block B_x of $T_E(\sigma)$ has mq rows, because σ has m occurrences of k-mer xb for each of the q choices of $b \in \Omega$.

Suppose by way of contradiction that at least $m+1$ rows of B_x end in the same symbol $a \in \Omega$. Upon cyclically rotating the table one column to the right and sorting the rows (which leaves the table invariant), at least $m+1$ rows begin with k-mer ax, which contradicts that exactly m rows start with each k-mer. Thus, at most m rows of B_x end in the same symbol. Since B_x has mq rows and each of the q symbols occurs at the end of at most m of them, in fact, each symbol must occur at the end of exactly m of them. Let w_x be the word formed by the last character in the rows of B_x, from top to bottom. Then $w_x \in \mathcal{W}_{m,q}$. $\mathrm{EBWT}(\sigma)$ is the last column of the table, which is the concatenation of the q^{k-1} words w_x for $x \in \Omega^{k-1}$ in order by x.

Inverse: Given $w \in (\mathcal{W}_{m,q})^{q^{k-1}}$, we show that $\mathrm{EBWT}^{-1}(w) \in \mathcal{MC}(m,q,k)$.

For $j \in \{1,\ldots,k\}$, we use induction to show that exactly mq^{k-j} rows of $T_{IE}(w)$ start with each $x \in \Omega^j$. See Examples 8.2–8.3, in which $mq^{k-1} = 2 \cdot 2^{2-1} = 4$ rows start with each of 0 and 1 and $mq^{k-2} = 2 \cdot 2^{2-2} = 2$ rows start with each of 00, 01, 10, and 11.

Base case $j = 1$: Each letter in Ω occurs exactly mq^{k-1} times in w. Upon cyclically shifting the last column, w, of the table to the beginning and sorting rows to obtain the first column, this gives exactly mq^{k-1} occurrences of each letter in the first column. This proves the case $j = 1$.

Induction step: Given $j \in \{2, \dots, k\}$, suppose the statement holds for case $j - 1$. For each $x \in \Omega^{j-1}$, block B_x consists of mq^{k-j+1} consecutive rows. Split B_x into into q^{k-j} subblocks of mq consecutive rows (we may do this since $j \leq k$). The last column of each subblock forms an interval in w of length mq taken from $\mathcal{W}_{m,q}$, so each $a \in \Omega$ occurs at the end of exactly m rows in each subblock. Thus, in the whole table, exactly $m \cdot q^{k-j}$ rows start in $(k-1)$-mer x and end in character a. On rotating the last column to the beginning, exactly mq^{k-j} rows start with k-mer ax.

The case $j = k$ shows that every k-mer in Ω^k is a prefix of exactly m rows of $T_{IE}(w)$. Thus, every k-mer occurs exactly m times in $\mathrm{EBWT}^{-1}(w)$, so $\mathrm{EBWT}^{-1}(w) \in \mathcal{MC}(m, q, k)$. □

Theorem 8.8(a) may be used to generate multicyclic de Bruijn sequences:

- To generate all sequences: $\mathcal{MC}(m, q, k) = \{\mathrm{EBWT}^{-1}(w) : w \in \mathcal{W}_{m,q}\}$; see Table 6.
- To select a uniform random element of $\mathcal{MC}(m, q, k)$, do $n = q^{k-1}$ rolls of a fair $|\mathcal{W}_{m,q}| = (mq)!/m!^q$-sided die to pick $w_0, \dots, w_{n-1} \in \mathcal{W}_{m,q}$. Compute $\mathrm{EBWT}^{-1}(w_0 \dots w_{n-1})$.

8.5. Original vs. Extended Burrows-Wheeler Transformation

To compute the original Burrows-Wheeler Transform $\mathrm{BWT}(s)$ of $s \in \Omega^n$ (see [1]): form an $n \times n$ table, $T_B(s)$, whose rows are $\rho^0(s), \dots, \rho^{n-1}(s)$ sorted into lexicographic order. Read the last column from top to bottom to form a linear sequence, $\mathrm{BWT}(s)$. This is similar to computing $\mathrm{EBWT}((s))$, but s does not have to be primitive, and the table is always square.

