

# Burrows-Wheeler transforms and de Bruijn words

Peter M. Higgins

Department of Mathematical Sciences, University of Essex

Bar Ilan University November 2012

The Burrows-Wheeler transform  $BW(w)$  of a word  $w$  is a certain word of length  $|w|$  over the same alphabet that is used as a tool of lossless data compression.

## Definitions

Let  $A = \{a_0 < a_1 < \dots\}$  be a finite alphabet,  $A^+(A^*)$  the free semigroup (resp. free monoid) over  $A$ . A word  $w \in A^+$  is *primitive* if  $w$  is not a power of some other word. A *conjugate* of  $w$  is a word  $w' = vu$  where  $w = uv$ . The first word in the lexicographic order of the conjugacy class of a primitive word  $w$  is the *Lyndon word* of the *necklace* (conjugacy class)  $n(w)$  of  $w$ .

The BW transform of two conjugate words is the same. Inverting BW transforms can be done in linear time but in general inversion gives rise to multisets of necklaces rather than just a single conjugacy class of a primitive word.

The Burrows-Wheeler transform  $BW(w)$  of a word  $w$  is a certain word of length  $|w|$  over the same alphabet that is used as a tool of lossless data compression.

## Definitions

Let  $A = \{a_0 < a_1 < \dots\}$  be a finite alphabet,  $A^+(A^*)$  the free semigroup (resp. free monoid) over  $A$ . A word  $w \in A^+$  is *primitive* if  $w$  is not a power of some other word. A *conjugate* of  $w$  is a word  $w' = vu$  where  $w = uv$ . The first word in the lexicographic order of the conjugacy class of a primitive word  $w$  is the *Lyndon word* of the *necklace* (conjugacy class)  $n(w)$  of  $w$ .

The BW transform of two conjugate words is the same. Inverting BW transforms can be done in linear time but in general inversion gives rise to multisets of necklaces rather than just a single conjugacy class of a primitive word.

The Burrows-Wheeler transform  $BW(w)$  of a word  $w$  is a certain word of length  $|w|$  over the same alphabet that is used as a tool of lossless data compression.

## Definitions

Let  $A = \{a_0 < a_1 < \dots\}$  be a finite alphabet,  $A^+(A^*)$  the free semigroup (resp. free monoid) over  $A$ . A word  $w \in A^+$  is *primitive* if  $w$  is not a power of some other word. A *conjugate* of  $w$  is a word  $w' = vu$  where  $w = uv$ . The first word in the lexicographic order of the conjugacy class of a primitive word  $w$  is the *Lyndon word* of the *necklace* (conjugacy class)  $n(w)$  of  $w$ .

The BW transform of two conjugate words is the same. Inverting BW transforms can be done in linear time but in general inversion gives rise to multisets of necklaces rather than just a single conjugacy class of a primitive word.

## Example

$M = \{aab, ab, abb\}$  ( $M$  is a multi-set of necklaces, with each necklace represented by its Lyndon word). Form the dictionary of all words  $u^{\frac{l}{|u|}}$  where  $u \in n(v)$  ( $v \in M$ ) and  $l$  is the least common multiple of the lengths of the  $n(v)$ : here  $l = 3 \times 2 = 6$ .

a	a	<u>b</u>	a	a	b
a	b	<u>a</u>	a	b	a
a	<u>b</u>	a	b	a	b
a	b	<u>b</u>	a	b	b
b	a	<u>a</u>	b	a	a
b	<u>a</u>	b	a	b	a
b	a	<u>b</u>	b	a	b
b	b	<u>a</u>	b	b	a

Then  $BW(M) = w = babbaaba$ , the final column of the table. Note the roots of the words are not in alphabetical order:  $baa$  precedes  $ba$  but the Lyndon roots are  $aab < ab < abb$ .

## Example

$M = \{aab, ab, abb\}$  ( $M$  is a multi-set of necklaces, with each necklace represented by its Lyndon word). Form the dictionary of all words  $u^{\lfloor l/|u|}$  where  $u \in n(v)$  ( $v \in M$ ) and  $l$  is the least common multiple of the lengths of the  $n(v)$ : here  $l = 3 \times 2 = 6$ .

a	a	<u>b</u>	a	a	<b>b</b>
a	b	<u>a</u>	a	b	<b>a</b>
a	<u>b</u>	a	b	a	<b>b</b>
a	b	<u>b</u>	a	b	<b>b</b>
b	a	<u>a</u>	b	a	<b>a</b>
b	<u>a</u>	b	a	b	<b>a</b>
b	a	<u>b</u>	b	a	<b>b</b>
b	b	<u>a</u>	b	b	<b>a</b>

Then  $BW(M) = w = babbaaba$ , the final column of the table. Note the roots of the words are not in alphabetical order:  $baa$  precedes  $ba$  but the Lyndon roots are  $aab < ab < abb$ .

