Dynamic Burrows-Wheeler Transform

Mikaël Salson, Thierry Lecroq, Martine Léonard and Laurent Mouchard

Functional Medical Imaging
LITIS
University of Rouen
France

Algorithm Design Section
Dpt Computer Science
King's College London
England

Stringology 2009
A research workshop of the Israel Science Foundation

Motivation

Recent Advances in High-Throughput Sequencing

Index large dynamic texts (genomes) to speed up approximate pattern matching (myriads of short patterns, at most one error).

Motivation

Recent Advances in High-Throughput Sequencing

Index large dynamic texts (genomes) to speed up approximate pattern matching (myriads of short patterns, at most one error).

Two possible efficient solutions

	FM-Index BWT + Suffix Array Ferragina-Manzini FOCS'00	LZ-Index LZ-factorisation Navarro SPIRE'04
Advantages	Compressed indexes Approximate string matching	
DNA, 50 Mb	19 to 34 MB	44 MB

Motivation

Recent Advances in High-Throughput Sequencing

Index large dynamic texts (genomes) to speed up approximate pattern matching (myriads of short patterns, at most one error).

Two possible efficient solutions

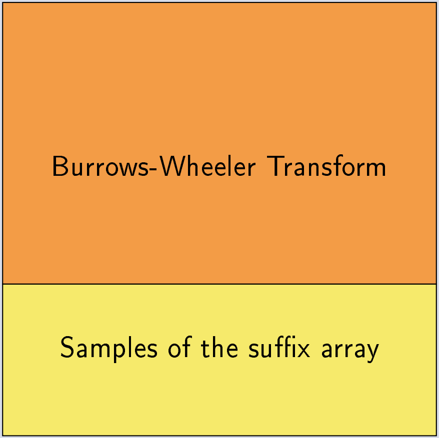
	FM-Index BWT + Suffix Array Ferragina-Manzini FOCS'00	LZ-Index LZ-factorisation Navarro SPIRE'04
Advantages	Compressed indexes Approximate string matching	
DNA, 50 Mb	19 to 34 MB	44 MB

Our main motivation: index large dynamic texts

Single modification in the text → Creation of a new index from scratch?

Our choice

FM-Index: more efficient and probably (?) easier to update



BWT is used to **quickly**:

- retrieve the original text;
- search for patterns.

We need the whole structure;

SA is used to **quickly**:

- find the position of each occurrence of a pattern.

We can sample the SA and recover the missing values.

Burrows-Wheeler Transform of $T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$}$

The big picture

The Burrows-Wheeler Transform is a permutation of a text T that eases the compression of T.

Burrows-Wheeler Transform of $T = \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ C & T & C & T & G & C & \$ \end{array}$

The big picture

The Burrows-Wheeler Transform is a permutation of a text T that eases the compression of T.

unsorted cyclic shifts

	unsorted $T^{[i]}$						
0	C	T	C	T	G	C	$
1	T	C	T	G	C	$	C
2	C	T	G	C	$	C	T
3	T	G	C	$	C	T	C
4	G	C	$	C	T	C	T
5	C	$	C	T	C	T	G
6	$	C	T	C	T	G	C

Burrows-Wheeler Transform of $T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$}$

The big picture

The Burrows-Wheeler Transform is a permutation of a text T that eases the compression of T.

unsorted cyclic shifts

	unsorted $T^{[i]}$						
0	C	T	C	T	G	C	$
1	T	C	T	G	C	$	C
2	C	T	G	C	$	C	T
3	T	G	C	$	C	T	C
4	G	C	$	C	T	C	T
5	C	$	C	T	C	T	G
6	$	C	T	C	T	G	C

sorted cyclic shifts, BWT

	F ↓	sorted $T^{[i]}$					L ↓
0	$	C	T	C	T	G	C
1	C	$	C	T	C	T	G
2	C	T	C	T	G	C	$
3	C	T	G	C	$	C	T
4	G	C	$	C	T	C	T
5	T	C	T	G	C	$	C
6	T	G	C	$	C	T	C
	↓ sorted						↓ unsorted

Burrows-Wheeler Transform of $T = \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ C & T & C & T & G & C & \$ \end{array}$

The big picture

The Burrows-Wheeler Transform is a permutation of a text T that eases the compression of T.

unsorted cyclic shifts

	unsorted $T^{[i]}$						
0	C	T	C	T	G	C	\$
1	T	C	T	G	C	\$	C
2	C	T	G	C	\$	C	T
$\boxed{3}$	T	G	C	\$	C	T	C
4	G	C	\$	C	T	C	T
5	C	\$	C	T	C	T	G
6	\$	C	T	C	T	G	C

sorted cyclic shifts, BWT and suffix array

	F ↓	sorted $T^{[i]}$					*L* ↓	*SA* ↓
0	\$	C	T	C	T	G	C	6
1	C	\$	C	T	C	T	G	5
2	C	T	C	T	G	C	\$	0
3	C	T	G	C	\$	C	T	2
4	G	C	\$	C	T	C	T	4
5	T	C	T	G	C	\$	C	1
6	T	G	C	\$	C	T	C	$\boxed{3}$
	↓ sorted						↓ unsorted	