Given $w \in \Omega^n$, we construct the inverse BWT table $T_{IB}(w)$ by the same algorithm as for $T_{IE}(w)$, except we stop at n columns rather than when the last column equals w. Table $T_{IB}(w)$ is defined for all w, but $\mathrm{BWT}^{-1}(w)$ may or may not exist. If $T_{IB}(w)$ consists of the n rotations of its first row, sorted into order, then one or more inverses $\mathrm{BWT}^{-1}(w)$ exist (take any row of the table as the inverse); otherwise, an inverse $\mathrm{BWT}^{-1}(w)$ does not exist.

Both s and its rotations have the same BWT table and the same transforms (see [14, Prop. 1]): $T_B(\rho^j(s)) = T_B(s)$ and $\mathrm{BWT}(\rho^j(s)) = \mathrm{BWT}(s)$ for $j = 0, \dots, |s| - 1$. Given $w = \mathrm{BWT}(s)$, the inverse may only be computed up to an unknown rotation, $\rho^j(s)$. As an aside, we note the following, but we do not use it: in order to recover a linear sequence s from the inverse, Burrows and Wheeler [1, Sec. 4.1] introduce a terminator character (which we denote '\$') that does not occur in the input. To encode, compute

$w = \mathrm{BWT}(s\$)$. To decode, compute $\mathrm{BWT}^{-1}(w)$, select the rotation that puts '\$' at the end, and delete '\$' to recover s. Terminator characters are common in implementations of the BWT, but since our application is cyclic sequences rather than linear sequences, we do not use a terminator.

If s is primitive, then $T_B(s) = T_E((s))$ and $\mathrm{BWT}(s) = \mathrm{EBWT}((s))$. More generally, any sequence s can be decomposed as $s = t^d$ with t primitive and $d \geq 0$, for which we have:

Theorem 8.9. *Let $s = t^d$, with t primitive and $d \geq 0$. Then* $\mathrm{BWT}(s) = \mathrm{EBWT}((t)^d) = {a_1}^d {a_2}^d \cdots {a_r}^d$*, where $r = |t|$ and* $\mathrm{BWT}(t) = \mathrm{EBWT}((t)) = a_1 \ldots a_r$.

Theorem 8.10.

(a) Given $\mathrm{EBWT}(\sigma) = a_1 \ldots a_r$*, then* $\mathrm{EBWT}(\sigma^d) = {a_1}^d \ldots {a_r}^d$.
(b) Given $\mathrm{EBWT}^{-1}(a_1 \ldots a_r) = \sigma$*, then* $\mathrm{EBWT}^{-1}({a_1}^d \ldots {a_r}^d) = \sigma^d$.

In the following, "$\mathrm{BWT}^{-1}(w)$ exists" means that at least one inverse exists, not that a unique inverse exists.

Theorem 8.11. *(a)* $\mathrm{BWT}^{-1}(w)$ *exists iff* $\mathrm{EBWT}^{-1}(w)$ *has the form $(t)^d$, where t is primitive and $d \geq 0$ (specifically $d = |w|/|t|$). In this case,* $\mathrm{BWT}^{-1}(w) = t^d$ *(or any rotation of it).*

(b) $s_1 = \mathrm{BWT}^{-1}(a_1 a_2 \cdots a_r)$ *exists iff* $s_2 = \mathrm{BWT}^{-1}({a_1}^d {a_2}^d \cdots {a_r}^d)$ *exists. If these exist, then up to a rotation, $s_2 = {s_1}^d$.*

Theorem 8.11(b) is essentially equivalent to [14, Prop. 2].

The above three theorems are trivial when the input is null. In the proofs below, we assume the input sequences have positive length.

Proof of Theorem 8.9. Let $n = |s| = rd$. There are $|t| = r$ distinct rotations of s. In the sequence $\rho^j(s)$ for $j = 0, \ldots, n-1$, each distinct rotation is generated d times. These are sorted into order to form $T_B(s)$, giving r blocks of d consecutive equal rows. Therefore, the last column of the table has the form ${a_1}^d {a_2}^d \cdots {a_r}^d$. The BWT construction ensures this is a permutation of the input $s = t^d$, so $a_1 \ldots a_r$ is a permutation of t.