## Example

$M = \{aab, ab, abb\}$  ( $M$  is a multi-set of necklaces, with each necklace represented by its Lyndon word). Form the dictionary of all words  $u^{\frac{l}{|u|}}$  where  $u \in n(v)$  ( $v \in M$ ) and  $l$  is the least common multiple of the lengths of the  $n(v)$ : here  $l = 3 \times 2 = 6$ .

a	a	<u>b</u>	a	a	<b>b</b>
a	b	<u>a</u>	a	b	<b>a</b>
a	<u>b</u>	a	b	a	<b>b</b>
a	b	<u>b</u>	a	b	<b>b</b>
b	a	<u>a</u>	b	a	<b>a</b>
b	<u>a</u>	b	a	b	<b>a</b>
b	a	<u>b</u>	b	a	<b>b</b>
b	b	<u>a</u>	b	b	<b>a</b>

Then  $BW(M) = w = babbaaba$ , the final column of the table. Note the roots of the words are not in alphabetical order:  $baa$  precedes  $ba$  but the Lyndon roots are  $aab < ab < abb$ .

## Definition

The Standard Permutation,  $\pi(w)$  is formed by mapping each letter's set of positions in the first column to those in the last while preserving order:

$$\pi(w) = \pi_a(w) \cup \pi_b(w) = \begin{pmatrix} 0123 \\ 1457 \end{pmatrix} \cup \begin{pmatrix} 4567 \\ 0236 \end{pmatrix} = (014)(25)(376)$$

To recover  $M$ , take each cycle in  $\pi(w)$  by replacing each integer  $m$  by the letter  $c$  where  $m \in \text{dom} \pi_c$ : in this case we write  $a$  whenever we see a number from  $\{0, 1, 2, 3\}$  and  $b$  otherwise:

$$M = \{aab, ab, abb\}.$$

Burrows and Wheeler [1] developed the transform to code (single) words; this generalized form, which allows inversion of an arbitrary word, first appears in Mantaci et. al [8].



## Definition

The Standard Permutation,  $\pi(w)$  is formed by mapping each letter's set of positions in the first column to those in the last while preserving order:

$$\pi(w) = \pi_a(w) \cup \pi_b(w) = \begin{pmatrix} 0123 \\ 1457 \end{pmatrix} \cup \begin{pmatrix} 4567 \\ 0236 \end{pmatrix} = (014)(25)(376)$$

To recover  $M$ , take each cycle in  $\pi(w)$  by replacing each integer  $m$  by the letter  $c$  where  $m \in \text{dom} \pi_c$ : in this case we write  $a$  whenever we see a number from  $\{0, 1, 2, 3\}$  and  $b$  otherwise:

$$M = \{aab, ab, abb\}.$$

Burrows and Wheeler [1] developed the transform to code (single) words; this generalized form, which allows inversion of an arbitrary word, first appears in Mantaci et. al [8].

## Definition

The Standard Permutation,  $\pi(w)$  is formed by mapping each letter's set of positions in the first column to those in the last while preserving order:

$$\pi(w) = \pi_a(w) \cup \pi_b(w) = \begin{pmatrix} 0123 \\ 1457 \end{pmatrix} \cup \begin{pmatrix} 4567 \\ 0236 \end{pmatrix} = (014)(25)(376)$$

To recover  $M$ , take each cycle in  $\pi(w)$  by replacing each integer  $m$  by the letter  $c$  where  $m \in \text{dom}\pi_c$ : in this case we write  $a$  whenever we see a number from  $\{0, 1, 2, 3\}$  and  $b$  otherwise:

$$M = \{aab, ab, abb\}.$$

Burrows and Wheeler [1] developed the transform to code (single) words; this generalized form, which allows inversion of an arbitrary word, first appears in Mantaci et. al [8].

The permutation  $\pi$  acts to map each column of the table onto its predecessor (in cyclic order) that is:  $a_{ij} = a_{i\pi.j-1}$ .

Hence given  $w = w(M) \in A^*$ , we may construct  $\pi(w)$  and recover the table. We pass from  $M$  to  $w$  and back to  $M$  via this common table: for a given multiset  $M$  we have  $T(M) = T = T(w(M))$ .

Conversely, beginning with a given word  $w$ , we may find  $\pi(w)$  and form a table  $T(w)$ . The roots of the rows of  $T(w)$  form the necklaces of a multiset  $M(w)$  and  $T(M(w)) = T = T(w)$ .

The permutation  $\pi$  acts to map each column of the table onto its predecessor (in cyclic order) that is:  $a_{ij} = a_{i\pi.j-1}$ .

Hence given  $w = w(M) \in A^*$ , we may construct  $\pi(w)$  and recover the table. We pass from  $M$  to  $w$  and back to  $M$  via this common table: for a given multiset  $M$  we have  $T(M) = T = T(w(M))$ .

Conversely, beginning with a given word  $w$ , we may find  $\pi(w)$  and form a table  $T(w)$ . The roots of the rows of  $T(w)$  form the necklaces of a multiset  $M(w)$  and  $T(M(w)) = T = T(w)$ .

The permutation  $\pi$  acts to map each column of the table onto its predecessor (in cyclic order) that is:  $a_{ij} = a_{i\pi.j-1}$ .