Burrows-Wheeler Transform of $T = \overset{0}{\text{C}}\,\overset{1}{\text{T}}\,\overset{2}{\text{C}}\,\overset{3}{\text{T}}\,\overset{4}{\text{G}}\,\overset{5}{\text{C}}\,\overset{6}{\$}$

The big picture

The Burrows-Wheeler Transform is a permutation of a text T that eases the compression of T.

unsorted cyclic shifts

	unsorted $T^{[i]}$						
0	C	T	C	T	G	C	\$
1	T	C	T	G	C	\$	C
2	C	T	G	C	\$	C	T
$\boxed{3}$	T	G	C	\$	C	T	C
4	G	C	\$	C	T	C	T
5	C	\$	C	T	C	T	G
6	\$	C	T	C	T	G	C

sorted cyclic shifts, BWT and suffix array

	F ↓		sorted $T^{[i]}$				L ↓	SA ↓	ISA ↓
0	\$	C	T	C	T	G	C	6	2
1	C	\$	C	T	C	T	G	5	5
2	C	T	C	T	G	C	\$	0	3
3	C	T	G	C	\$	C	T	2	6
4	G	C	\$	C	T	C	T	4	4
5	T	C	T	G	C	\$	C	1	1
6	T	G	C	\$	C	T	C	$\boxed{3}$	0
	↓ sorted						↓ unsorted		

Which column do we have to consider?

First column $F = \$$ C C C G T T

- gathers identical symbols;
- easy to compress $\$r_3\text{CG}r_2\text{T}$;
- recovering T from F is impossible.

Which column do we have to consider?

First column $F = \$ \text{ C C C G T T}$

- gathers identical symbols;
- easy to compress $\$r_3\text{CG}r_2\text{T}$;
- recovering T from F is impossible.

Last column $L = \text{C G } \$ \text{ T T C C}$

- has a tendancy to create packets of identical symbols:
- eases compression;
- is recovering T from L possible?

FM-index: a dynamic data structure?

$$T = \begin{array}{ccccccc} 0&1&2&3&4&5&6\\ \mathrm{C}&\mathrm{T}&\mathrm{C}&\mathrm{T}&\mathrm{G}&\mathrm{C}&\$ \end{array} \quad \rightarrow \quad T' = \begin{array}{cccccccc} 0&1&2&3&4&5&6&7\\ \mathrm{C}&\mathrm{T}&G&\mathrm{C}&\mathrm{T}&\mathrm{G}&\mathrm{C}&\$ \end{array}$$

BWT **Easy (Laurent!)**

$$BWT(\mathrm{CTCTGC\$}) = \mathrm{C\ G\ \$\ T\ T\ C\ C}$$

$$BWT(\mathrm{CT}G\mathrm{CTGC\$}) = \mathrm{C\ G}\ G\ \mathrm{\$\ T\ T\ C\ C}$$

FM-index: a dynamic data structure?

$$T = \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \mathrm{C} & \mathrm{T} & \mathrm{C} & \mathrm{T} & \mathrm{G} & \mathrm{C} & \$ \end{array} \quad \rightarrow \quad T' = \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \mathrm{C} & \mathrm{T} & G & \mathrm{C} & \mathrm{T} & \mathrm{G} & \mathrm{C} & \$ \end{array}$$

BWT **Easy (Laurent!)**

$$BWT(\mathrm{CTCTGC\$}) = \mathrm{C\ G\ \$\ T\ T\ C\ C}$$

$$BWT(\mathrm{CT}G\mathrm{CTGC\$}) = \mathrm{C\ G}\ G\ \mathrm{\$\ T\ T\ C\ C}$$

Suffix Array **Hard (Mikaël!)**

$$SA(\mathrm{CTCTGC\$}) = 6\ 5\ 0\ 2\ 4\ 1\ 3$$

$$SA(\mathrm{CT}G\mathrm{CTGC\$}) = 7\ 6\ 3\ 0\ 5\ 2\ 4\ 1$$

FM-index: a dynamic data structure?

$$T = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{T}}\,\overset{4}{\mathrm{G}}\,\overset{5}{\mathrm{C}}\,\overset{6}{\$} \quad \rightarrow \quad T' = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{G}\,\overset{3}{\mathrm{C}}\,\overset{4}{\mathrm{T}}\,\overset{5}{\mathrm{G}}\,\overset{6}{\mathrm{C}}\,\overset{7}{\$}$$

BWT **Easy (Laurent!)**

$$BWT(\mathrm{CTCTGC\$}) = \mathrm{C\ G\ \$\ T\ T\ C\ C}$$
$$BWT(\mathrm{CT}G\mathrm{CTGC\$}) = \mathrm{C\ G\ }G\mathrm{\ \$\ T\ T\ C\ C}$$

Suffix Array **Hard (Mikaël!)**

$$SA(\mathrm{CTCTGC\$}) = 6\ 5\ 0\ 2\ 4\ 1\ 3$$
$$SA(\mathrm{CT}G\mathrm{CTGC\$}) = 7\ 6\ 3\ 0\ 5\ 2\ 4\ 1$$

BWT update looks easier than SA update

- BWT and SA are strongly related;
- BWT update will help for SA update.

