Space efficient locating with the r-index

Gagie, Navarro, Prezza
SODA 2017 & JACM 2018

Locate on a BWT-index

We consider the problem of finding the position in the text of all pattern occurrences.

Recall we only have the first and last column (F and L)

T = swiss miss miss missing

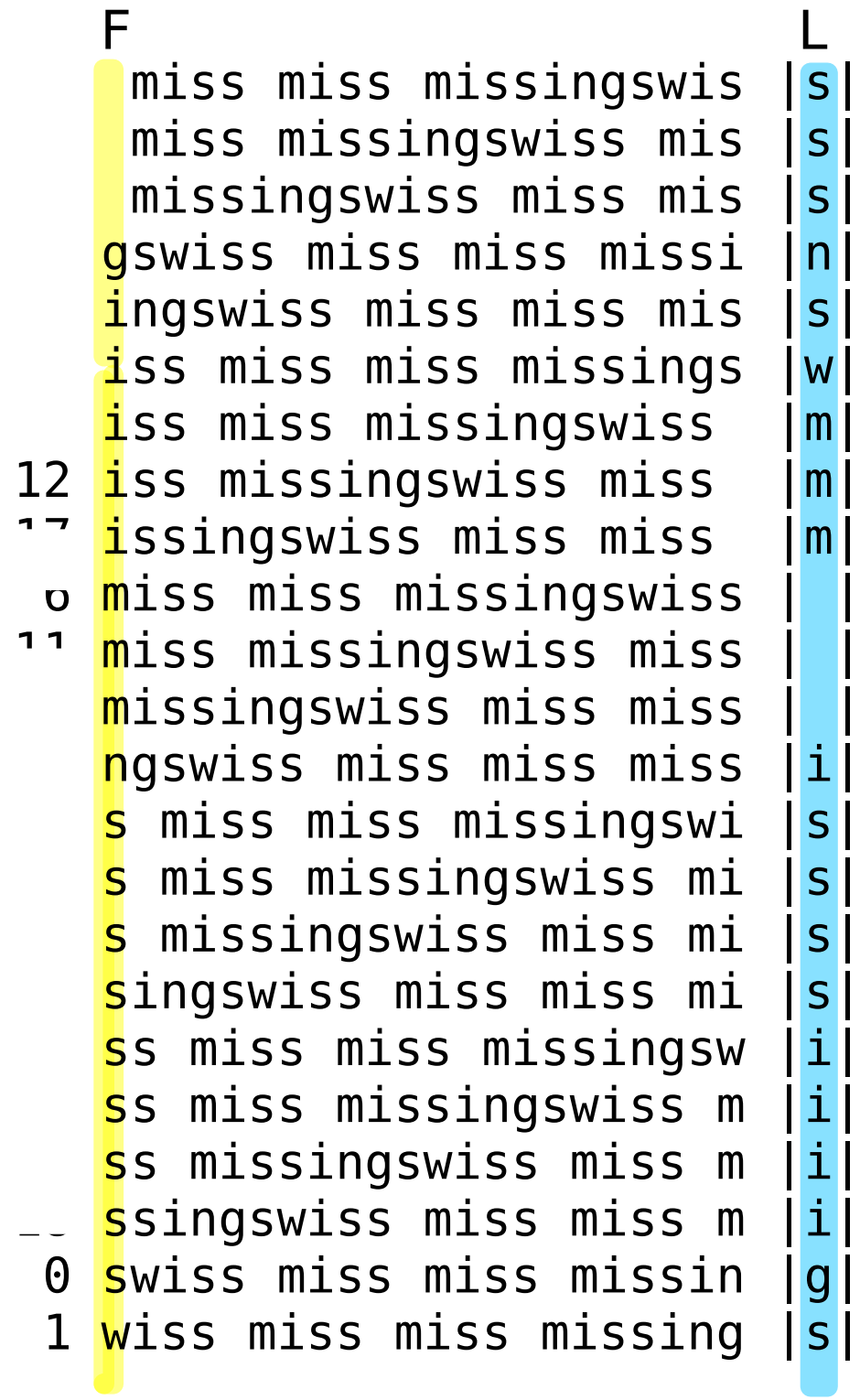
	F	L
5	miss miss missingswis	s
10	miss missingswiss mis	s
15	missingswiss miss mis	s
22	gswiss miss miss missi	n
20	ingswiss miss miss mis	s
2	iss miss miss missings	w
7	iss miss missingswiss	m
12	iss missingswiss miss	m
17	issingswiss miss miss	m
6	miss miss missingswiss	
11	miss missingswiss miss	
16	missingswiss miss miss	
21	ngswiss miss miss miss	i
4	s miss miss missingswi	s
9	s miss missingswiss mi	s
14	s missingswiss miss mi	s
19	singswiss miss miss mi	s
3	ss miss miss missingsw	i
8	ss miss missingswiss m	i
13	ss missingswiss miss m	i
18	ssingswiss miss miss m	i
0	swiss miss miss missin	g
1	wiss miss miss missing	s

Simple solution: uniform sampling
store one out of t SA values

To find the position of an occ
use te LF map to move backward
untill we reach a stored value

the parametr t induces a trade-off:

extra space: (n/t)log n bits
locate time: O(t) per occurence



When the input is highly compressible (for example consists of many variants of the same sequence) it is more convenient to use an index of size O(r) words where r is the number of runs in the BWT.