Hence given  $w = w(M) \in A^*$ , we may construct  $\pi(w)$  and recover the table. We pass from  $M$  to  $w$  and back to  $M$  via this common table: for a given multiset  $M$  we have  $T(M) = T = T(w(M))$ .

Conversely, beginning with a given word  $w$ , we may find  $\pi(w)$  and form a table  $T(w)$ . The roots of the rows of  $T(w)$  form the necklaces of a multiset  $M(w)$  and  $T(M(w)) = T = T(w)$ .

## Theorem

[Gessel & Reutenauer][4] [See also 8,5] The inverse Burrows-Wheeler map  $BW^{-1} : A^* \rightarrow \mathcal{M}$ , is a bijection onto the collection  $\mathcal{M}$  of all multi-sets of necklaces over  $A$ .

# Semigroup of the permutation

Any permutation  $\pi$  on the ordered set  $[n] = \{0 < 1 < \dots < n-1\}$  can be expressed as the disjoint union of order-preserving partial one-to-one mappings on  $[n]$  and there is a unique maximal such decomposition in which the domains of the mappings are intervals.

## Example

Take  $n = 13$  and  $\pi = (0196385210111274)$ :

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 9 & 10 & 8 & 0 & 2 & 3 & 4 & 5 & 6 & 1 & 1 & 2 & 7 \end{pmatrix}$$

the right hand boundary of the intervals correspond to the descents in the number sequence formed by the range as shown above.

$$\pi_a = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 9 & 10 \end{pmatrix} \pi_b = \begin{pmatrix} 3 \\ 8 \end{pmatrix} \pi_c = \begin{pmatrix} 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 0 & 2 & 3 & 4 & 5 & 6 & 1 & 1 & 2 \end{pmatrix} \pi_d = \begin{pmatrix} 12 \\ 7 \end{pmatrix}.$$

# Semigroup of the permutation

Any permutation  $\pi$  on the ordered set  $[n] = \{0 < 1 < \dots < n-1\}$  can be expressed as the disjoint union of order-preserving partial one-to-one mappings on  $[n]$  and there is a unique maximal such decomposition in which the domains of the mappings are intervals.

## Example

Take  $n = 13$  and  $\pi = (0196385210111274)$ :

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 9 & 10 & 8 & 0 & 2 & 3 & 4 & 5 & 6 & 1 & 1 & 2 & 7 \end{pmatrix}$$

the right hand boundary of the intervals correspond to the descents in the number sequence formed by the range as shown above.

$$\pi_a = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 9 & 10 \end{pmatrix} \pi_b = \begin{pmatrix} 3 \\ 8 \end{pmatrix} \pi_c = \begin{pmatrix} 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 0 & 2 & 3 & 4 & 5 & 6 & 1 & 1 & 2 \end{pmatrix} \pi_d = \begin{pmatrix} 12 \\ 7 \end{pmatrix}.$$



# Semigroup of the permutation

Any permutation  $\pi$  on the ordered set  $[n] = \{0 < 1 < \dots < n-1\}$  can be expressed as the disjoint union of order-preserving partial one-to-one mappings on  $[n]$  and there is a unique maximal such decomposition in which the domains of the mappings are intervals.

## Example

Take  $n = 13$  and  $\pi = (0196385210111274)$ :

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 9 & 10 & 8 & 0 & 2 & 3 & 4 & 5 & 6 & 1 & 1 & 2 & 7 \end{pmatrix}$$

the right hand boundary of the intervals correspond to the descents in the number sequence formed by the range as shown above.

$$\pi_a = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 9 & 10 \end{pmatrix} \pi_b = \begin{pmatrix} 3 \\ 8 \end{pmatrix} \pi_c = \begin{pmatrix} 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 0 & 2 & 3 & 4 & 5 & 6 & 1 & 1 & 2 \end{pmatrix} \pi_d = \begin{pmatrix} 12 \\ 7 \end{pmatrix}.$$

# Semigroup of the permutation

Any permutation  $\pi$  on the ordered set  $[n] = \{0 < 1 < \dots < n-1\}$  can be expressed as the disjoint union of order-preserving partial one-to-one mappings on  $[n]$  and there is a unique maximal such decomposition in which the domains of the mappings are intervals.

## Example

Take  $n = 13$  and  $\pi = (0196385210111274)$ :

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 1 & 9 & 10 & 8 & 0 & 2 & 3 & 4 & 5 & 6 & 1 & 1 & 2 & 7 \end{pmatrix}$$

the right hand boundary of the intervals correspond to the descents in the number sequence formed by the range as shown above.

$$\pi_a = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 9 & 10 \end{pmatrix} \pi_b = \begin{pmatrix} 3 \\ 8 \end{pmatrix} \pi_c = \begin{pmatrix} 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 0 & 2 & 3 & 4 & 5 & 6 & 1 & 1 & 2 \end{pmatrix} \pi_d = \begin{pmatrix} 12 \\ 7 \end{pmatrix}.$$

These mappings generate a subsemigroup  $S$  of  $POI_n$ , the inverse monoid of all partial one-to-one and order-preserving mappings on the chain  $[n]$ , which is an *aperiodic* semigroup (all subgroups are trivial).