FM-index: a dynamic data structure?

$$T = \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \mathrm{C} & \mathrm{T} & \mathrm{C} & \mathrm{T} & \mathrm{G} & \mathrm{C} & \$ \end{array} \quad \rightarrow \quad T' = \begin{array}{cccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \mathrm{C} & \mathrm{T} & G & \mathrm{C} & \mathrm{T} & \mathrm{G} & \mathrm{C} & \$ \end{array}$$

BWT **Easy (Hum, let me think about it!)**

$$BWT(\mathrm{CTCTGC\$}) = \mathrm{C\ G\ \$\ T\ T\ C\ C}$$
$$BWT(\mathrm{CT}G\mathrm{CTGC\$}) = \mathrm{C\ G}\ G\ \mathrm{\$\ T\ T\ C\ C}$$

FM-index: a dynamic data structure?

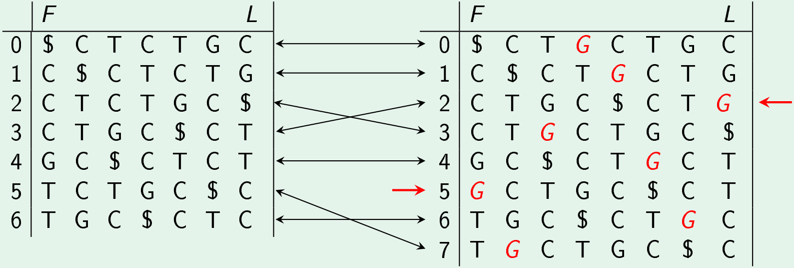
$$T = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{T}}\,\overset{4}{\mathrm{G}}\,\overset{5}{\mathrm{C}}\,\overset{6}{\$} \quad \rightarrow \quad T' = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{G}\,\overset{3}{\mathrm{C}}\,\overset{4}{\mathrm{T}}\,\overset{5}{\mathrm{G}}\,\overset{6}{\mathrm{C}}\,\overset{7}{\$}$$

BWT

Easy (Hum, let me think about it!)

$$BWT(\mathrm{CTCTGC\$}) = \mathrm{C\ G\ \$\ T\ T\ C\ C}$$
$$BWT(\mathrm{CT}G\mathrm{CTGC\$}) = \mathrm{C\ G}\ G\ \mathrm{\$\ T\ T\ C\ C}$$

Conceptual matrices



	F						*L*
0	$	C	T	C	T	G	C
1	C	$	C	T	C	T	G
2	C	T	C	T	G	C	$
3	C	T	G	C	$	C	T
4	G	C	$	C	T	C	T
5	T	C	T	G	C	$	C
6	T	G	C	$	C	T	C

	F							*L*
0	$	C	T	*G*	C	T	G	C
1	C	$	C	T	*G*	C	T	G
2	C	T	G	C	$	C	T	*G*
3	C	T	*G*	C	T	G	C	$
4	G	C	$	C	T	*G*	C	T
5	*G*	C	T	G	C	$	C	T
6	T	G	C	$	C	T	*G*	C
7	T	*G*	C	T	G	C	$	C

From $BWT(T)$=CG\$TTCC back to T=CTCTGC\$

Right-to-left reconstruction of T from L

i	L
0	C
1	G
2	\$
3	T
4	T
5	C
6	C

From $BWT(T)$=CG\$TTCC back to T=CTCTGC\$

Right-to-left reconstruction of T from L

i	F	L
0	\$	C
1	C	G
2	C	\$
3	C	T
4	G	T
5	T	C
6	T	C

Deduce sorted F from unsorted L

From $BWT(T)=\text{CG\$TTCC}$ back to $T=\text{CTCTGC\$}$

Right-to-left reconstruction of T from L

i	F	L	LF
0	\$	C	
1	C	G	
2	C	$\boxed{\$}$	
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

$\boxed{\$}$

From $BWT(T)$=CG$TTCC back to T=CTCTGC$

Right-to-left reconstruction of T from L

i	F	L	LF
0	$	C	
1	C	G	
2	C	$\boxed{\$}$	
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

$

From $BWT(T)=\mathrm{CG\$TTCC}$ back to $T=\mathrm{CTCTGC\$}$

Right-to-left reconstruction of T from L

i	F	L	LF
0	$\boxed{\$}$	C	
1	C	G	
2	C	$\boxed{\$}$	
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

$\$$

From $BWT(T)$=CG$TTCC back to T=CTCTGC$

Right-to-left reconstruction of T from L

i	F	L	LF
0	$	C	
1	C	G	
2	C	$	0
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

$

From $BWT(T)$=CG\$TTCC back to T=CTCTGC\$

Right-to-left reconstruction of T from L

i	F	L	LF
0	$\boxed{\$}$	C	
1	C	G	
2	C	\$	0
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

\$

From $BWT(T)$=CG\$TTCC back to T=CTCTGC\$

Right-to-left reconstruction of T from L

i	F	L	LF
0	\$	$\boxed{\text{C}}$	
1	C	G	
2	C	\$	0
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

$\boxed{\text{C}}$ \$

From $BWT(T)$=CG$TTCC back to T=CTCTGC$

Right-to-left reconstruction of T from L

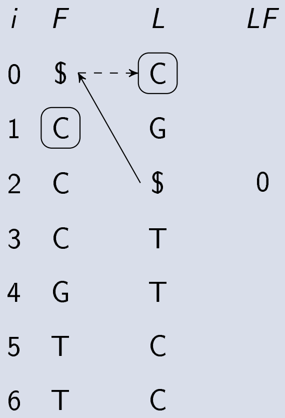
i	F	L	LF
0	$	$\boxed{\text{C}}$	
1	C	G	
2	C	$	0
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