In this setting storing (n/t) SA entries space dominates the index size: using BWT properties we can save space by storing only 2r SA entries

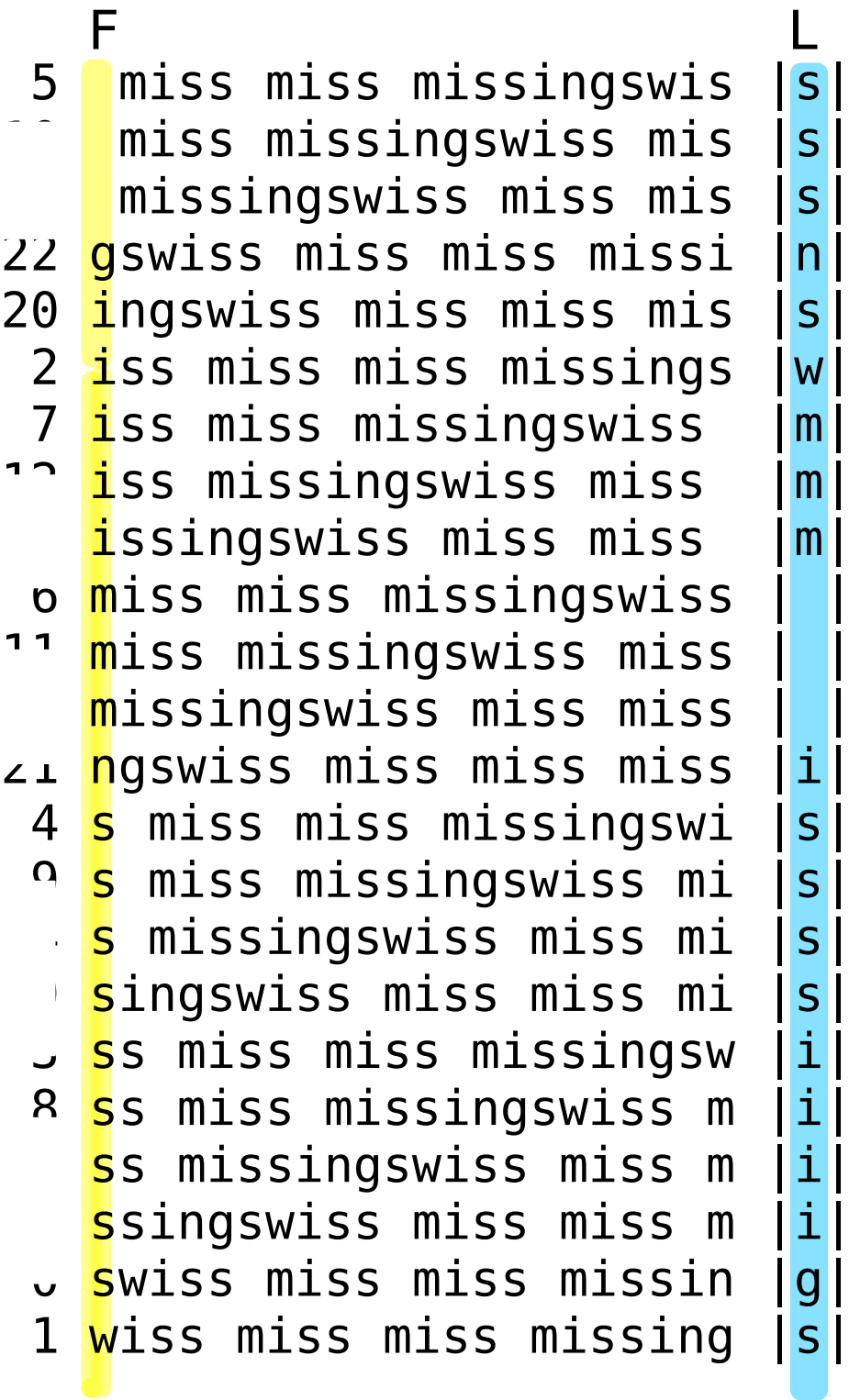
The resulting index is called the r-index
[Gagie, Prezza, Navarro 2018]

r-Index: locate 1st occurrence

Toehold Lemma:
to locate the lexicographically first occurrence of a pattern we only need the SA entries for rows containing the first occ of a run in L

