

Burrows-Wheeler Transform and Suffix Arrays

Pavel Pevzner

Department of Computer Science and Engineering
University of California at San Diego

Algorithms on Strings
Data Structures and Algorithms

This slide deck is incomplete.
For the complete set of frames,
please see our videos in the
[Algorithms on Strings](#) course on [Coursera](#)
([Algorithms and Data Structures](#) Specialization)

Outline

- **Burrows-Wheeler Transform**
- Inverting Burrows-Wheeler Transform
- Using BWT for Pattern Matching
- Suffix Arrays
- Approximate Pattern Matching

Text Compression by Run-Length Encoding

- **Run-length encoding** compresses a run of n identical symbols:

Text

GGGGGGGGGGCCCCCCCCCCCCAAAAAATTTTTTTTTTTTTTTTTTTTCCCCCG

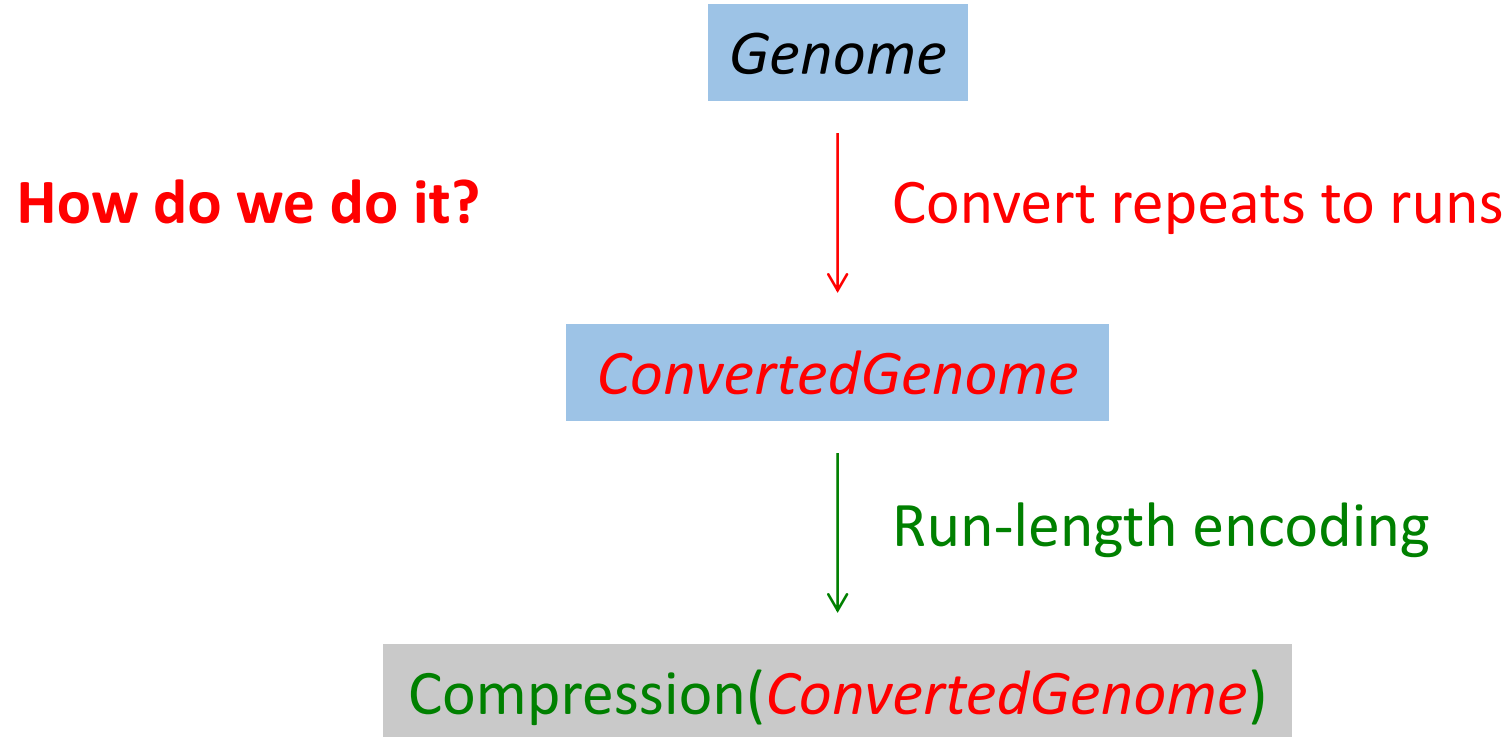
↓

10G11C7A15T5C1G

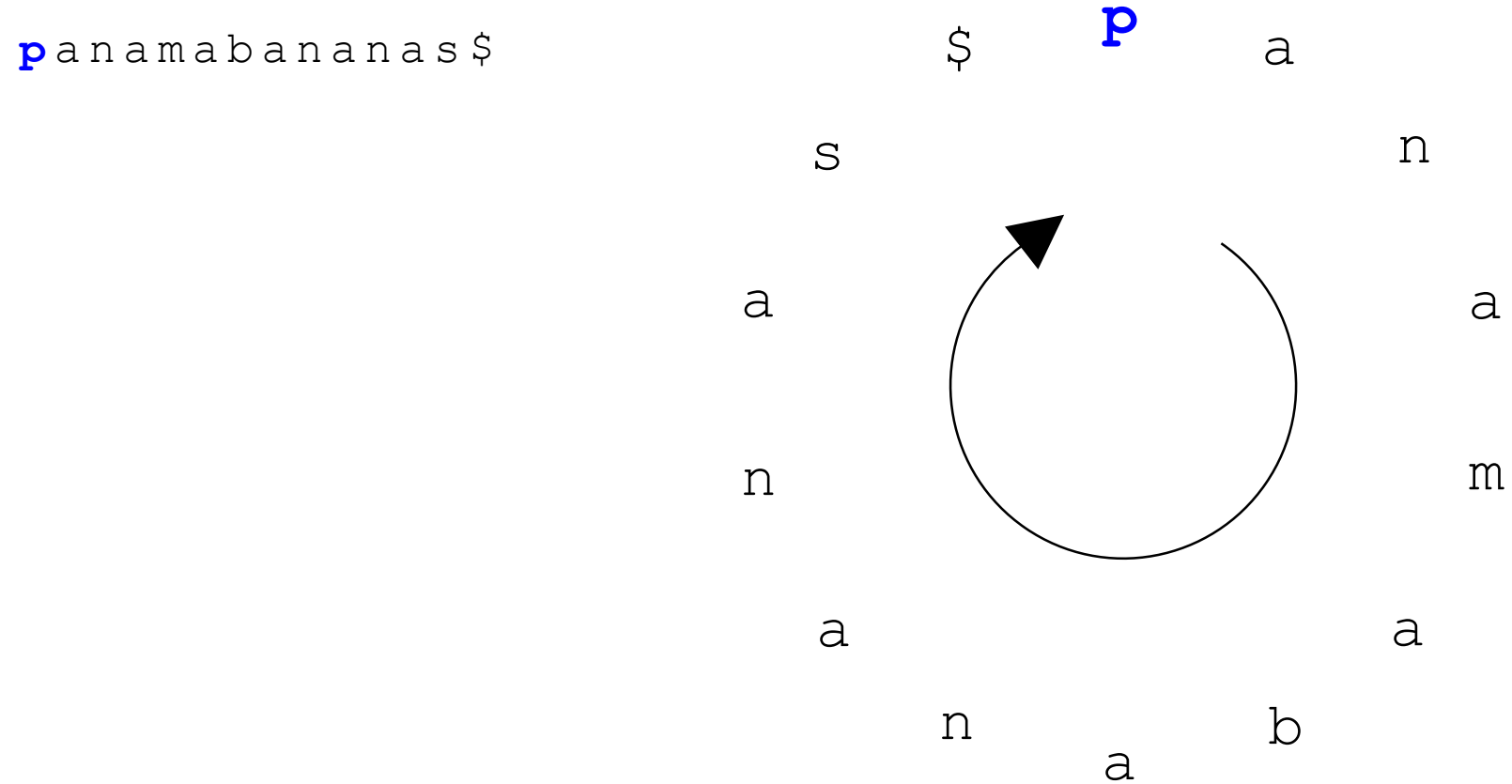
- genomes don't have lots of runs... but they do have lots of repeats:

ACTGACCGAACTGAGTATCCGACTGAACTGATCAGTACTGACATTGC

Idea: Converting Repeats to Runs

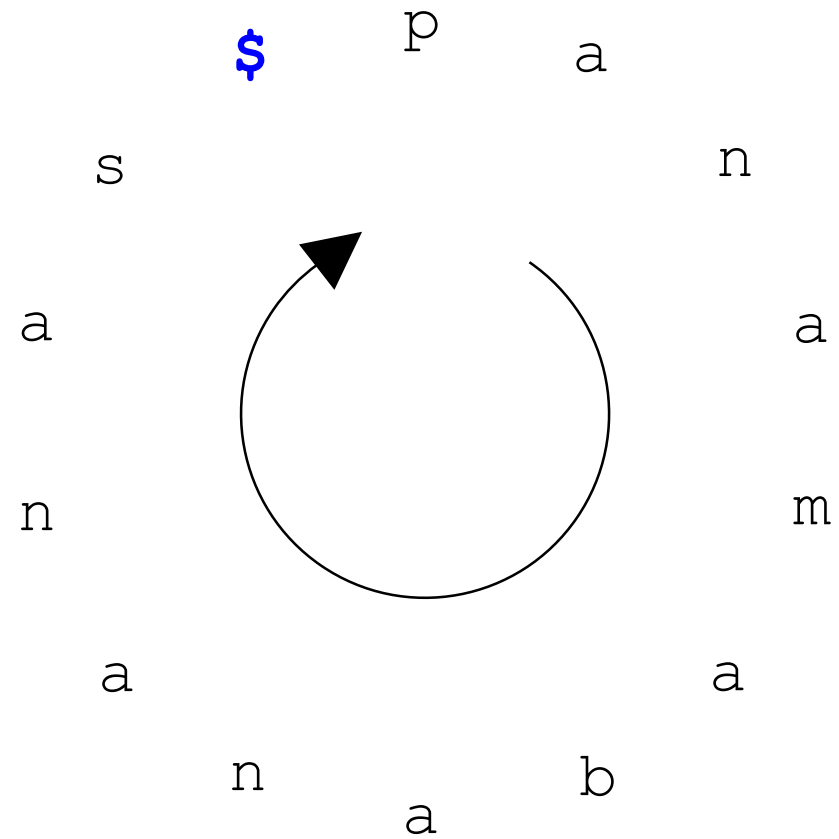


Forming All Cyclic Rotations of *Text*



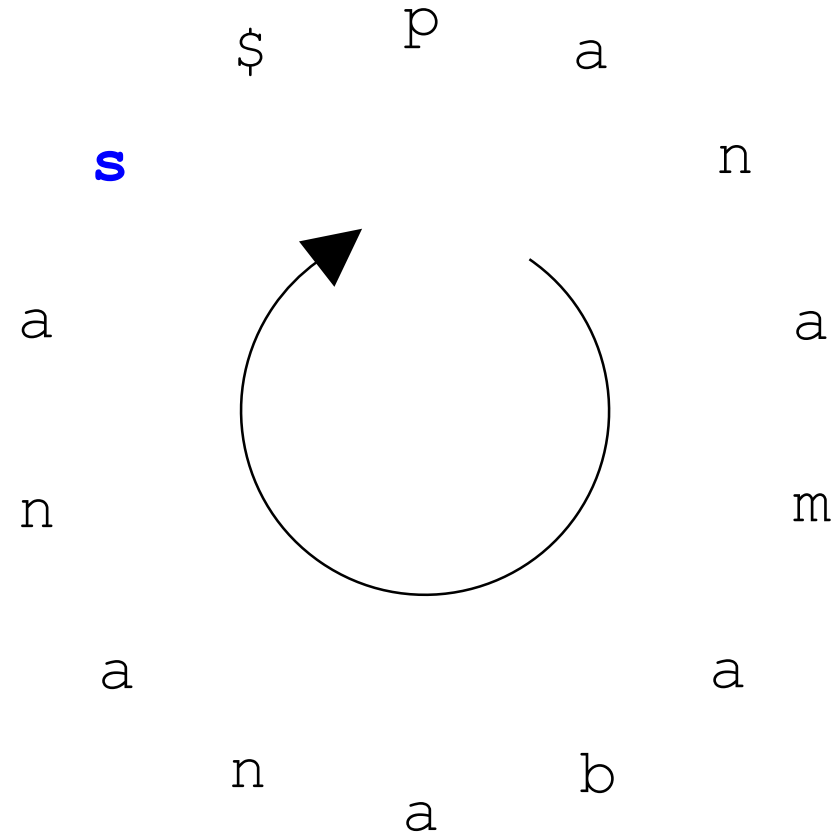
Cyclic Rotations

panamabananas\$
\$panamabananas



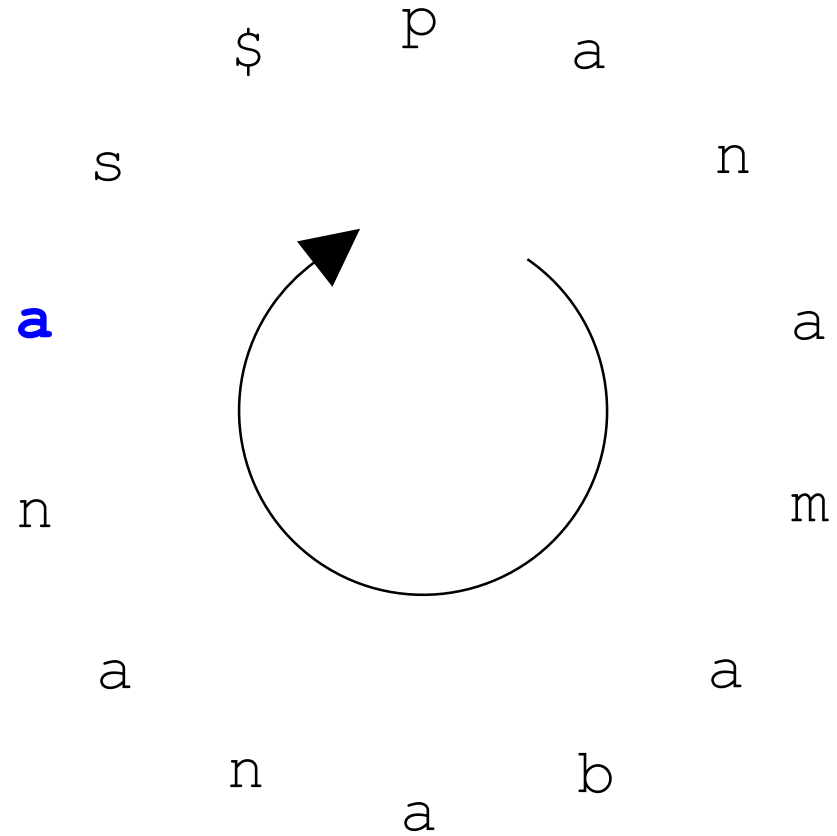
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana



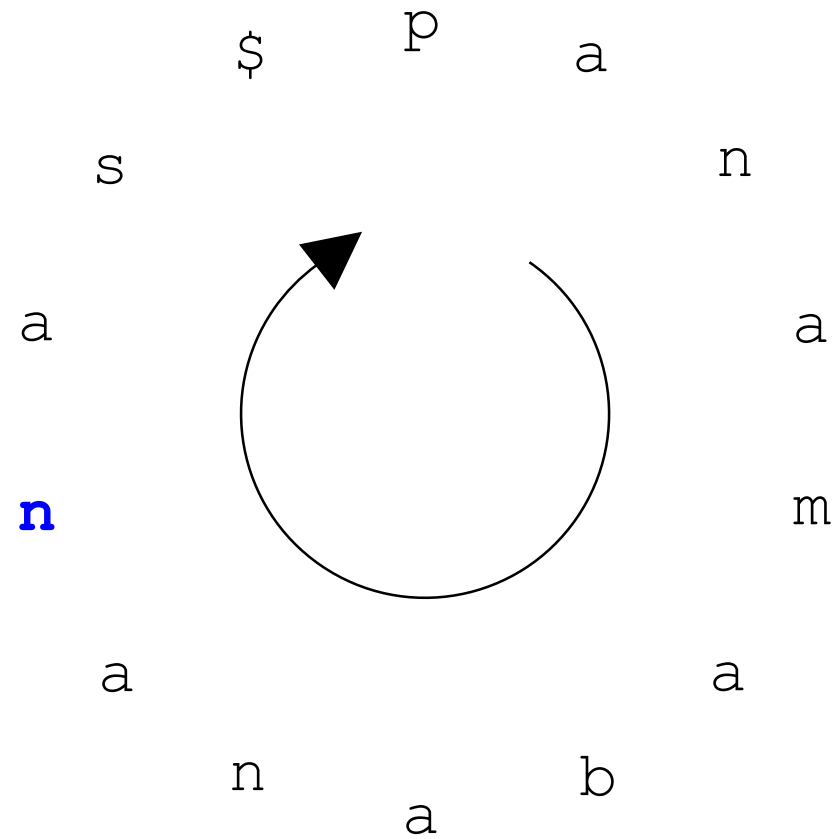
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana



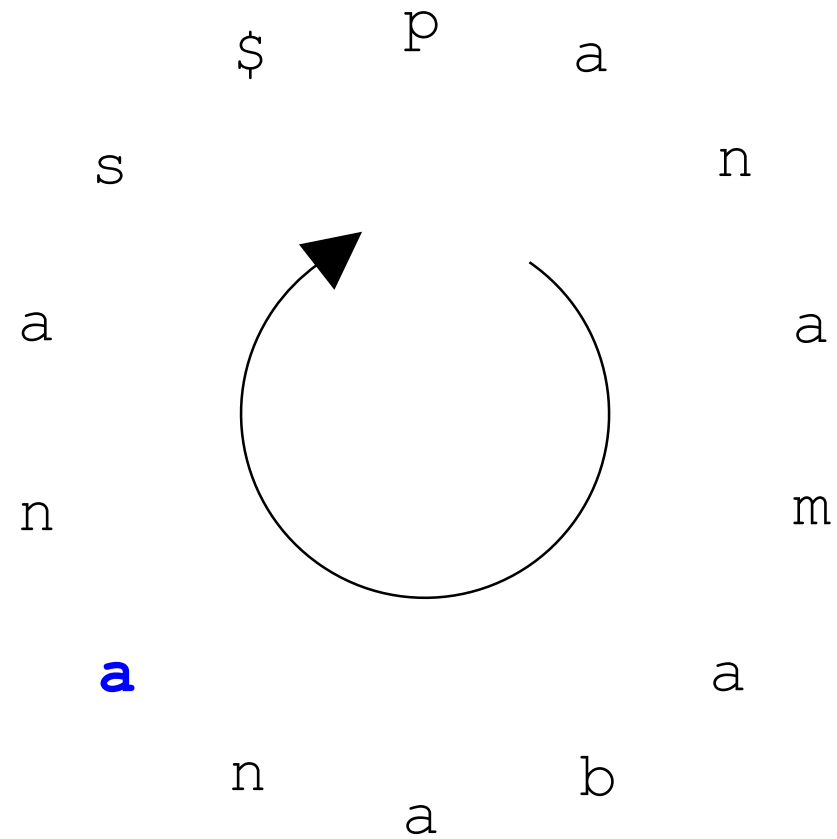
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana



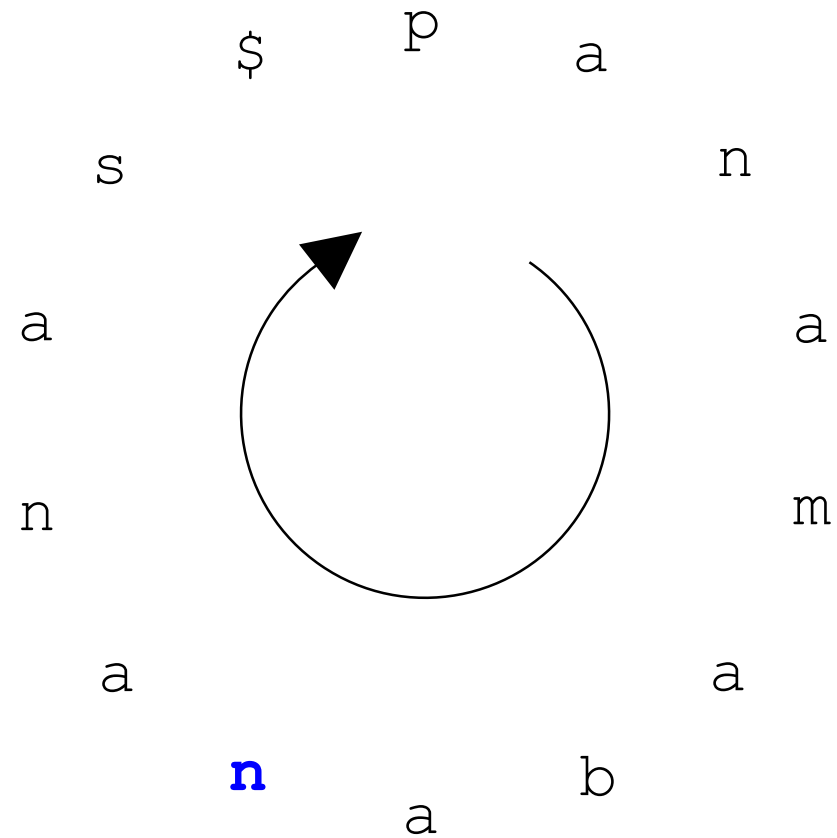
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamaban



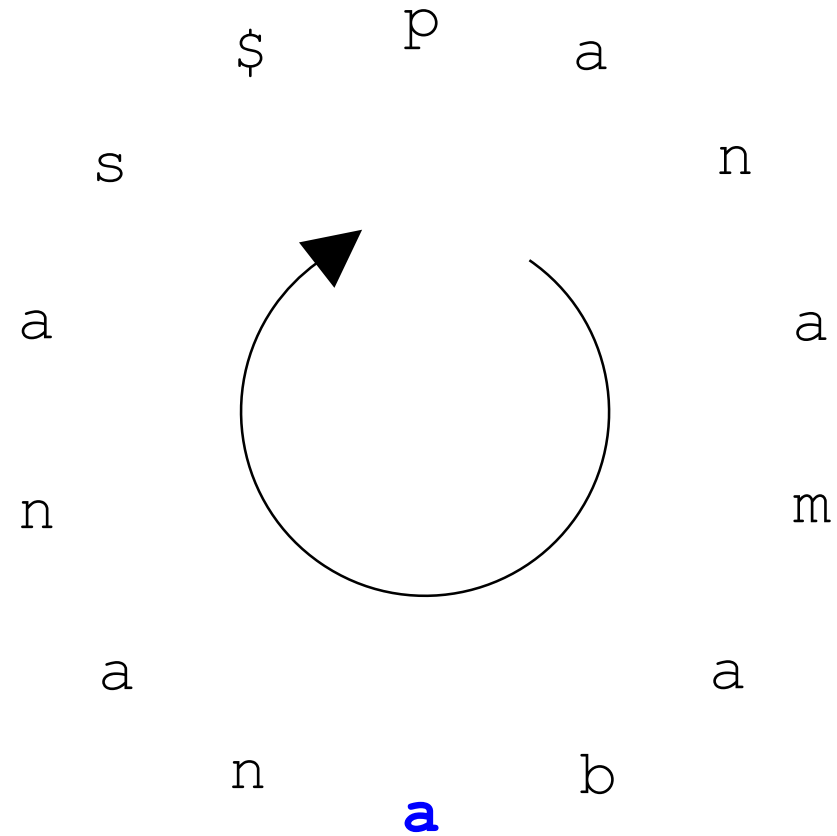
Cyclic Rotations

```
panamabananas$  
$panamabananas  
s$panamabanana  
as$panamabanana  
nas$panamabanana  
anas$panamaban  
nanas$panamaba
```