C $

From $BWT(T)$=CG$TTCC back to T=CTCTGC$

Right-to-left reconstruction of T from L



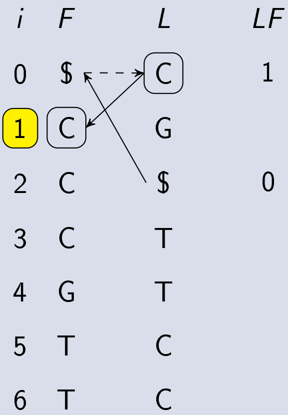
i	F	L	LF
0	$	C	
1	C	G	
2	C	$	0
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

C $

From $BWT(T)$=CG\$TTCC back to T=CTCTGC\$

Right-to-left reconstruction of T from L



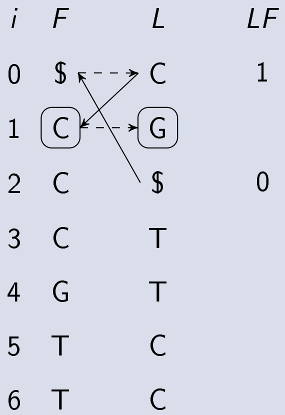
i	F	L	LF
0	\$	C	1
1	C	G	
2	C	\$	0
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

C \$

From $BWT(T)$=CG$TTCC back to T=CTCTGC$

Right-to-left reconstruction of T from L



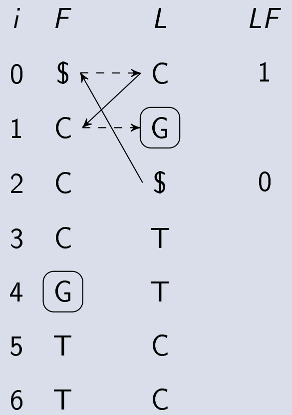
i	F	L	LF
0	$	C	1
1	$\boxed{\text{C}}$	$\boxed{\text{G}}$	
2	C	$	0
3	C	T	
4	G	T	
5	T	C	
6	T	C	

Right-to-left suffixes of T

$\boxed{\text{G}}$C $

From $BWT(T)=\mathrm{CG\$TTCC}$ back to $T=\mathrm{CTCTGC\$}$

Right-to-left reconstruction of T from L



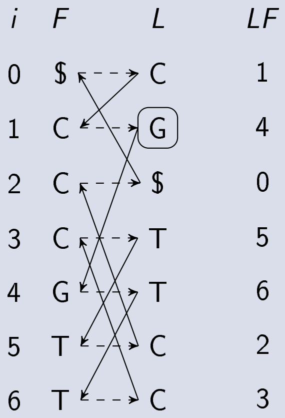
i	F	L	LF
0	\$	C	1
1	C	$\boxed{G}$	
2	C	\$	0
3	C	T	
4	$\boxed{G}$	T	
5	T	C	
6	T	C	

Reconstructed text T

G C \$

From $BWT(T)=\text{CG\$TTCC}$ back to $T=\text{CTCTGC\$}$

Right-to-left reconstruction of T from L



i	F	L	LF
0	\$	C	1
1	C	$\boxed{\text{G}}$	4
2	C	\$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3

Reconstructed text T

C T C T G C \$

A visual representation of *LF* for $T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$}$

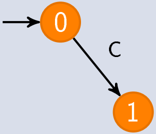
Considering *LF* as an automaton

i	F	L	LF
0	\$	C	1
1	C	G	4
2	C	\$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3

A visual representation of LF for $T = \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \mathrm{C} & \mathrm{T} & \mathrm{C} & \mathrm{T} & \mathrm{G} & \mathrm{C} & \$ \end{array}$

Considering *LF* as an automaton

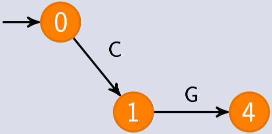
i	*F*	*L*	*LF*
0	$	C	1
1	C	G	4
2	C	$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3



A visual representation of *LF* for $T = \overset{0}{\text{C}}\,\overset{1}{\text{T}}\,\overset{2}{\text{C}}\,\overset{3}{\text{T}}\,\overset{4}{\text{G}}\,\overset{5}{\text{C}}\,\overset{6}{\text{\$}}$

Considering *LF* as an automaton

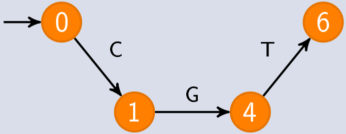
i	*F*	*L*	*LF*
0	$	C	1
1	C	G	4
2	C	$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3



A visual representation of *LF* for $T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$}$

Considering *LF* as an automaton

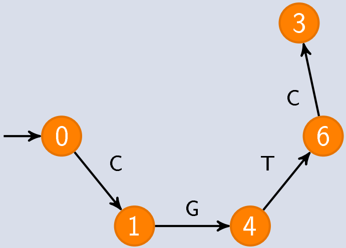
i	F	L	LF
0	\$	C	1
1	C	G	4
2	C	\$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3



A visual representation of *LF* for $T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$}$