I stumbled upon the BW schema by regarding the action of cycling through the conjugates of a primitive word as being induced by the action of the relevant 'moving' letter, which then represented an order-preserving one-to-one partial mapping on the dictionary of conjugates.

I solved problem of when two words generate isomorphic semigroups.

These mappings generate a subsemigroup  $S$  of  $POI_n$ , the inverse monoid of all partial one-to-one and order-preserving mappings on the chain  $[n]$ , which is an *aperiodic* semigroup (all subgroups are trivial).

I stumbled upon the BW schema by regarding the action of cycling through the conjugates of a primitive word as being induced by the action of the relevant 'moving' letter, which then represented an order-preserving one-to-one partial mapping on the dictionary of conjugates.

I solved problem of when two words generate isomorphic semigroups.

These mappings generate a subsemigroup  $S$  of  $POI_n$ , the inverse monoid of all partial one-to-one and order-preserving mappings on the chain  $[n]$ , which is an *aperiodic* semigroup (all subgroups are trivial).

I stumbled upon the BW schema by regarding the action of cycling through the conjugates of a primitive word as being induced by the action of the relevant 'moving' letter, which then represented an order-preserving one-to-one partial mapping on the dictionary of conjugates.

I solved problem of when two words generate isomorphic semigroups.

However  $S$  can be realised without reference to the notion of mappings, or of order, as  $S$  is the *syntactic semigroup* of the cyclic semigroup generated by the seed word  $u$ .

### Definition

Let  $L \subseteq A^+$ , the free semigroup on a finite set  $A$ . The *syntactic congruence*  $\sim$  on  $A^+$  is defined by

$$u \sim v \Leftrightarrow (puv \in L \leftrightarrow pvq) \forall p, q \in A^*.$$

However  $S$  can be realised without reference to the notion of mappings, or of order, as  $S$  is the *syntactic semigroup* of the cyclic semigroup generated by the seed word  $u$ .

### Definition

Let  $L \subseteq A^+$ , the free semigroup on a finite set  $A$ . The *syntactic congruence*  $\sim$  on  $A^+$  is defined by

$$u \sim v \Leftrightarrow (puv \in L \leftrightarrow pvq) \forall p, q \in A^*.$$

## Definition

The semigroup  $S_L = A^+ / \sim$  is called the *syntactic semigroup of  $L$* :  $\sim$  is the coarsest congruence that has its classes entirely contained in  $L$  and  $S_L$  is finite if and only if  $L$  is a *regular language*, which is one that is recognized by a finite state automaton.

## Theorem

*The semigroup  $S(u)$  generated by the letters acting by conjugation on the necklace of a primitive word  $u$  is isomorphic to the syntactic semigroup  $A^+ / \sim$  of the language  $L = \langle u \rangle$ , the monogenic subsemigroup of  $A^+$  generated by  $u$ .*



## Definition

The semigroup  $S_L = A^+ / \sim$  is called the *syntactic semigroup of  $L$* :  $\sim$  is the coarsest congruence that has its classes entirely contained in  $L$  and  $S_L$  is finite if and only if  $L$  is a *regular language*, which is one that is recognized by a finite state automaton.

## Theorem

*The semigroup  $S(u)$  generated by the letters acting by conjugation on the necklace of a primitive word  $u$  is isomorphic to the syntactic semigroup  $A^+ / \sim$  of the language  $L = \langle u \rangle$ , the monogenic subsemigroup of  $A^+$  generated by  $u$ .*

## Definition

A binary de Bruijn word  $w$  of length  $2^n$  is one that has every member of  $A^n$  among its cyclic factors ( $A = \{a, b\}$ ).

## Example

A de Bruijn word of length  $2^4 = 16$  is  $w = \text{aaaa babb bbab aabb}$ .

The first  $n$  columns of the conjugate table  $T$  of a de Bruijn word  $w$  consists of the dictionary of  $A^n$ . Let  $u \in A^{n-1}$  so that  $ua, ub$  represent the beginning of two successive rows of  $T$ . If these rows both ended in the same letter  $c$ , then the cyclic conjugates of these two rows would give two distinct rows that each began with the same  $n$ -factor,  $cu$ .

## Definition

A binary de Bruijn word  $w$  of length  $2^n$  is one that has every member of  $A^n$  among its cyclic factors ( $A = \{a, b\}$ ).

## Example

A de Bruijn word of length  $2^4 = 16$  is  $w = \text{aaaa babb bbab aabb}$ .

The first  $n$  columns of the conjugate table  $T$  of a de Bruijn word  $w$  consists of the dictionary of  $A^n$ . Let  $u \in A^{n-1}$  so that  $ua, ub$  represent the beginning of two successive rows of  $T$ . If these rows both ended in the same letter  $c$ , then the cyclic conjugates of these two rows would give two distinct rows that each began with the same  $n$ -factor,  $cu$ .