Let $\sigma = (t)^d$. Since t is primitive, $T_E(\sigma)$ consists of each rotation of t repeated on d consecutive rows, sorted into order. This is identical to the first r columns of $T_B(s)$. Conversely, $T_B(s) = [T_E(\sigma) \ \cdots \ T_E(\sigma)]$ (d copies of $T_E(\sigma)$ horizontally). Thus, the last columns of $T_B(s)$ and $T_E(\sigma)$ are the same, so $\mathrm{BWT}(s) = \mathrm{EBWT}(\sigma)$. □

Proof of Theorem 8.10. (a) Replicate each row of $T_E(\sigma)$ on d consecutive rows to form $T_E(\sigma^d)$. Given that the last column of $T_E(\sigma)$ is $\mathrm{EBWT}(\sigma) = a_1 \ldots a_r$, then the last column of $T_E(\sigma^d)$ is $\mathrm{EBWT}(\sigma^d) = {a_1}^d \ldots {a_r}^d$.

(b) Applying EBWT^{-1} to both equations in (a) gives (b). □

Proof of Theorem 8.11. (a) Suppose $s = \mathrm{BWT}^{-1}(w)$ exists. Decompose $s = t^d$ where t is primitive and $d \geq 0$. Then $w = \mathrm{BWT}(t^d)$, which, by Theorem 8.9, equals $\mathrm{EBWT}((t)^d)$, so $\mathrm{EBWT}^{-1}(w) = (t)^d$.

Conversely, suppose $\mathrm{EBWT}^{-1}(w) = (t)^d$, where t is primitive. By Theorem 8.9, $\mathrm{BWT}(t^d) = \mathrm{EBWT}((t)^d) = w$, so $\mathrm{BWT}^{-1}(w) = t^d$.

Note that $\mathrm{BWT}^{-1}(w)$ and linearizations of cycles of $\mathrm{EBWT}^{-1}(w)$ are only defined up to a rotation, so other rotations of s and t may be used.

(b) Let $w = a_1 \ldots a_r$ and $w' = {a_1}^d \ldots {a_r}^d$. Let $\sigma = \mathrm{EBWT}^{-1}(w)$. By Theorem 8.10, $\mathrm{EBWT}^{-1}(w') = \sigma^d$. If σ has the form $(t)^h$ for some primitive cycle t and integer $h \geq 0$, then $\sigma^d = (t)^{hd}$. By part (a), $\mathrm{BWT}^{-1}(w) = t^h$ and $\mathrm{BWT}^{-1}(w') = t^{hd}$ (or any rotations of these), so $\mathrm{BWT}^{-1}(w')$ is the d^{th} power of $\mathrm{BWT}^{-1}(w)$. But if σ doesn't have that form, then $\mathrm{BWT}^{-1}(w)$ and $\mathrm{BWT}^{-1}(w')$ don't exist. □

Set $\mathcal{D} = \{w \in \Omega^+ : w$ does not have the form ${a_1}^i {a_2}^i \cdots$ for any $i > 1\}$.

Theorem 8.12. *Let $(s) \in \mathcal{C}^{(d)}(m, q, k)$, with $m, q, k \geq 1$. Then* $\mathrm{BWT}(s)$ *has the form ${a_1}^d \ldots {a_r}^d$ where $a_1 \ldots a_r \in (\mathcal{W}_{m/d,q})^{q^{k-1}} \cap \mathcal{D}$.*

Proof. Since s has order d, it has the form $s = t^d$ with t primitive. By Lemma 3.1, $t \in \mathcal{C}^{(1)}(m/d, q, k)$. Since t is primitive, $\mathrm{BWT}(t) = \mathrm{EBWT}((t))$; say this is $u = a_1 \ldots a_r$. By Theorem 8.8(a), $u \in (\mathcal{W}_{m/d,q})^{q^{k-1}}$, and by Theorem 8.9, $\mathrm{BWT}(s) = {a_1}^d \ldots {a_r}^d$

Now suppose by way of contradiction that $u = {b_1}^i \ldots {b_{r/i}}^i$ with $i | r$ and $i > 1$. By Theorem 8.10(b), $\mathrm{EBWT}^{-1}({b_1}^i \ldots {b_{r/i}}^i)$ has at least i cycles; but $\mathrm{EBWT}^{-1}(u) = (t)$ has exactly one cycle, a contradiction. Thus, $u \in \mathcal{D}$. □