Considering *LF* as an automaton

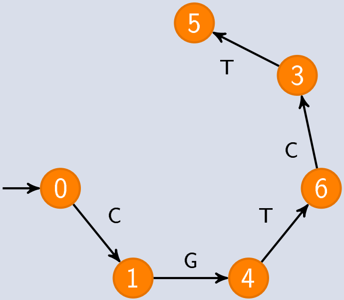
i	*F*	*L*	*LF*
0	$	C	1
1	C	G	4
2	C	$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3



A visual representation of *LF* for $T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$}$

Considering *LF* as an automaton

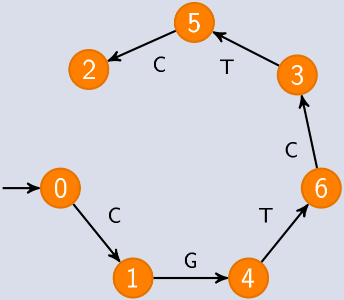
i	*F*	*L*	*LF*
0	$	C	1
1	C	G	4
2	C	$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3



A visual representation of LF for $T = \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ C & T & C & T & G & C & \$ \end{array}$

Considering *LF* as an automaton

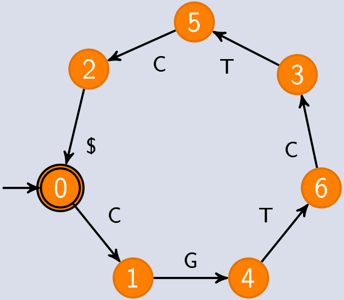
i	F	L	LF
0	\$	C	1
1	C	G	4
2	C	\$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3



A visual representation of *LF* for $T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$}$

Considering *LF* as an automaton

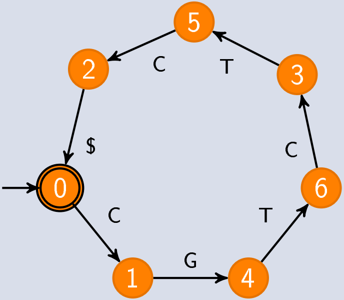
i	*F*	*L*	*LF*
0	\$	C	1
1	C	G	4
2	C	\$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3



A visual representation of *LF* for $T = \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ \mathsf{C} & \mathsf{T} & \mathsf{C} & \mathsf{T} & \mathsf{G} & \mathsf{C} & \$ \end{array}$

Considering *LF* as an automaton

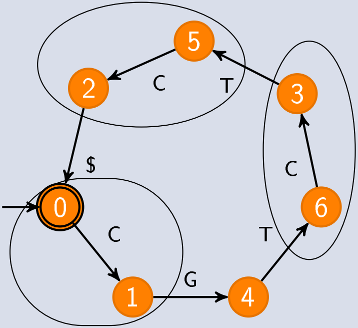
i	*F*	*L*	*LF*
0	$	C	1
1	C	G	4
2	C	$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3

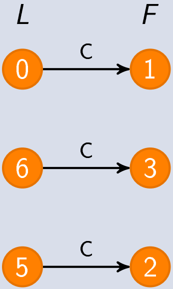


A visual representation of *LF* for $T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$}$

Considering *LF* as an automaton

i	*F*	*L*	*LF*
0	\$	C	1
1	C	G	4
2	C	\$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3

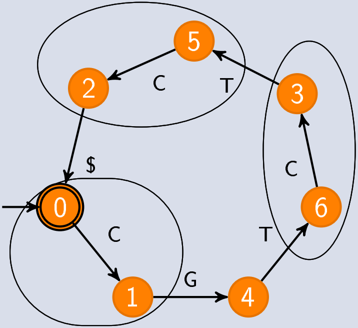


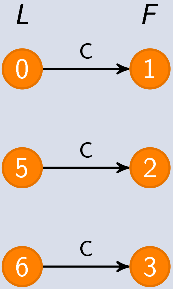


A visual representation of *LF* for $T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$}$

Considering *LF* as an automaton

i	*F*	*L*	*LF*
0	\$	C	1
1	C	G	4
2	C	\$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3

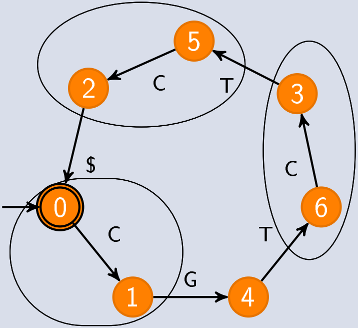


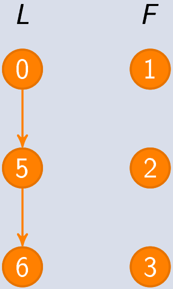


A visual representation of *LF* for $T = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{T}}\,\overset{4}{\mathrm{G}}\,\overset{5}{\mathrm{C}}\,\overset{6}{\$}$

Considering *LF* as an automaton

i	*F*	*L*	*LF*
0	$	C	1
1	C	G	4
2	C	$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3

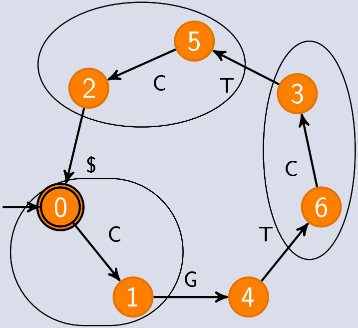


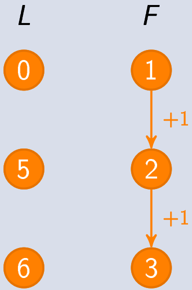


A visual representation of *LF* for $T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$}$