## Definition

A binary de Bruijn word  $w$  of length  $2^n$  is one that has every member of  $A^n$  among its cyclic factors ( $A = \{a, b\}$ ).

## Example

A de Bruijn word of length  $2^4 = 16$  is  $w = \text{aaaa babb bbab aabb}$ .

The first  $n$  columns of the conjugate table  $T$  of a de Bruijn word  $w$  consists of the dictionary of  $A^n$ . Let  $u \in A^{n-1}$  so that  $ua, ub$  represent the beginning of two successive rows of  $T$ . If these rows both ended in the same letter  $c$ , then the cyclic conjugates of these two rows would give two distinct rows that each began with the same  $n$ -factor,  $cu$ .

This is impossible for a de Bruijn word  $w$ , and indeed, writing  $\alpha = ab$  and  $\beta = ba$ , we may prove:

### Theorem

*The de Bruijn words are the words of the cycles that lie among the inverse Burrows-Wheeler transforms of  $\{\alpha, \beta\}^{2^{n-1}}$ .*

### Example

Take  $v = \beta^4 \alpha \beta^3 = ba \cdot ba \cdot ba \cdot ba \cdot ab \cdot ba \cdot ba \cdot ba$ . The successive images of the members of  $X_{16}$  under  $\pi_v$  are:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 13 & 5 & 7 & 8 & 11 & 13 & 15 & 0 & 2 & 4 & 6 & 9 & 10 & 12 & 14 & \end{pmatrix}$$

which gives that  $\pi$  is the cycle

$\pi = (0137151412925116131048)$ , yielding the Lyndon de Bruijn word

$$w = aaaa bbbb aaba bbab$$

This is impossible for a de Bruijn word  $w$ , and indeed, writing  $\alpha = ab$  and  $\beta = ba$ , we may prove:

### Theorem

*The de Bruijn words are the words of the cycles that lie among the inverse Burrows-Wheeler transforms of  $\{\alpha, \beta\}^{2^{n-1}}$ .*

### Example

Take  $v = \beta^4 \alpha \beta^3 = ba \cdot ba \cdot ba \cdot ba \cdot ab \cdot ba \cdot ba \cdot ba$ . The successive images of the members of  $X_{16}$  under  $\pi_v$  are:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 1 & 3 & 5 & 7 & 8 & 11 & 13 & 15 & 0 & 2 & 4 & 6 & 9 & 10 & 12 & 14 \end{pmatrix}$$

which gives that  $\pi$  is the cycle

$\pi = (0137151412925116131048)$ , yielding the Lyndon de Bruijn word

$$w = aaaa bbbb aaba bbab$$

This is impossible for a de Bruijn word  $w$ , and indeed, writing  $\alpha = ab$  and  $\beta = ba$ , we may prove:

### Theorem

The de Bruijn words are the words of the cycles that lie among the inverse Burrows-Wheeler transforms of  $\{\alpha, \beta\}^{2^{n-1}}$ .

### Example

Take  $v = \beta^4 \alpha \beta^3 = ba \cdot ba \cdot ba \cdot ba \cdot ab \cdot ba \cdot ba \cdot ba$ . The successive images of the members of  $X_{16}$  under  $\pi_v$  are:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 1 & 3 & 5 & 7 & 8 & 11 & 13 & 15 & 0 & 2 & 4 & 6 & 9 & 10 & 12 & 14 \end{pmatrix}$$

which gives that  $\pi$  is the cycle

$\pi = (0137151412925116131048)$ , yielding the Lyndon de Bruijn word

$$w = aaaa bbbb aaba bbab$$

This is impossible for a de Bruijn word  $w$ , and indeed, writing  $\alpha = ab$  and  $\beta = ba$ , we may prove:

### Theorem

*The de Bruijn words are the words of the cycles that lie among the inverse Burrows-Wheeler transforms of  $\{\alpha, \beta\}^{2^{n-1}}$ .*

### Example

Take  $v = \beta^4 \alpha \beta^3 = ba \cdot ba \cdot ba \cdot ba \cdot ab \cdot ba \cdot ba \cdot ba$ . The successive images of the members of  $X_{16}$  under  $\pi_v$  are:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 13 & 5 & 7 & 8 & 11 & 13 & 15 & 0 & 2 & 4 & 6 & 9 & 10 & 12 & 14 & \end{pmatrix}$$

which gives that  $\pi$  is the cycle

$\pi = (0137151412925116131048)$ , yielding the Lyndon de Bruijn word

$$w = aaaa bbbb aaba bbab$$



This is impossible for a de Bruijn word  $w$ , and indeed, writing  $\alpha = ab$  and  $\beta = ba$ , we may prove:

### Theorem

*The de Bruijn words are the words of the cycles that lie among the inverse Burrows-Wheeler transforms of  $\{\alpha, \beta\}^{2^{n-1}}$ .*