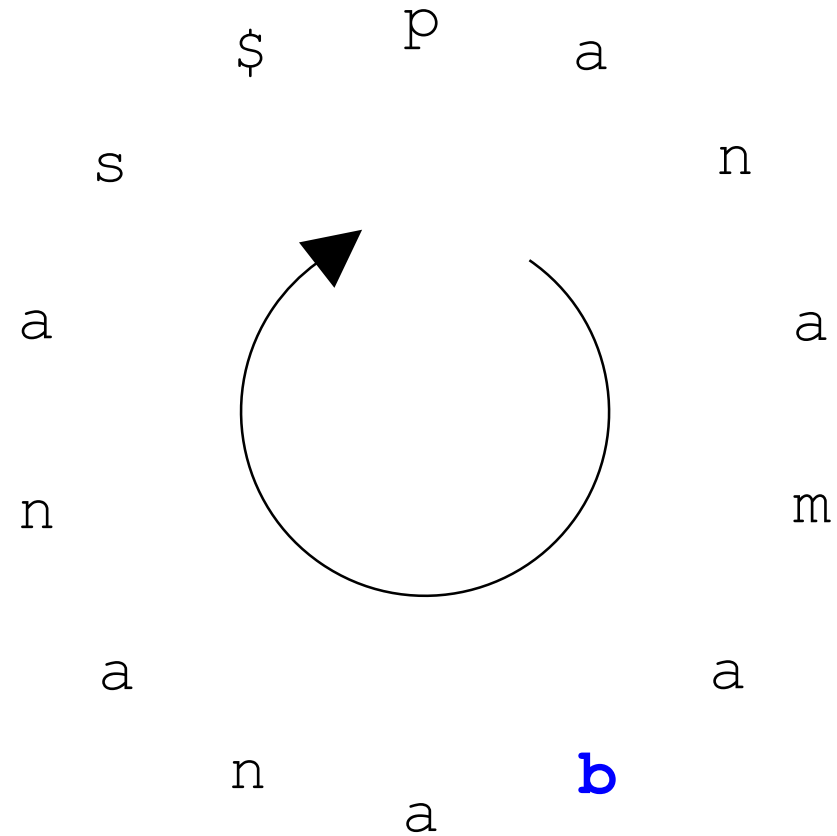
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamabana
nanas\$panamaba
ananas\$panamab



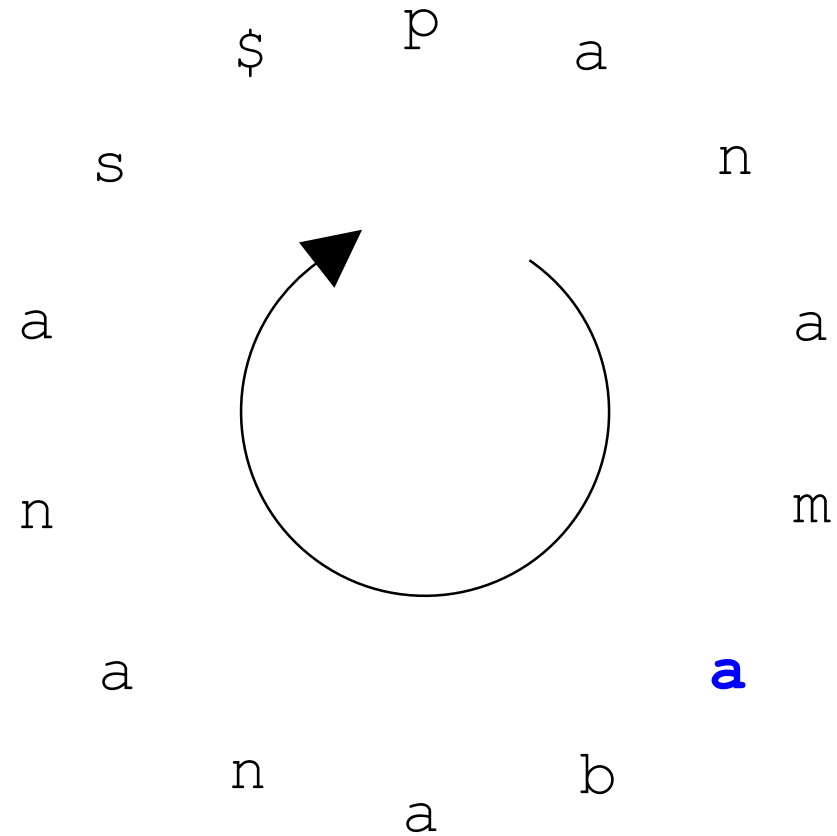
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamabana
nanas\$panamaba
ananas\$panamab
bananas\$panama



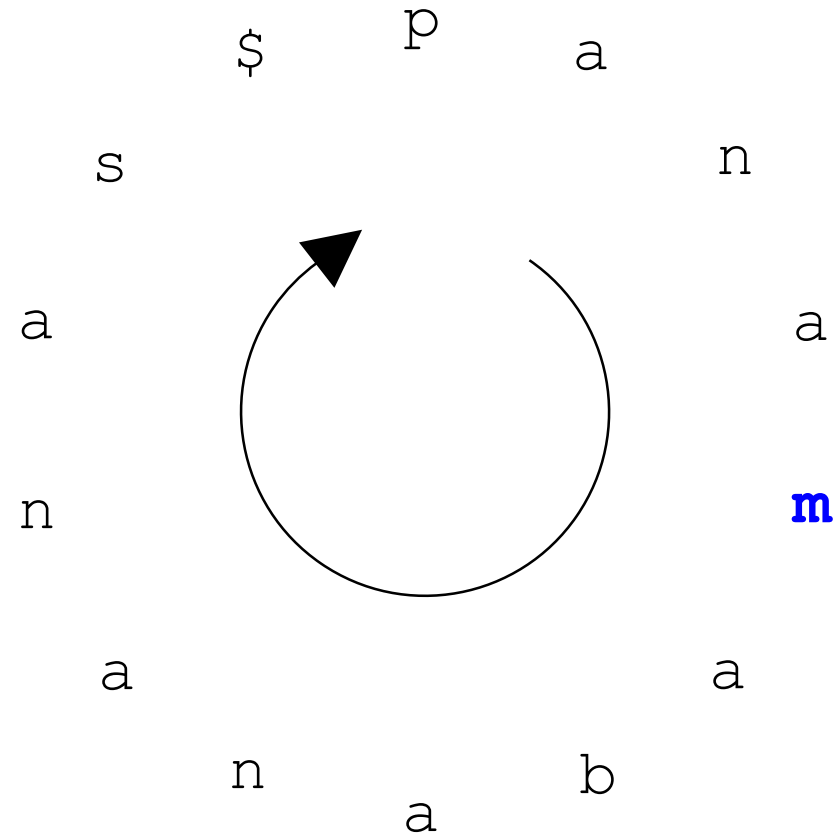
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanan
nas\$panamabana
anas\$panamaba
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam



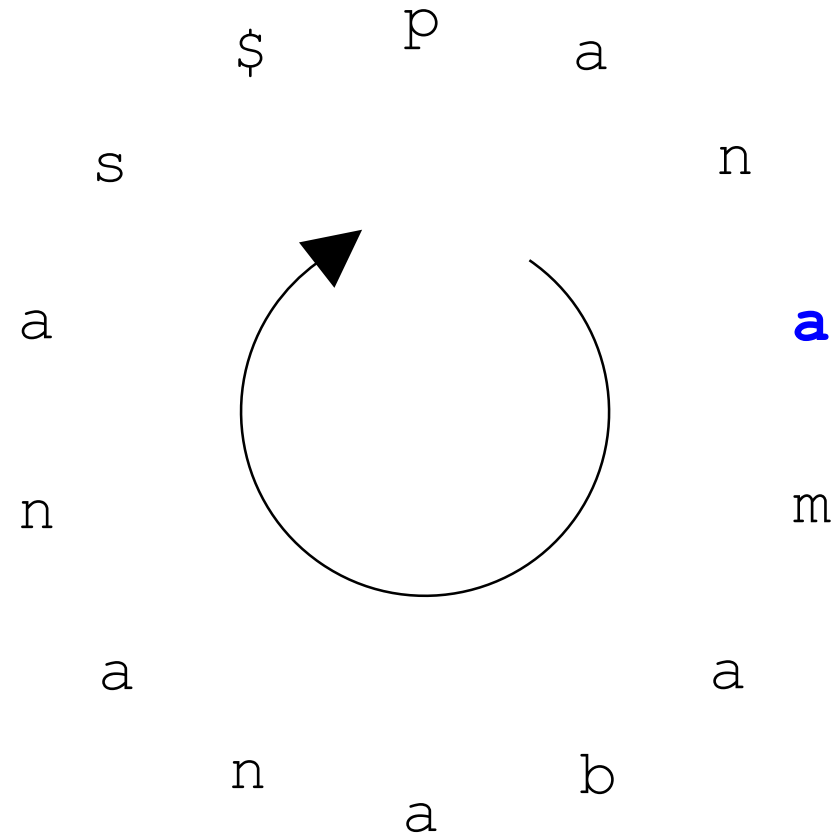
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamabana
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana



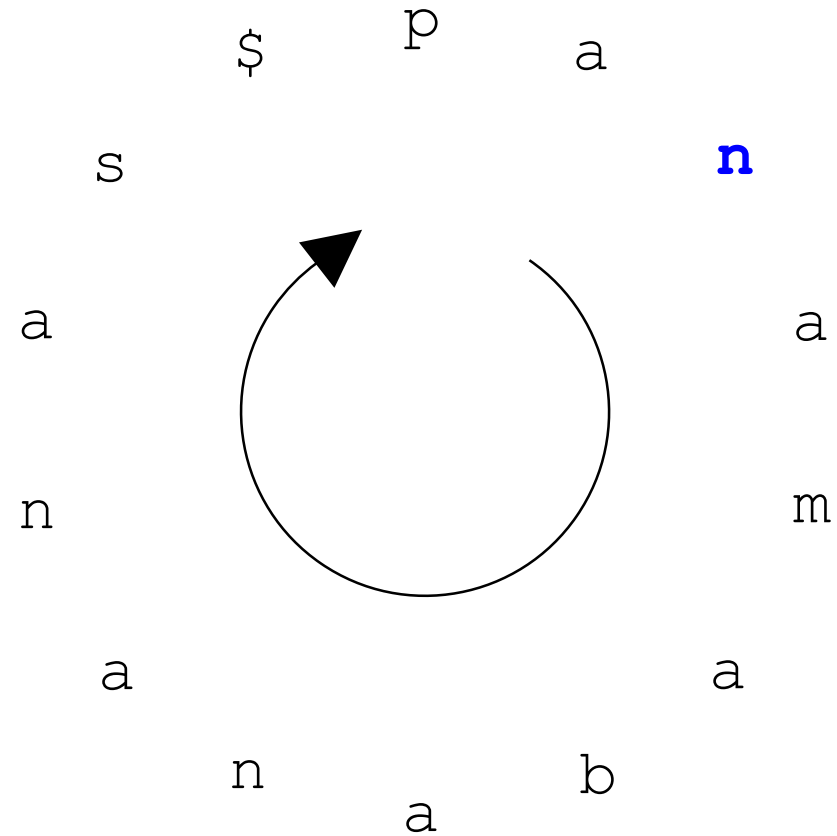
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamabana
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan



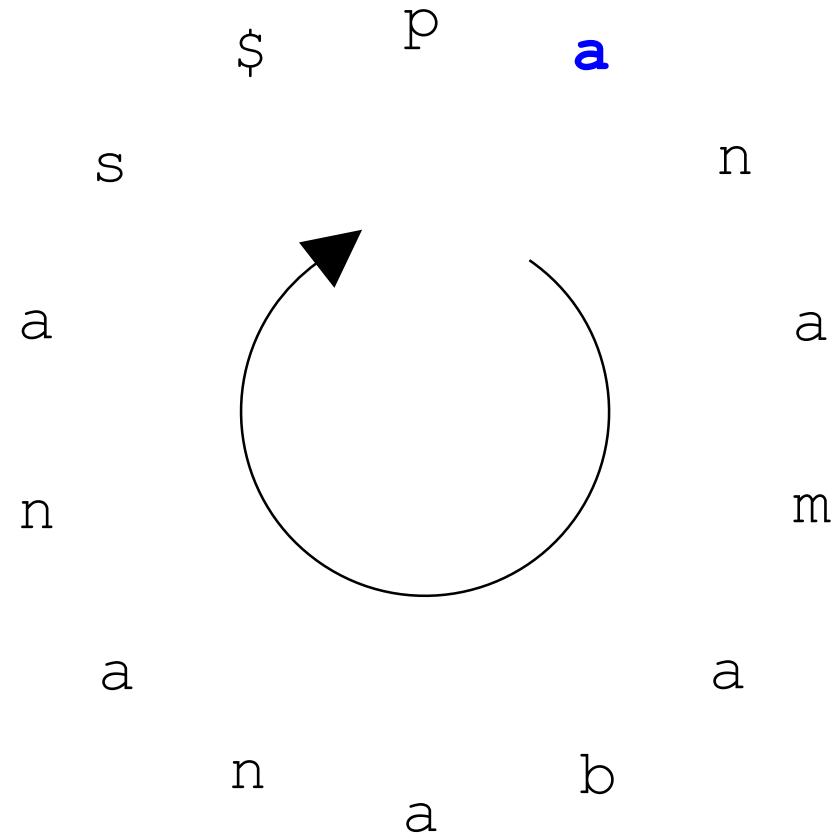
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamaban
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa



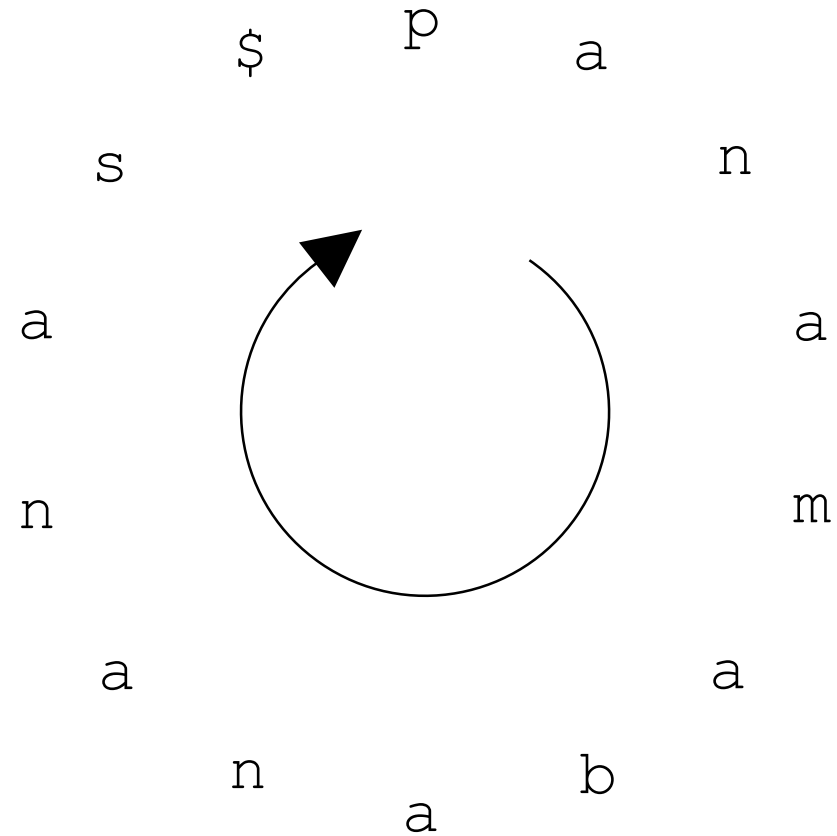
Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamabana
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa
a**n****a****m****a****b****a****n****a****n****a****s****\$****p**




Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamabana
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa
anamabananas\$p



Sorting Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanan
as\$panamabanan
nas\$panamabana
anas\$panamaban
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa
anamabananas\$p



\$panamabananas

Sort the strings
lexicographically
(\$ comes first)

Sorting Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabananas
as\$panamabanan
nas\$panamabana
anas\$panamaban
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa
anamabananas\$p

\$panamabanas
abananas\$panam

Sort the strings
lexicographically
(\$ comes first)

Sorting Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanana
as\$panamabanana
nas\$panamabana
anas\$panamaba
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa
anamabananas\$p

\$panamabananas
abananas\$panam
amabananas\$pan

Sort the strings
lexicographically
(\$ comes first)

Sorting Cyclic Rotations

panamabananas\$	→	\$ panamabananas
\$panamabananas	→	a bananas\$panam
s\$panamabanana	→	am abananas\$pan
as\$panamabanana	→	anam abananas\$p
nas\$panamabana		
anas\$panamaban		
anas\$panamaba		
nanas\$panamaba		
ananas\$panamab		
ananas\$panamab		
bananas\$panama		
abananas\$panam		
mabananas\$pana		
amabananas\$pan		
namabananas\$pa		
anamabananas\$p		

Sort the strings
lexicographically
(\$ comes first)

Sorting Cyclic Rotations

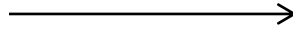
panamabananas\$
\$panamabananas
s\$panamabananas
as\$panamabananas
nas\$panamabananas
anas\$panamabananas
nanas\$panamabananas
bananas\$panamabananas
abanas\$panamabananas
mabananas\$panamabananas
amabananas\$panamabananas
namabananas\$panamabananas
anamabananas\$panamabananas

\$panamabananas
abananas\$panamabananas
amabananas\$panamabananas
anamabananas\$panamabananas
ananas\$panamabananas

Sort the strings
lexicographically
(\$ comes first)

Sorting Cyclic Rotations

panamabananas\$
\$panamabananas
s\$panamabanan
as\$panamabanan
nas\$panamabana
anas\$panamaban
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa
anamabananas\$p

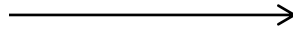


\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$panamab
anas\$panamaban
as\$panamabanan
babananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabanan

Sort the strings
lexicographically
(\$ comes first)

BWT(panamabananas\$)=smnbnbnaaaa\$a

panamabananas\$
\$panamabananas
s\$panamabananas
as\$panamabanan
nas\$panamabana
anas\$panamaban
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa
anamabananas\$p



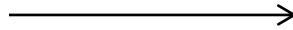
\$panamabananas**s**
abananas\$pana**m**
amabananas\$pa**n**
anamabananas\$b**p**
ananas\$panama**b**
anas\$panamaba**n**
as\$panamabana**n**
bananas\$panam**a**
mabananas\$pan**a**
namabananas\$bpa**a**
nanas\$panamab**a**
nas\$panamaban**a**
panamabananas\$b
s\$panamabanan**a**

All cyclic rotations of
“panamabananas\$”

Burrows-Wheeler Transform (BWT):
Last column = **smnpbnnaaaaa\$a**

BWT(p**an**a**m**a**b**a**n**a**s**\$)=s**m**n**p**b**n**n**a****a****a****a**\$**a**

panamabananas\$
\$panamabananas
s\$panamabanan
as\$panamabanan
nas\$panamabana
anas\$panamaban
nanas\$panamaba
ananas\$panamab
bananas\$panama
abananas\$panam
mabananas\$pana
amabananas\$pan
namabananas\$pa
anamabananas\$pa



\$panamabananas**s**
abananas\$pana**m**
amabananas\$pa**n**
anamabananas\$p**p**
ananas\$panama**b**
anas\$panamaba**n**
as\$panamabana**n**
bananas\$panam**a**
mabananas\$pan**a**
namabananas\$p**a**
nanas\$panamab**a**
nas\$panamaban**a**
panamabananas\$**s**
s\$panamabanan**a**

All cyclic rotations of
“panamabananas\$”

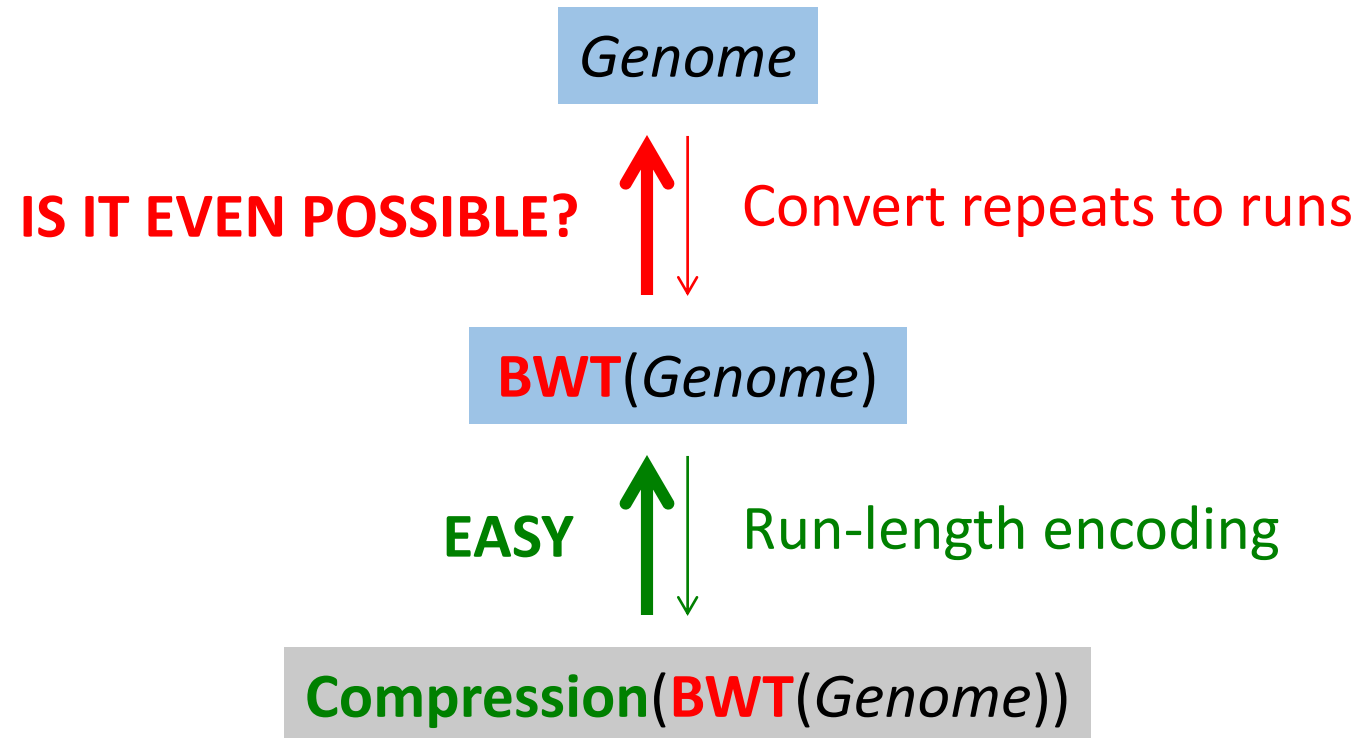
Burrows-Wheeler Transform (BWT):
Last column = **smnpbnnaaaaa\$a**