Example 8.13. Cyclic multi de Bruijn sequences in $\mathcal{C}(2, 2, 2)$ correspond to entries marked '*' in Table 6. Spaces highlight the structure from Theorem 8.8(a) but are not part of the sequences.

$$\begin{aligned}
\mathrm{BWT}(00010111) &= \mathrm{EBWT}((00010111)) &&= 1001\ 1010\\
\mathrm{BWT}(00011011) &= \mathrm{EBWT}((00011011)) &&= 1010\ 1100\\
\mathrm{BWT}(00011101) &= \mathrm{EBWT}((00011101)) &&= 1010\ 0110\\
\mathrm{BWT}(00100111) &= \mathrm{EBWT}((00100111)) &&= 1100\ 1010\\
\mathrm{BWT}(00110011) &= \mathrm{EBWT}((0011)(0011)) &&= 1100\ 1100 = 1^2 0^2 1^2 0^2\ .
\end{aligned}$$

By Theorem 8.9, BWT(00110011) is derived from this example with $m = 1$:

$$\mathrm{BWT}(0011) = \mathrm{EBWT}((0011)) = 10\ 10\ .$$

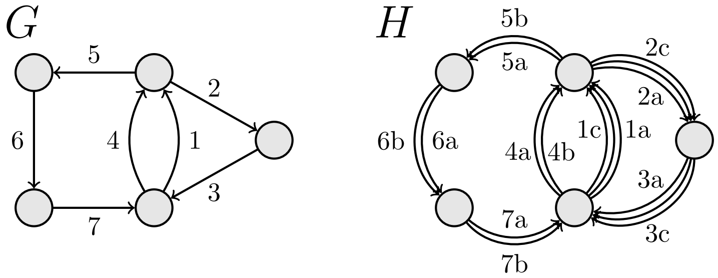


Figure 2: Given directed multigraph G and edge multiplicities $\nu_1 = \nu_2 = \nu_3 = 3$, $\nu_4 = \nu_5 = \nu_6 = \nu_7 = 2$, replace each edge e of G by ν_e edges to form H. For example, edge 2 in G yields edges $2a, 2b, 2c$ in H.

9. Partitioning a graph into aperiodic cycles

We give an alternate proof of Theorem 8.8(b) based on the graph $G(m, q, k)$ rather than the Extended Burrows-Wheeler Transform.

Let G be a multigraph with vertices $V(G)$ and edges $E(G)$.

Denote the sets of incoming and outgoing edges at vertex x by $\textsc{In}_G(x) = \{e \in E(G) : \textsc{Head}(e) = x\}$ and $\textsc{Out}_G(x) = \{e \in E(G) : \textsc{Tail}(e) = x\}$.

The *period* of a cycle on edges $(e_1, \ldots, e_n)$ is the smallest $p \in \{1, \ldots, n\}$ for which $e_i = e_{i+p}$ (subscripts taken mod n) for $i = 1, \ldots, n$. A cycle is *aperiodic* when its period is its length.

Let $\vec{\nu} = \langle \nu_e : e \in E(G) \rangle$ be an assignment of nonnegative integers to edges of G. Let $\mathcal{M}_G(\vec{\nu})$ be the collection of multisets of aperiodic cycles in G in which each edge e is used exactly ν_e times in the multiset. Different representations of a cycle, such as $(e_1, \ldots, e_n)$ vs. $(e_{i+1}, \ldots, e_n, e_1, \ldots, e_i)$, are considered equivalent. In this section, we will determine $|\mathcal{M}_G(\vec{\nu})|$. Set

$$\textsc{Indeg}_G(x, \nu) = \sum_{e \in \textsc{In}_G(x)} \nu_e \qquad \textsc{Outdeg}_G(x, \nu) = \sum_{e \in \textsc{Out}_G(x)} \nu_e \,.$$

A necessary condition for such a multiset to exist is that at each vertex $x \in V(G)$, $\textsc{Indeg}_G(x, \nu) = \textsc{Outdeg}_G(x, \nu)$, because the number of times vertex x is entered (resp., exited) among the cycles is given by $\textsc{Indeg}_G(x, \nu)$ (resp., $\textsc{Outdeg}_G(x, \nu)$). Below, we assume this holds.