### Example

Take  $v = \beta^4 \alpha \beta^3 = ba \cdot ba \cdot ba \cdot ba \cdot ab \cdot ba \cdot ba \cdot ba$ . The successive images of the members of  $X_{16}$  under  $\pi_v$  are:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 13 & 5 & 7 & 8 & 11 & 13 & 15 & 0 & 2 & 4 & 6 & 9 & 10 & 12 & 14 & \end{pmatrix}$$

which gives that  $\pi$  is the cycle

$\pi = (0137151412925116131048)$ , yielding the Lyndon de Bruijn word

$$w = aaaa bbbb aaba bbab$$

This is impossible for a de Bruijn word  $w$ , and indeed, writing  $\alpha = ab$  and  $\beta = ba$ , we may prove:

### Theorem

*The de Bruijn words are the words of the cycles that lie among the inverse Burrows-Wheeler transforms of  $\{\alpha, \beta\}^{2^{n-1}}$ .*

### Example

Take  $v = \beta^4 \alpha \beta^3 = ba \cdot ba \cdot ba \cdot ba \cdot ab \cdot ba \cdot ba \cdot ba$ . The successive images of the members of  $X_{16}$  under  $\pi_v$  are:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 \\ 13 & 5 & 7 & 8 & 11 & 13 & 15 & 0 & 2 & 4 & 6 & 9 & 10 & 12 & 14 & \end{pmatrix}$$

which gives that  $\pi$  is the cycle

$\pi = (0137151412925116131048)$ , yielding the Lyndon de Bruijn word

$$w = aaaa bbbb aaba bbab$$

More generally, take  $A = \{0 < 1 < \dots < k - 1\} = [k]$ .

### Definition

A set  $M$  of necklaces  $\{n_i\}_{1 \leq i \leq t}$  is a *de Bruijn set* of span  $n$  if  $|n_1| + |n_2| + \dots + |n_t| = k^n$  and every  $w \in A^n$  can be read (necessarily exactly once) among the cyclic conjugates of the necklaces.

Let  $G = \{i_1 i_2 \dots i_k : \{i_1, i_2, \dots, i_k\} = [k]\}$  (that is,  $G$  consists of all  $k!$  products of distinct members of the  $k$ -ary alphabet  $A$ ).

### Theorem

*The set of all BW transforms of de Bruijn sets  $M$  of span  $n$  over  $A$  is  $G^{k^{n-1}}$ .*

More generally, take  $A = \{0 < 1 < \dots < k - 1\} = [k]$ .

### Definition

A set  $M$  of necklaces  $\{n_i\}_{1 \leq i \leq t}$  is a *de Bruijn set* of span  $n$  if  $|n_1| + |n_2| + \dots + |n_t| = k^n$  and every  $w \in A^n$  can be read (necessarily exactly once) among the cyclic conjugates of the necklaces.

Let  $G = \{i_1 i_2 \dots i_k : \{i_1, i_2, \dots, i_k\} = [k]\} = [k]!$  (that is,  $G$  consists of all  $k!$  products of distinct members of the  $k$ -ary alphabet  $A$ ).

### Theorem

*The set of all BW transforms of de Bruijn sets  $M$  of span  $n$  over  $A$  is  $G^{k^{n-1}}$ .*

More generally, take  $A = \{0 < 1 < \dots < k - 1\} = [k]$ .

### Definition

A set  $M$  of necklaces  $\{n_i\}_{1 \leq i \leq t}$  is a *de Bruijn set* of span  $n$  if  $|n_1| + |n_2| + \dots + |n_t| = k^n$  and every  $w \in A^n$  can be read (necessarily exactly once) among the cyclic conjugates of the necklaces.

Let  $G = \{i_1 i_2 \dots i_k : \{i_1, i_2, \dots, i_k\} = [k]\} = [k]!$  (that is,  $G$  consists of all  $k!$  products of distinct members of the  $k$ -ary alphabet  $A$ ).

### Theorem

*The set of all BW transforms of de Bruijn sets  $M$  of span  $n$  over  $A$  is  $G^{k^{n-1}}$ .*

## Example

Take  $A = \{a < b\}$  and put

$v = \beta\alpha^2\beta^2\alpha^2\beta = ba \cdot ab \cdot ab \cdot ba \cdot ba \cdot ab \cdot ab \cdot ba$  and we obtain:

$$\pi(v) = (0124937151413116128)(510);$$

The corresponding set of Lyndon words is

$\{aaaa \cdot baab \cdot bbba \cdot bb, ab\}$ , the cyclic 4-factors of which are all the  $2^4 = 16$  words of  $A^4$ , with  $\{abab, baba\}$  arising from the necklace with Lyndon word  $ab$ .

## Example

Take  $A = \{a < b\}$  and put

$v = \beta\alpha^2\beta^2\alpha^2\beta = ba \cdot ab \cdot ab \cdot ba \cdot ba \cdot ab \cdot ab \cdot ba$  and we obtain:

$$\pi(v) = (0124937151413116128)(510);$$

The corresponding set of Lyndon words is

$\{aaaa \cdot baab \cdot bbba \cdot bb, ab\}$ , the cyclic 4-factors of which are all the  $2^4 = 16$  words of  $A^4$ , with  $\{abab, baba\}$  arising from the necklace with Lyndon word  $ab$ .