Considering *LF* as an automaton

i	*F*	*L*	*LF*
0	$	C	1
1	C	G	4
2	C	$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3

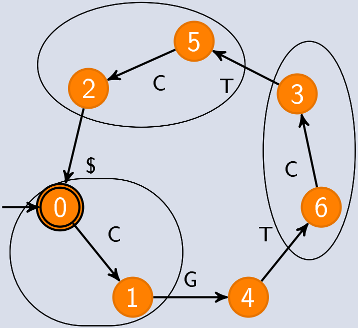


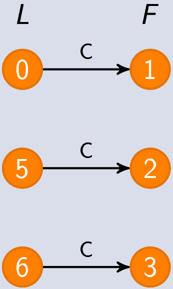


A visual representation of *LF* for $T = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{T}}\,\overset{4}{\mathrm{G}}\,\overset{5}{\mathrm{C}}\,\overset{6}{\$}$

Considering *LF* as an automaton

i	*F*	*L*	*LF*
0	$	C	1
1	C	G	4
2	C	$	0
3	C	T	5
4	G	T	6
5	T	C	2
6	T	C	3





Property

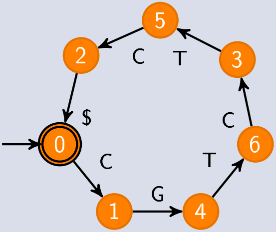
If $i_1 \xrightarrow{c} i_2$ $j_1 \xrightarrow{c} j_2$ then $i_1 < j_1 \iff i_2 < j_2$

Updating *LF* and the BWT

$$T = \begin{array}{ccccccc}0&1&2&3&4&5&6\\ C&T&C&T&G&C&\$\end{array} \rightarrow T' = \begin{array}{cccccccc}0&1&2&3&4&5&6&7\\ C&T&G&C&T&G&C&\$\end{array}$$

Stage 1: inserted letter *G* is between \$ and *L* (not in *L*)

i	*F*	*L*
0	\$	C
1	C	G
2	C	\$
3	C	T
4	G	T
5	T	C
6	T	C



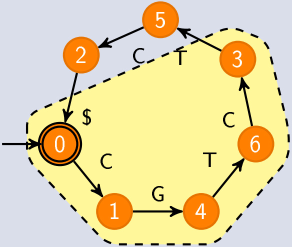
F							*L*
\$	C	T	*G*	C	T	G	C
C	\$	C	T	*G*	C	T	G
⋮							
G	C	\$	C	T	*G*	C	T
⋮							
T	G	C	\$	C	T	*G*	C

Updating *LF* and the BWT

$$T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$} \rightarrow T' = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{\color{red}{G}}\,\boxed{\overset{3}{C}\,\overset{4}{T}\,\overset{5}{G}\,\overset{6}{C}\,\overset{7}{\$}}$$

Stage 1: inserted letter *G* is between $ and *L* (not in *L*)

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	T
4	G	T
5	T	C
6	T	C



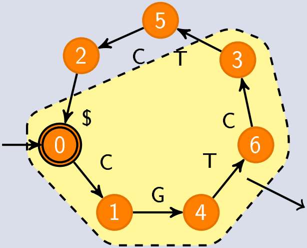
F							*L*
$	C	T	*G*	C	T	G	C
C	$	C	T	*G*	C	T	G
⋮							
G	C	$	C	T	*G*	C	T
⋮							
T	G	C	$	C	T	*G*	C

Updating LF and the BWT

$$T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$} \rightarrow T' = \overset{0}{C}\overset{1}{T}\overset{2}{G}\boxed{\overset{3}{C}\overset{4}{T}\overset{5}{G}\overset{6}{C}\overset{7}{\$}}$$

Stage 1: inserted letter *G* is between \$ and L (not in L)

i	F	L
0	\$	C
1	C	G
2	C	\$
3	C	T
4	G	T
5	T	C
6	T	C



Preserved zone

Best possible situation:

→ at position 1.

On average:

→ at position $n/2$.

Worst-case situation:

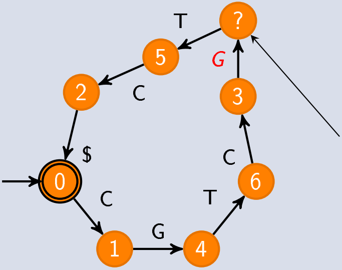
→ at position n.

Updating *LF* and the BWT

$$T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$} \to T' = \overset{0}{C}\overset{1}{T}\overset{2}{\color{red}G}\boxed{\overset{3}{C}\overset{4}{T}\overset{5}{G}\overset{6}{C}\overset{7}{\$}}$$

Stage 2: inserted letter *G* is in *L*

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	T
4	G	T
5	T	C
6	T	C



F   *L*

C T G C $ C T *G*

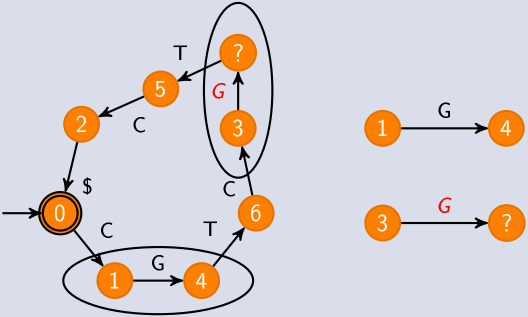
We have to consider all existing transitions labeled with *G*