Form a directed multigraph H with the same vertices as G, and with each edge e of G replaced by ν_e distinguishable edges on the same vertices as e. See Fig. 2. Since G is a multigraph, if there are r edges $e_1, \ldots, e_r$ from v to w in G, there will be $\sum_{i=1}^{r} \nu_{e_i}$ edges from v to w in H. By construction, at every vertex x, $\textsc{Indeg}_H(x) = \textsc{Indeg}_G(x, \nu)$ and $\textsc{Outdeg}_H(x) = \textsc{Outdeg}_G(x, \nu)$.

Combining this with $\textsc{Indeg}_G(x,\nu) = \textsc{Outdeg}_G(x,\nu)$ gives that H is balanced (at each vertex, the sum of the indegrees equals the sum of the outdegrees). Note that we do not require G or H to be connected.

Define a map $\pi : H \to G$ that preserves vertices and maps the ν_e edges of H corresponding to e back to e in G. Extend π to map multisets of cycles in H to multisets of cycles in G as follows:

$$\begin{aligned}\pi\big((e_{1,1},\ldots,e_{1,n_1})\cdots(e_{r,1},\ldots,e_{r,n_r})\big)\\ = \big(\pi(e_{1,1}),\ldots,\pi(e_{1,n_1})\big)\cdots\big(\pi(e_{r,1}),\ldots,\pi(e_{r,n_r})\big)\ .\end{aligned}$$

A *cycle partition* of H is a set of cycles in H, with each edge of H used once over the set. Let $\mathrm{CP}(H)$ denote the set of all cycle partitions of H.

An *edge successor map* of H is a function $f : E(H) \to E(H)$ such that for every edge $e \in E(H)$, $\textsc{Head}(e) = \textsc{Tail}(f(e))$, and for every vertex $x \in V(H)$, f restricts to a bijection $f : \textsc{In}_H(x) \to \textsc{Out}_H(x)$. Let $F(H)$ be the set of all edge successor maps of H.

Theorem 9.1. *There is a bijection between* $\mathrm{CP}(H)$ *and* $F(H)$*.*

Denote the edge successor map corresponding to cycle partition C by f_C, and the cycle partition corresponding to edge successor map f by C_f.

Proof. Let $C \in \mathrm{CP}(H)$. Construct $f_C \in F(H)$ as follows: for each edge $e \in E(H)$, set $f_C(e) = e'$ where e' is the unique edge following e in its cycle in C; note that $\textsc{Head}(e) = \textsc{Tail}(e')$, as required. Every edge appears exactly once in C and has exactly one image and one inverse in f_C. Thus, at each vertex x, f_C restricts to a bijection $f_C : \textsc{In}_H(x) \to \textsc{Out}_H(x)$, so $f_C \in F(H)$.

Conversely, given $f \in F(H)$, construct a cycle partition C_f as follows. f is a permutation of the finite set $E(H)$. Express this permutation in cycle form, C_f. Each permutation cycle has the form $(e_1, e_2, \ldots, e_n)$ with $f(e_i) = e_{i+1}$ (subscripts taken mod n). Since $\textsc{Head}(e_i) = \textsc{Tail}(f(e_i)) = \textsc{Tail}(e_{i+1})$, these permutation cycles are also graph cycles, so $C_f \in \mathrm{CP}(H)$.

By construction, the maps $f \mapsto C_f$ and $C \mapsto f_C$ are inverses. □

Corollary 9.2. *Let H be a finite balanced directed multigraph, with all edges distinguishable. The number of cycle partitions of H is*

$$|\,\mathrm{CP}(H)| = \prod_{x\in V(H)} (\textsc{Outdeg}(x))! \tag{23}$$

Proof. At each $x \in V(H)$, there are $(\textsc{Outdeg}(x))!$ bijections from $\textsc{In}(x)$ to $\textsc{Out}(x)$, so $|F(H)|$ is given by (23). By Theorem 9.1, the number of cycle partitions of H is also given by (23). □

Cycles in $\pi(C_f)$ are not necessarily aperiodic, but may be split into aperiodic cycles as follows. Replace each cycle $D = (e_1, \ldots, e_n)$ of $\pi(C_f)$ by n/p cycles $(e_1, \ldots, e_{n/p})$, where p is the period of D. This is well-defined; had we represented D by a different rotation, the result would be equivalent, but represented by a different rotation. Let $\textsc{Split}(\pi(C_f))$ denote splitting all cycles of $\pi(C_f)$ in this fashion; this is a multiset of aperiodic cycles in G.