## Example

Take  $A = \{a < b\}$  and put

$v = \beta\alpha^2\beta^2\alpha^2\beta = ba \cdot ab \cdot ab \cdot ba \cdot ba \cdot ab \cdot ab \cdot ba$  and we obtain:

$$\pi(v) = (0124937151413116128)(510);$$

The corresponding set of Lyndon words is

$\{aaaa \cdot baab \cdot bbba \cdot bb, ab\}$ , the cyclic 4-factors of which are all the  $2^4 = 16$  words of  $A^4$ , with  $\{abab, baba\}$  arising from the necklace with Lyndon word  $ab$ .



An interesting special case is where  $v = \alpha^{k^{n-1}}$ , where  $\alpha = (12 \cdots k-1)$

## Theorem

*$BW^{-1}(v)$  is the set of necklaces of Lyndon words of length dividing  $n$ . The lexicographic concatenation of these words (which appear in the table in dictionary order) is the first de Bruijn word of span  $n$  in the lexicographic order.*

The second statement follows from a theorem of Frederickson and Maiorana [3]; also there is a proof by Moreno [9].

An interesting special case is where  $v = \alpha^{k^{n-1}}$ , where  $\alpha = (12 \cdots k-1)$

## Theorem

*$BW^{-1}(v)$  is the set of necklaces of Lyndon words of length dividing  $n$ . The lexicographic concatenation of these words (which appear in the table in dictionary order) is the first de Bruijn word of span  $n$  in the lexicographic order.*

The second statement follows from a theorem of Frederickson and Maiorana [3]; also there is a proof by Moreno [9].

An interesting special case is where  $v = \alpha^{k^{n-1}}$ , where  $\alpha = (12 \cdots k-1)$

## Theorem

*$BW^{-1}(v)$  is the set of necklaces of Lyndon words of length dividing  $n$ . The lexicographic concatenation of these words (which appear in the table in dictionary order) is the first de Bruijn word of span  $n$  in the lexicographic order.*

The second statement follows from a theorem of Frederickson and Maiorana [3]; also there is a proof by Moreno [9].

## Example

Take  $k = 2$ ,  $n = 5$  and using  $\alpha = ab$  so that  $v = (ab)^{16}$ . Then

$$\pi(v) = (0)(124816)(36122417)(51020918)(714282519) \cdot$$

$$(1122132621)(1530292723)(31).$$

$$BW^{-1}(v) = a \cdot aaaab \cdot aaabb \cdot aabab \cdot aabbb \cdot ababb \cdot abbbb \cdot b$$

## Example

Take  $k = 2$ ,  $n = 5$  and using  $\alpha = ab$  so that  $v = (ab)^{16}$ . Then

$$\pi(v) = (0)(124816)(36122417)(51020918)(714282519) \cdot \\ (1122132621)(1530292723)(31).$$

$$BW^{-1}(v) = a \cdot aaaab \cdot aaabb \cdot aabab \cdot aabbb \cdot ababb \cdot abbbb \cdot b$$

## Example

Take  $k = 2$ ,  $n = 5$  and using  $\alpha = ab$  so that  $v = (ab)^{16}$ . Then

$$\pi(v) = (0)(124816)(36122417)(51020918)(714282519) \cdot$$

$$(1122132621)(1530292723)(31).$$

$$BW^{-1}(v) = a \cdot aaaab \cdot aaabb \cdot aabab \cdot aabbb \cdot ababb \cdot abbbb \cdot b$$

## Example

Put  $k = n = 3$  so that  $\alpha = abc$  say and  $v = \alpha^{k^{n-1}} = (abc)^9$ . We find that  $\pi(v) =$

$$\pi(v) = (0)(139)(2618)(41210)(51519)(72111) \cdot$$

$$(82420)(13)(141622)(172523)(26);$$

$$BW^{-1}(\alpha^9) = a \cdot aab \cdot aac \cdot abb \cdot abc \cdot acb \cdot acc \cdot b \cdot bbc \cdot bcc \cdot c.$$

## Example

Put  $k = n = 3$  so that  $\alpha = abc$  say and  $v = \alpha^{k^{n-1}} = (abc)^9$ . We find that  $\pi(v) =$

$$\pi(v) = (0)(139)(2618)(41210)(51519)(72111) \cdot$$

$$(82420)(13)(141622)(172523)(26);$$

$$BW^{-1}(\alpha^9) = a \cdot aab \cdot aac \cdot abb \cdot abc \cdot acb \cdot acc \cdot b \cdot bbc \cdot bcc \cdot c.$$



## Example

Put  $k = n = 3$  so that  $\alpha = abc$  say and  $v = \alpha^{k^{n-1}} = (abc)^9$ . We find that  $\pi(v) =$

$$\pi(v) = (0)(139)(2618)(41210)(51519)(72111) \cdot$$

$$(82420)(13)(141622)(172523)(26);$$

$$BW^{-1}(\alpha^9) = a \cdot aab \cdot aac \cdot abb \cdot abc \cdot acb \cdot acc \cdot b \cdot bbc \cdot bcc \cdot c.$$

The number of (not necessarily distinct) factors of a word of length  $n$  (over a  $k$ -letter alphabet) is  $\binom{n+1}{2} = \frac{1}{2}n(n+1)$ .