Proof:
by induction on the backward search steps

Note: we also need the select operation on column L



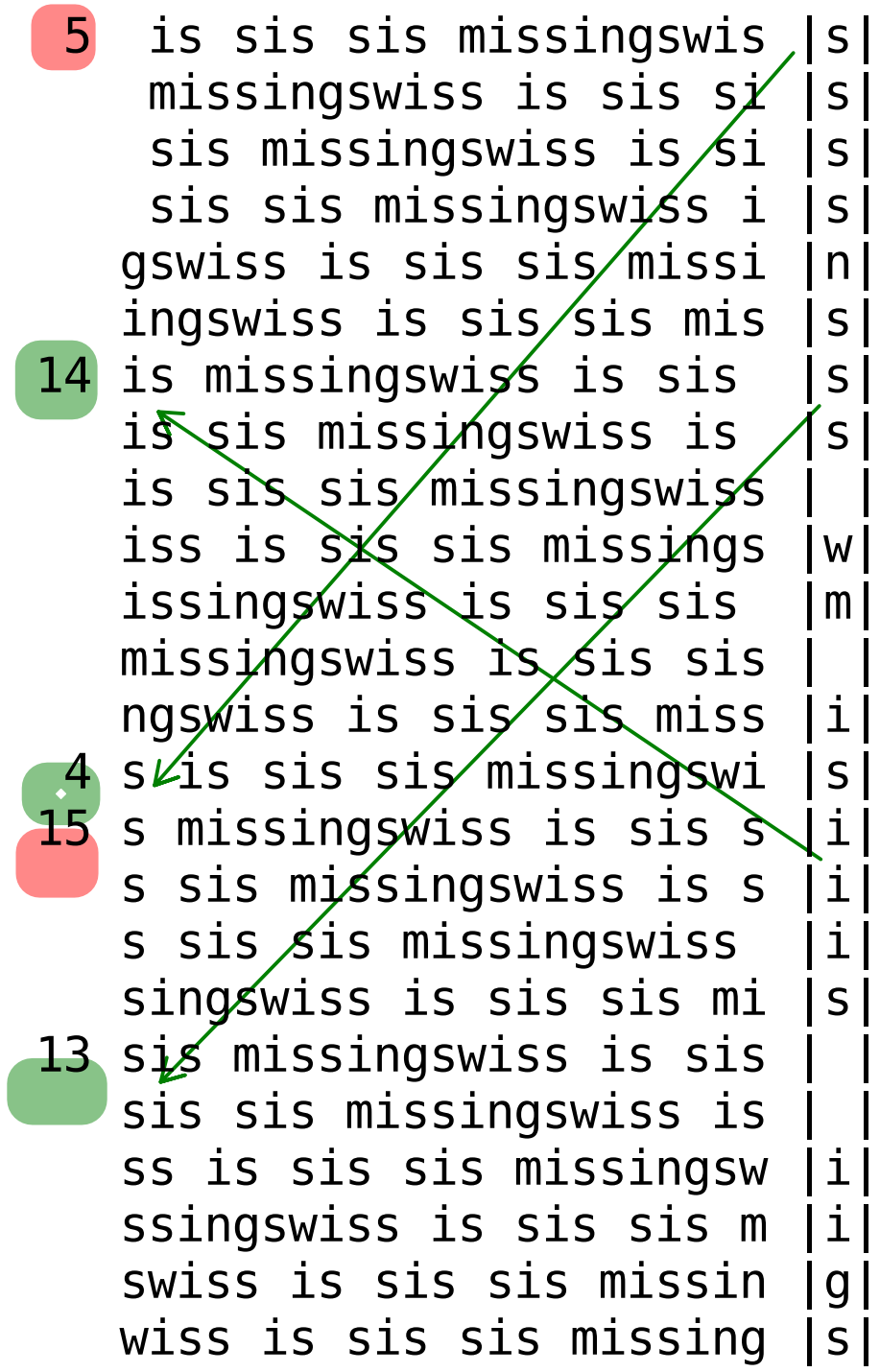
Example: searching "sis" in

T = swiss is sis sis missing

We only use the SA entries marked in red. The one in green are derived

Green arrows are applications of the LF map. Each LF application reduces the current position by 1

The first occurrence of "sis" is in text position 13



r-Index: locate next occurrence

```
                1    1    2
     0    5    0    5    0
T =  swiss miss miss missing
```

SA values at end/begin of runs
(15,22), (22,20), (20,2) (2,7)
(17,6) (16,21), (21,4) (19,3)
(18,0), (0,1)

Sorted pairs:
(0,1) (2,7), (15,22), (16,21)
(17,6) (18,0) (19,3) (20,2)
(21,4) (22,20)