Updating *LF* and the BWT

$$T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$} \rightarrow T' = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{G}\,\overset{3}{C}\,\overset{4}{T}\,\overset{5}{G}\,\overset{6}{C}\,\overset{7}{\$}$$

Stage 2: inserted letter *G* is in *L*

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	*G*
4	G	T
5	T	C
6	T	C

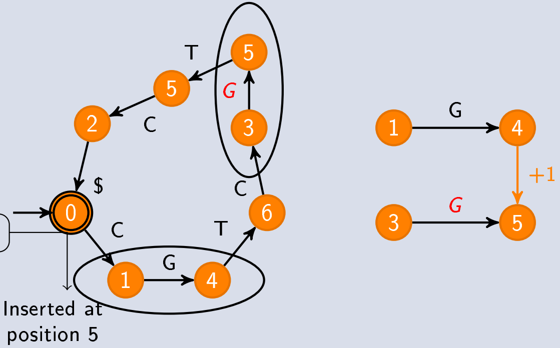


Updating *LF* and the BWT

$$T = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{T}}\,\overset{4}{\mathrm{G}}\,\overset{5}{\mathrm{C}}\,\overset{6}{\$} \rightarrow T' = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\color{red}{G}}\,\overset{3}{\mathrm{C}}\,\overset{4}{\mathrm{T}}\,\overset{5}{\mathrm{G}}\,\overset{6}{\mathrm{C}}\,\overset{7}{\$}$$

Stage 2: inserted letter *G* is in *L*

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	*G*
4	G	T
5	*G*	T
6	T	C
7	T	C

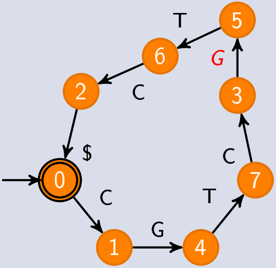


Updating *LF* and the BWT

$$T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$} \rightarrow T' = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{\color{red}G}\,\overset{3}{C}\,\overset{4}{T}\,\overset{5}{G}\,\overset{6}{C}\,\overset{7}{\$}$$

Stage 2: inserted letter *G* is in *L*

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	*G*
4	G	T
5	*G*	T
6	T	C
7	T	C

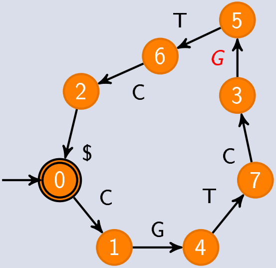


Updating *LF* and the BWT

$$T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$} \rightarrow T' = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{G}\,\overset{3}{C}\,\overset{4}{T}\,\overset{5}{G}\,\overset{6}{C}\,\overset{7}{\$}$$

Stage 2: inserted letter *G* is in *L*

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	*G*
4	G	T
5	*G*	T
6	T	C
7	T	C



F							*L*
G	C	T	G	C	$	C	T

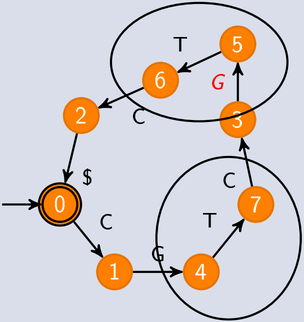
We have to consider all existing transitions labeled with T.

Updating *LF* and the BWT

$$T = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{T}}\,\overset{4}{\mathrm{G}}\,\overset{5}{\mathrm{C}}\,\overset{6}{\$} \rightarrow T' = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{G}\,\overset{3}{\mathrm{C}}\,\overset{4}{\mathrm{T}}\,\overset{5}{\mathrm{G}}\,\overset{6}{\mathrm{C}}\,\overset{7}{\$}$$

Stage 3: inserted letter *G* is in *F*: new row

i	*F*	*L*
0	\$	C
1	C	G
2	C	\$
3	C	*G*
4	G	T
5	*G*	T
6	T	C
7	T	C



Now, by T, we go to 6 from state 5 instead of 3.

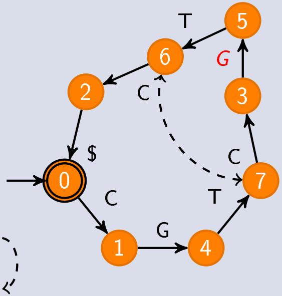
We check if the property is satisfied by considering the other edges labelled by T.