We may count the number of *distinct* factors  $f_w$  of  $w \in A^n$ ; for long words, there must be repeats of short factors. This leads to an upper bound for the function:

$$f(n) = \max\{f_w : w \in A^n\}$$

Since factors of de Bruijn words of span  $m$  have no repeats of their factors of length at least  $m$  (as their  $m$ -prefixes are distinct) we can derive the form of a lower bound for  $f(n)$  yielding:

$$\frac{1}{2}n^2 - f(n) = O(n \log n).$$

The number of (not necessarily distinct) factors of a word of length  $n$  (over a  $k$ -letter alphabet) is  $\binom{n+1}{2} = \frac{1}{2}n(n+1)$ .

We may count the number of *distinct* factors  $f_w$  of  $w \in A^n$ ; for long words, there must be repeats of short factors. This leads to an upper bound for the function:

$$f(n) = \max\{f_w : w \in A^n\}$$

Since factors of de Bruijn words of span  $m$  have no repeats of their factors of length at least  $m$  (as their  $m$ -prefixes are distinct) we can derive the form of a lower bound for  $f(n)$  yielding:

$$\frac{1}{2}n^2 - f(n) = O(n \log n).$$

The number of (not necessarily distinct) factors of a word of length  $n$  (over a  $k$ -letter alphabet) is  $\binom{n+1}{2} = \frac{1}{2}n(n+1)$ .

We may count the number of *distinct* factors  $f_w$  of  $w \in A^n$ ; for long words, there must be repeats of short factors. This leads to an upper bound for the function:

$$f(n) = \max\{f_w : w \in A^n\}$$

Since factors of de Bruijn words of span  $m$  have no repeats of their factors of length at least  $m$  (as their  $m$ -prefixes are distinct) we can derive the form of a lower bound for  $f(n)$  yielding:

$$\frac{1}{2}n^2 - f(n) = O(n \log n).$$

The number of (not necessarily distinct) factors of a word of length  $n$  (over a  $k$ -letter alphabet) is  $\binom{n+1}{2} = \frac{1}{2}n(n+1)$ .

We may count the number of *distinct* factors  $f_w$  of  $w \in A^n$ ; for long words, there must be repeats of short factors. This leads to an upper bound for the function:

$$f(n) = \max\{f_w : w \in A^n\}$$

Since factors of de Bruijn words of span  $m$  have no repeats of their factors of length at least  $m$  (as their  $m$ -prefixes are distinct) we can derive the form of a lower bound for  $f(n)$  yielding:

$$\frac{1}{2}n^2 - f(n) = O(n \log n).$$







The number of (not necessarily distinct) factors of a word of length  $n$  (over a  $k$ -letter alphabet) is  $\binom{n+1}{2} = \frac{1}{2}n(n+1)$ .




We may count the number of *distinct* factors  $f_w$  of  $w \in A^n$ ; for long words, there must be repeats of short factors. This leads to an upper bound for the function:

$$f(n) = \max\{f_w : w \in A^n\}$$

Since factors of de Bruijn words of span  $m$  have no repeats of their factors of length at least  $m$  (as their  $m$ -prefixes are distinct) we can derive the form of a lower bound for  $f(n)$  yielding:

$$\frac{1}{2}n^2 - f(n) = O(n \log n).$$

-  Burrows M. D.J. Wheeler, *A block sorting data compression algorithm*, Technical Report, DIGITAL System Center, 1994.
-  Crochemore, M.J. Desarmenien and D. Perrin, *A note on the Burrows-Wheeler transformation*, Theoretical Computer Science, **Vol. 332** Issue 1-3, February, 2005.
-  Fredericksen H., J Maiorana, *Necklaces of beads in  $k$  colors and  $k$ -ary de bruijn sequences*, Discrete Maths. **23** (1978) 207-210.
-  Gessel, I.M., C. Reutenauer, *Counting permutations with a given cycle structure and descent set*, J. of Combinatorial Theory, Series A **64** 189-215.
-  Higgins, P.M., *The semigroup of conjugates of a word*, International Journal of Algebra and Computation, **Vol. 16**, No. 6 (2006), 1015-1029.
-  Higgins, P.M., *Burrows-Wheeler transformations and de Bruijn words*, Theoretical Computer Science, 457 (2012), 128-136.

-  Lothaire, M. 'Combinatorics on Words', Cambridge University Press, (2002).
-  Mantaci, S. A. Restivo, G. Rosone, M. Sciortino, *An extension of the Burrows-Wheeler Transform*, Theoretical Computer Science, 387(3) (2007) 298-415.
-  Moreno E. *On the theorem of Fredericksen and Maiorana about de Bruijn sequences*, *Advances in Applied Mathematics* **33** (2) 413-415.