	F	L
5	miss miss missingswis	s
10	miss missingswiss mis	s
15	missingswiss miss mis	s
22	gswiss miss miss missi	n
20	ingswiss miss miss mis	s
2	iss miss miss missings	w
7	iss miss missingswiss	m
12	iss missingswiss miss	m
17	issingswiss miss miss	m
6	miss miss missingswiss	
11	miss missingswiss miss	
16	missingswiss miss miss	
21	ngswiss miss miss miss	i
4	s miss miss missingswi	s
9	s miss missingswiss mi	s
14	s missingswiss miss mi	s
19	singswiss miss miss mi	s
3	ss miss miss missingsw	i
8	ss miss missingswiss m	i
13	ss missingswiss miss m	i
18	ssingswiss miss miss m	i
0	swiss miss miss missin	g
1	wiss miss miss missing	s

r-Index: locate next occurrence

```
                1    1    2
     0    5    0    5    0
T =  swiss miss miss missing
```

Sorted pairs:
(0,1) (2,7), (15,22), (16,21)
(17,6) (18,0) (19,3) (20,2)
(21,4) (22,20)

Lemma: L(p)=L(p+1) q=LF(p)
q+1=L(p+1)

rows ending with the same
symbol stay together!

F	L
miss miss missingswis	s
miss missingswiss mis	s
missingswiss miss mis	s
gswiss miss miss missi	n
ingswiss miss miss mis	s
iss miss miss missings	w
iss miss missingswiss	m
iss missingswiss miss	m
issingswiss miss miss	m
miss miss missingswiss	
miss missingswiss miss	
missingswiss miss miss	
ngswiss miss miss miss	i
s miss miss missingswi	s
s miss missingswiss mi	s
s missingswiss miss mi	s
singswiss miss miss mi	s
ss miss miss missingsw	i
ss miss missingswiss m	i
ss missingswiss miss m	i
ssingswiss miss miss m	i
swiss miss miss missin	g
wiss miss miss missing	s

r-Index: locate next occurrence

```
                1    1    2
     0    5    0    5    0
T =  swiss miss miss missing
```

Sorted pairs:
(0,1) (2,7), (15,22), (16,21)
(17,6) (18,0) (19,3) (20,2)
(21,4) (22,20)

Given the text position of a row using a predecessor query of the sorted pairs we can retrieve the text position of the next row

F	L
miss miss missingswis	s
miss missingswiss mis	s
missingswiss miss mis	s
gswiss miss miss missi	n
ingswiss miss miss mis	s
iss miss miss missings	w
iss miss missingswiss	m
iss missingswiss miss	m
issingswiss miss miss	m
miss miss missingswiss	
miss missingswiss miss	
missingswiss miss miss	
ngswiss miss miss miss	i
s miss miss missingswi	s
s miss missingswiss mi	s
s missingswiss miss mi	s
singswiss miss miss mi	s
ss miss miss missingsw	i
ss miss missingswiss m	i
ss missingswiss miss m	i
ssingswiss miss miss m	i
swiss miss miss missin	g
wiss miss miss missing	s

r-Index: locate next occurrence

```
                  1    1    2
     0    5    0    5    0
T =  swiss miss miss missing
```

Sorted pairs:
(0,1) (2,7), (15,22), (16,21)
(17,6) (18,0) (19,3) (20,2)
(21,4) (22,20)

Examples:
16->21 (from the 4th pair)
 3->2(LF) ->7(2nd pair)->8
10->2(pred)->7+(10-2) ->15

	F	L
10	miss miss missingswis	s
?	miss missingswiss mis	s
	missingswiss miss mis	s
	gswiss miss miss missi	n
	ingswiss miss miss mis	s
	iss miss miss missings	w
2	iss miss missingswiss	m
7	iss missingswiss miss	m
	issingswiss miss miss	m
	miss miss missingswiss	
	miss missingswiss miss	
	missingswiss miss miss	
16	ngswiss miss miss miss	i
?	s miss miss missingswi	s
	s miss missingswiss mi	s
	s missingswiss miss mi	s
	singswiss miss miss mi	s
3	ss miss miss missingsw	i
?	ss miss missingswiss m	i
	ss missingswiss miss m	i
	ssingswiss miss miss m	i
	swiss miss miss missin	g
	wiss miss miss missing	s

The space/time bounds for the r-index are:

$O(r)$ words $\qquad$ $O((|p| + occ) \log \log n)$ time

$O(r \log \log n)$ words $\qquad$ $O(|p| + occ)$ time (optimal)

Timeline of BWT based indexing

1994 BWT (Burrows, Wheeler)
1997 bzip2 (Seward)
2000 Backward search (Ferragina, GM)
2003 Wavelet Trees (Gupta, Grossi, Vitter)
2017 r-index (Gagie, Navarro, Prezza) ← It took 17 years to devise a space efficient locate
Next topic! → 2017 Wheeler-Graphs (Gagie, GM, Siren)