Theorem 9.3. *If* $\textsc{Indeg}_G(x, \nu) = \textsc{Outdeg}_G(x, \nu)$ *at every* $x \in V(G)$*, then*

$$|\mathcal{M}_G(\vec{\nu})| = \frac{\prod_{x \in V(G)} (\textsc{Outdeg}_G(x, \nu))!}{\prod_{e \in E(G)} \nu_e!}. \tag{24}$$

Proof. By Corollary 9.2, the number of cycle partitions of H is

$$\prod_{x \in V(H)} (\textsc{Outdeg}_H(x))! = \prod_{x \in V(G)} (\textsc{Outdeg}_G(x, \nu))! \tag{25}$$

In $F(H)$, define $f \equiv f'$ iff every edge $e \in V(H)$ satisfies $\pi(f(e)) = \pi(f'(e))$. Although C_f and $C_{f'}$ may have different cycle structures, on splitting the cycles, $\textsc{Split}(\pi(C_f)) = \textsc{Split}(\pi(C_{f'}))$. The size of each equivalence class is $\prod_{e \in E(G)} \nu_e!$. Dividing (25) by this gives (24). □

Example 9.4. In Fig. 2, these have different cycle structures:

$$\begin{aligned} C_f &= (4a, 5a, 6a, 7a)(1a, 2a, 3a, 1b, 2b, 3b)(1c, 5b, 6b, 7b, 4b, 2c, 3c) \\ C_{f'} &= (4a, 5a, 6a, 7a)(1a, 2a, 3a)(1b, 2b, 3b)(1c, 5b, 6b, 7b, 4b, 2c, 3c) \end{aligned}$$

However, they are in the same equivalence class. For all edges except $3a$ and $3b$, we have $f(e) = f'(e)$, so $\pi(f(e)) = \pi(f'(e))$. For edges $3a$ and $3b$,

$$\begin{array}{lll} f(3a) = 1b & f'(3a) = 1a & \pi(f(3a)) = \pi(f'(3a)) = 1 \\ f(3b) = 1a & f'(3b) = 1b & \pi(f(3b)) = \pi(f'(3b)) = 1 \end{array}$$

so $f \equiv f'$. Applying π and splitting the cycles gives

$$\textsc{Split}(\pi(C_f)) = \textsc{Split}(\pi(C_{f'})) = (4, 5, 6, 7)(1, 2, 3)(1, 2, 3)(1, 5, 6, 7, 4, 2, 3).$$

More generally in Fig. 2, let $\nu_1 = \nu_2 = \nu_3 = A$ and $\nu_4 = \nu_5 = \nu_6 = \nu_7 = B$. Fig. 2 shows H for $A = 3$ and $B = 2$. The values of $\textsc{Outdeg}_G(x, \nu)$ clockwise from the rightmost vertex are $A, A + B, B, B, A + B$. We obtain

$$|\mathcal{M}_G(\vec{\nu})| = \frac{A!(A+B)!^2 B!^2}{\prod_{i=1}^{7} \nu_i!} = \frac{A!(A+B)!^2 B!^2}{A!^3 B!^4} = \frac{(A+B)!^2}{A!^2 B!^2} = \binom{A+B}{A}^2.$$

Finally, we apply (24) to count multicyclic de Bruijn sequences. Let $G = G(1, q, k)$ and $H = G(m, q, k)$. This has $\nu_e = m$ for all $e \in \Omega^k$. Each vertex in H has indegree and outdegree mq. Thus,

$$|\mathcal{M}_G(\vec{\nu})| = \frac{\prod_{x \in \Omega^{k-1}} (mq)!}{\prod_{\alpha \in \Omega^k} m!} = \frac{(mq)!^{q^{k-1}}}{m!^{q^k}} = W(m, q, k) \ . \tag{26}$$

Each edge cycle $(e_1, \ldots, e_n)$ over G yields a cyclic sequence of length n over Ω by taking the first (or last) letter of the k-mer labelling each edge. The edge cycle is aperiodic in G iff the cyclic sequence of the cycle is aperiodic. Thus, $|\mathcal{MC}(m, q, k)|$ is given by (26), which agrees with Theorem 8.8(b).