Applying BWT to the Double Helix Paper by Watson&Crick

nd Corey (1). They kindly made their manuscript availa a
nd criticism, especially on interatomic distances. We a
nd cytosine. The sequence of bases on a single chain d a
nd experimentally (3,4) that the ratio of the amounts o u
nd for this reason we shall not comment on it. We wish a
nd guanine (purine) with cytosine (pyrimidine). In oth a
nd ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin a
nd its water content is rather high. At lower water co a
nd pyrimidine bases. The planes of the bases are perpe a
nd stereochemical arguments. It has not escaped our no a
nd that only specific pairs of bases can bond together u
nd the atoms near it is close to Furberg's 'standard co a
nd the bases on the inside, linked together by hydrogen a
nd the bases on the outside. In our opinion, this stru a
nd the other a pyrimidine for bonding to occur. The hy a
nd the phosphates on the outside. The configuration of a
nd the ration of guanine to cytosine, are always very c a
nd the same axis (see diagram). We have made the usual u
nd their co-workers at King's College, London. One of a

“and” is a frequent repeat in English texts

Going Back From BWT(*Genome*) to *Genome*



Outline

- Burrows-Wheeler Transform
- **Inverting Burrows-Wheeler Transform**
- Using BWT for Pattern Matching
- Suffix Arrays
- Approximate Pattern Matching

Reconstructing banana from annb\$aa

\$ b a n a n **a**
a \$ b a n a **n**
a n a \$ b a **n**
a n a n a \$ **b**
b a n a n a \$
n a \$ b a n **a**
n a n a \$ b **a**

Reconstructing banana

\$ b a n a n a
a **\$** b a n a n
a n a **\$** b a n
a n a n a **\$** b
b a n a n a **\$**
n a **\$** b a n a
n a n a **\$** b a

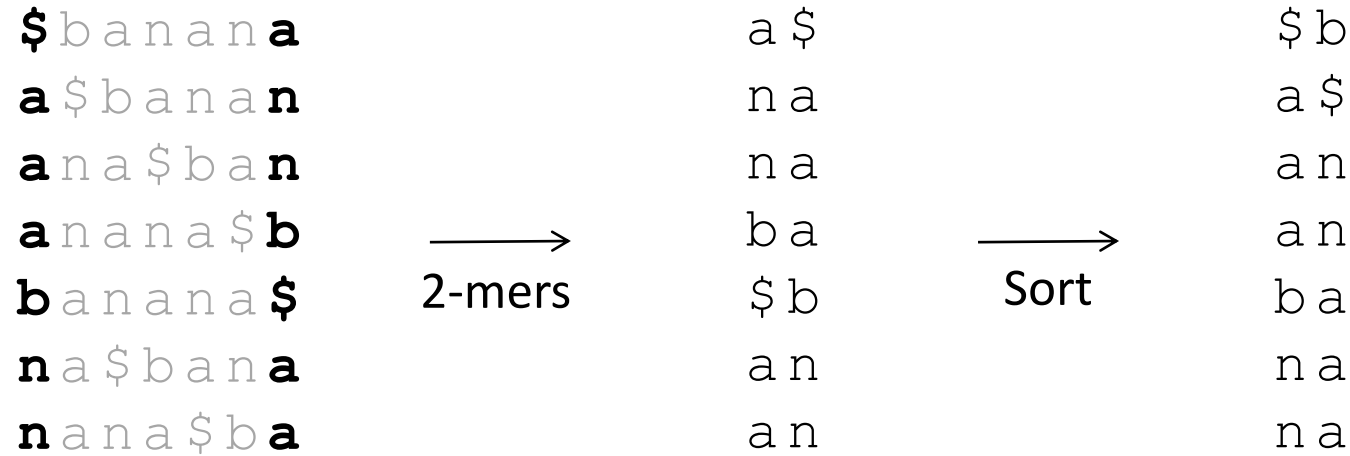
- Sorting all elements of “annb\$aa” gives first column of BWT matrix.

Reconstructing banana

\$ b a n a n a a		a \$
a \$ b a n a n n		n a
a n a \$ b a n n		n a
a n a n a \$ b	→	b a
b a n a n a \$	2-mers	\$ b
n a \$ b a n a		a n
n a n a \$ b a		a n

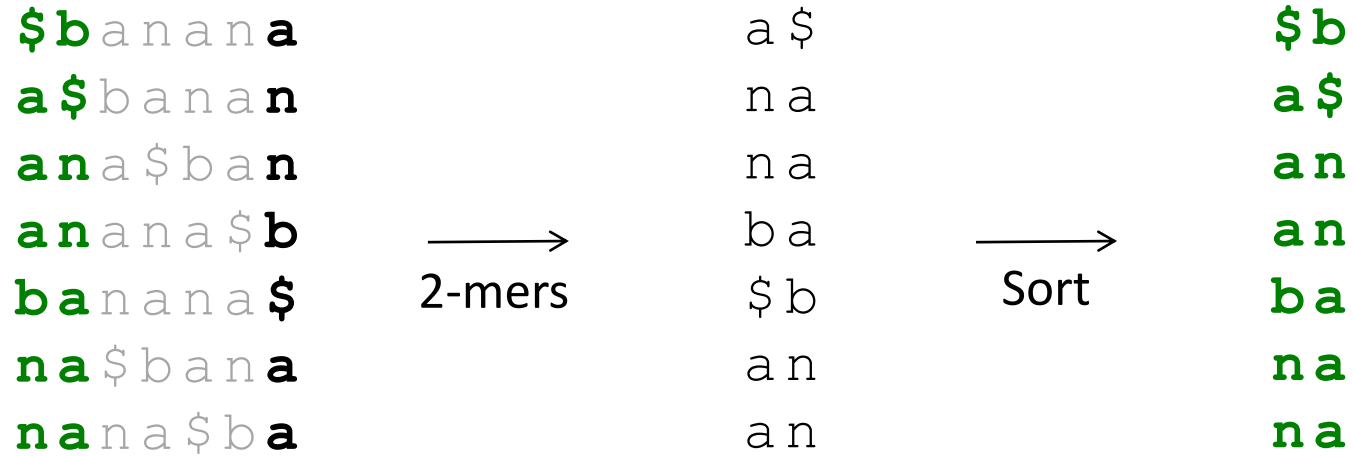
- We now know 2-mer composition of the circular string banana\$

Reconstructing banana



- We now know 2-mer composition of the circular string banana\$
- Sorting gives us the first 2 columns of the matrix.