Updating *LF* and the BWT

$$T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$} \rightarrow T' = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{G}\,\overset{3}{C}\,\overset{4}{T}\,\overset{5}{G}\,\overset{6}{C}\,\overset{7}{\$}$$

Stage 3: inserted letter *G* is in *F*: new row

i	*F*	*L*
0	\$	C
1	C	G
2	C	\$
3	C	*G*
4	G	T
5	*G*	T
6	T	C
7	T	C

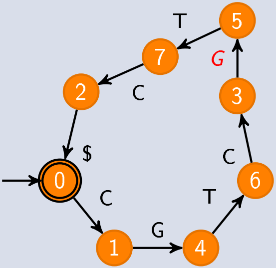


Updating LF and the BWT

$$T = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{\mathrm{C}}\,\overset{3}{\mathrm{T}}\,\overset{4}{\mathrm{G}}\,\overset{5}{\mathrm{C}}\,\overset{6}{\$} \rightarrow T' = \overset{0}{\mathrm{C}}\,\overset{1}{\mathrm{T}}\,\overset{2}{G}\,\overset{3}{\mathrm{C}}\,\overset{4}{\mathrm{T}}\,\overset{5}{\mathrm{G}}\,\overset{6}{\mathrm{C}}\,\overset{7}{\$}$$

Stage 3: inserted letter G is in F: new row

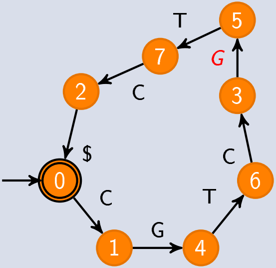
i	F	L
0	\$	C
1	C	G
2	C	\$
3	C	G
4	G	T
5	G	T
6	T	C
7	T	C



Updating *LF* and the BWT

$$T = \overset{0}{\mathrm{C}}\overset{1}{\mathrm{T}}\overset{2}{\mathrm{C}}\overset{3}{\mathrm{T}}\overset{4}{\mathrm{G}}\overset{5}{\mathrm{C}}\overset{6}{\$} \rightarrow T' = \overset{0}{\mathrm{C}}\overset{1}{\mathrm{T}}\overset{2}{G}\overset{3}{\mathrm{C}}\overset{4}{\mathrm{T}}\overset{5}{\mathrm{G}}\overset{6}{\mathrm{C}}\overset{7}{\$}$$

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	*G*
4	G	T
5	*G*	T
6	T	C
7	T	C



F *L*

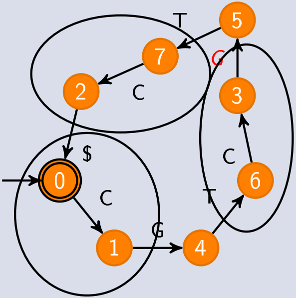
T *G* C T G C $ C

We have to consider all existing transitions labeled with C.

Updating *LF* and the BWT

$$T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$} \rightarrow T' = \overset{0}{C}\overset{1}{T}\overset{2}{G}\overset{3}{C}\overset{4}{T}\overset{5}{G}\overset{6}{C}\overset{7}{\$}$$

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	*G*
4	G	T
5	*G*	T
6	T	C
7	T	C

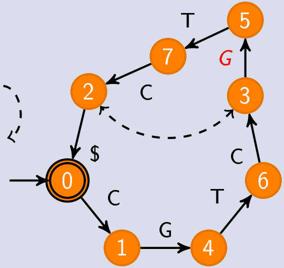


blablabla

Updating *LF* and the BWT

$$T = \overset{0}{\mathsf{C}}\,\overset{1}{\mathsf{T}}\,\overset{2}{\mathsf{C}}\,\overset{3}{\mathsf{T}}\,\overset{4}{\mathsf{G}}\,\overset{5}{\mathsf{C}}\,\overset{6}{\$} \rightarrow T' = \overset{0}{\mathsf{C}}\,\overset{1}{\mathsf{T}}\,\overset{2}{G}\,\overset{3}{\mathsf{C}}\,\overset{4}{\mathsf{T}}\,\overset{5}{\mathsf{G}}\,\overset{6}{\mathsf{C}}\,\overset{7}{\$}$$

i	*F*	*L*
0	$	C
1	C	G
2	C	$
3	C	*G*
4	G	T
5	*G*	T
6	T	C
7	T	C

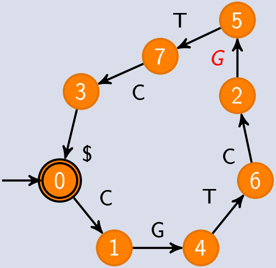


Same principle:
$6 < 7$ but $3 > 2$

Updating LF and the BWT

$$T = \overset{0}{C}\overset{1}{T}\overset{2}{C}\overset{3}{T}\overset{4}{G}\overset{5}{C}\overset{6}{\$} \rightarrow T' = \overset{0}{C}\overset{1}{T}\overset{2}{G}\overset{3}{C}\overset{4}{T}\overset{5}{G}\overset{6}{C}\overset{7}{\$}$$

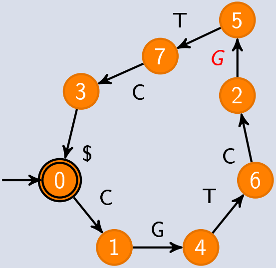
i	F	L
0	\$	C
1	C	G
2	C	G
3	C	\$
4	G	T
5	G	T
6	T	C
7	T	C



Updating *LF* and the BWT

$$T = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{C}\,\overset{3}{T}\,\overset{4}{G}\,\overset{5}{C}\,\overset{6}{\$} \rightarrow T' = \overset{0}{C}\,\overset{1}{T}\,\overset{2}{G}\,\overset{3}{C}\,\overset{4}{T}\,\overset{5}{G}\,\overset{6}{C}\,\overset{7}{\$}$$

i	*F*	*L*
0	$	C
1	C	G
2	C	*G*
3	C	$
4	G	T
5	*G*	T
6	T	C
7	T	C



We have only one edge labelled $, therefore the property is satisfied.