References

[1] M. Burrows and D.J. Wheeler, *A Block-sorting Lossless Data Compression Algorithm*, Systems Research Center Research Report 124, Digital Equipment Corporation, 1994.

[2] K.-T. Chen, R.H. Fox, and R.C. Lyndon, *Free differential calculus. IV. The quotient groups of the lower central series*, Ann. of Math. (2) **68** (1958), 81–95. MR0102539

[3] R. Dawson and I.J. Good, *Exact Markov probabilities from oriented linear graphs*, Ann. Math. Stat. **28** (1957), 946–956. MR0093819

[4] N.G. de Bruijn, *A combinatorial problem*, Nederl. Akad. Wetensch., Proc. **49** (1946), 758–764. Equivalent to Indag. Math. **8** (1946), 461–467. MR0018142

[5] N.G. de Bruijn, *Acknowledgement of priority to C. Flye Sainte-Marie on the counting of circular arrangements of 2^n zeros and ones that show each n-letter word exactly once*, Technological University Eindhoven, T.H.-Report 75-WSK-06, June 1975.

[6] N.G. de Bruijn and D.A. Klarner, *Multisets of aperiodic cycles*, SIAM J. Algebraic Discrete Methods **3** (1982), no. 3, 359–368. MR0666861

[7] A. de Rivière, *Question 48*, l'Intermédiare des Mathématiciens **1** (1894), 19–20.

[8] H. Fredricksen, *A survey of full length nonlinear shift register cycle algorithms*, SIAM Rev. **24** (1982), no. 2, 195–221. MR0652466

[9] I.M. Gessel and C. Reutenauer, *Counting permutations with given cycle structure and descent set*, J. Combin. Theory Ser. A **64** (1993), no. 2, 189–215. MR1245159

[10] P.M. Higgins, *Burrows-Wheeler transformations and de Bruijn words*, Theoret. Comput. Sci. **457** (2012), 128–136. MR2961204

[11] D. Kandel, Y. Matias, R. Unger, and P. Winkler, *Shuffling biological sequences*, Discrete Appl. Math. **71** (1996), no. 1–3, 171–185. MR1420298

[12] R.C. Lyndon, *On Burnside's problem*, Trans. Amer. Math. Soc. **77** (1954), 202–215. MR0064049

[13] S. Mantaci, A. Restivo, G. Rosone, and M. Sciortino, *An extension of the Burrows Wheeler transform and applications to sequence comparison and data compression*, Combinatorial pattern matching, 16th annual symposium, CPM 2005, Berlin: Springer, 2005, pp. 178–189 (English). MR2215399

[14] S. Mantaci, A. Restivo, G. Rosone, and M. Sciortino, *An extension of the Burrows-Wheeler transform*, Theoret. Comput. Sci. **387** (2007), no. 3, 298–312. MR2364570

[15] V. Al. Osipov, *Wavelet Analysis on Symbolic Sequences and Two-Fold de Bruijn Sequences*, J. Stat. Phys. **164** (2016), no. 1, 142–165. MR3509051

[16] C.F. Sainte-Marie, *Solution to question 48*, l'Intermédiare des Mathématiciens **1** (1894), 107–110.

[17] R.P. Stanley, *Enumerative combinatorics. Vol. 2*, Cambridge University Press, Cambridge, 1999. MR1676282

[18] W.T. Tutte, *The dissection of equilateral triangles into equilateral triangles*, Proc. Cambridge Philos. Soc. **44** (1948), 463–482. MR0027521

[19] T. van Aardenne-Ehrenfest and N.G. de Bruijn, *Circuits and trees in oriented linear graphs*, Simon Stevin **28** (1951), 203–217. MR0047311

Glenn Tesler
Department of Mathematics
University of California
San Diego
La Jolla, CA 92093-0112
USA
E-mail address: gptesler@math.ucsd.edu

Received 16 August 2016