Reconstructing banana



- We now know 2-mer composition of the circular string banana\$
- Sorting gives us the first 2 columns of the matrix.

Reconstructing banana

\$b a n a n a

a **\$b** a n a n

a n a **\$b** a n

a n a n a **\$b**

b a n a n a **\$**

n a **\$b** a n a

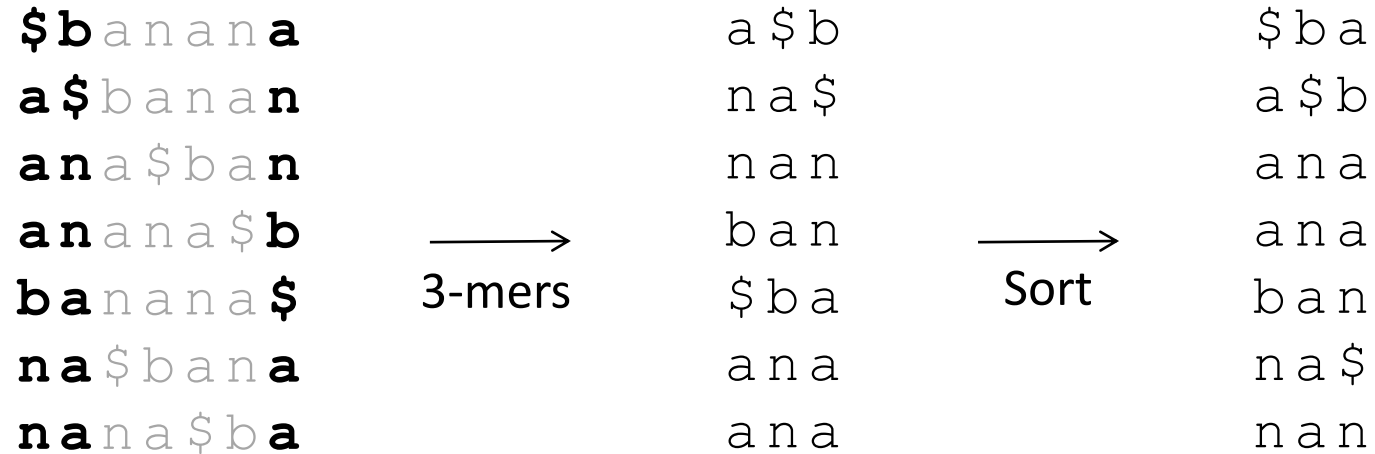
n a n a **\$b** a

Reconstructing banana

\$ b a n a n a		a \$ b
a \$ b a n a n		n a \$
a n a \$ b a n		n a n
a n a n a \$ b	→	b a n
b a n a n a \$	3-mers	\$ b a
n a \$ b a n a		a n a
n a n a \$ b a		a n a

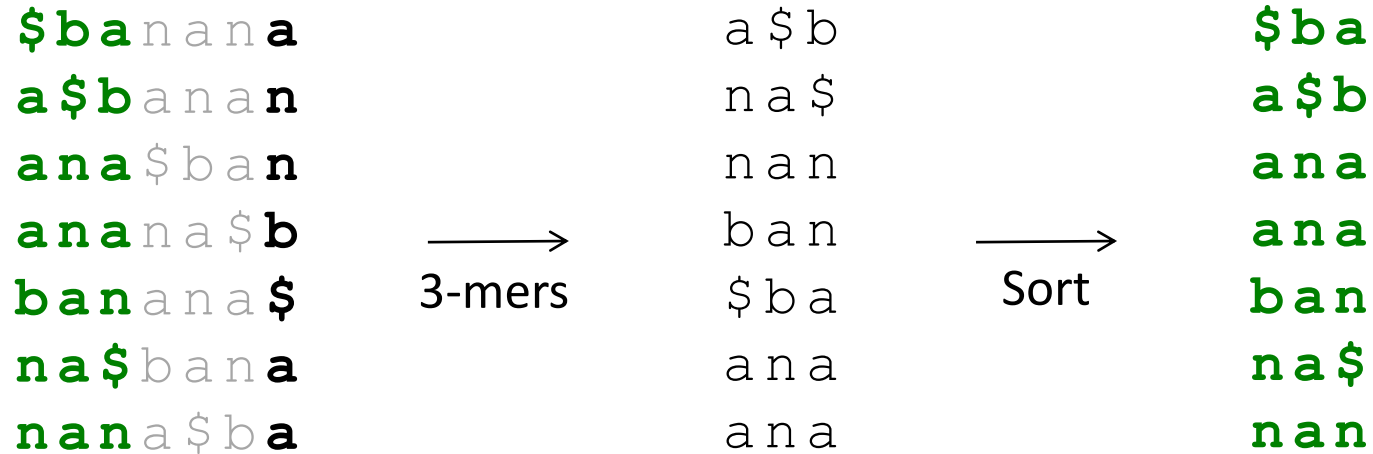
- We now know 3-mer composition of the circular string banana\$

Reconstructing banana



- We now know 3-mer composition of the circular string banana\$
- Sorting gives us the first 3 columns of the matrix.

Reconstructing banana



- We now know 3-mer composition of the circular string banana\$
- Sorting gives us the first 3 columns of the matrix.

Reconstructing banana

\$banana

a**\$b**anan

ana**\$b**an

ana**\$b**ana

ban**\$b**ana

na**\$b**ana

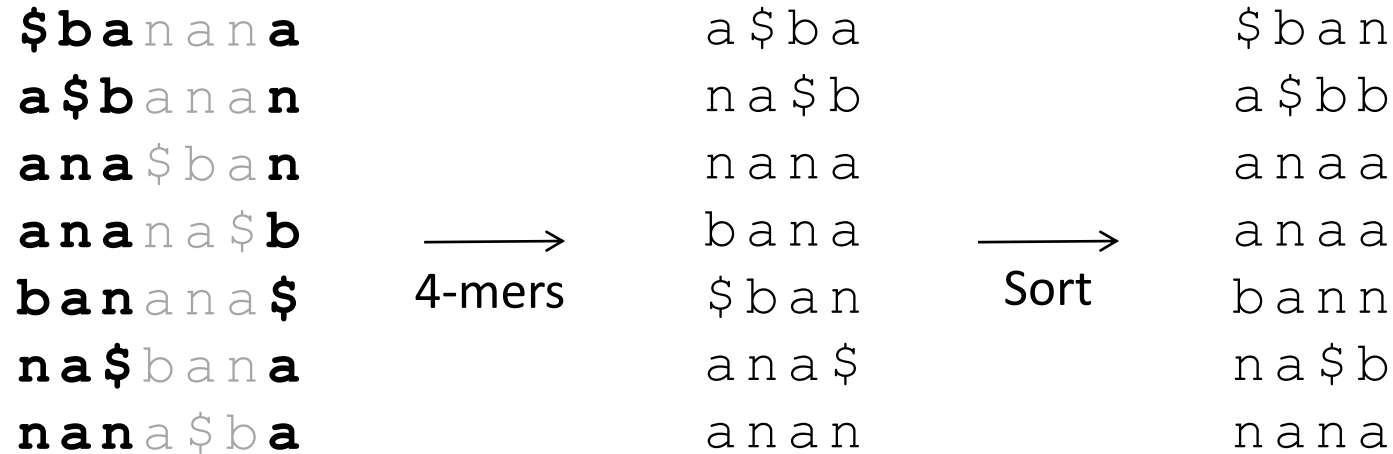
nan**\$b**ana

Reconstructing banana

\$b anana		a\$ba
a\$b anan		na\$b
ana \$ban		nana
ana na\$b	→	bana
ban ana\$	4-mers	\$ban
na \$bana		ana\$
nan a\$ba		anan

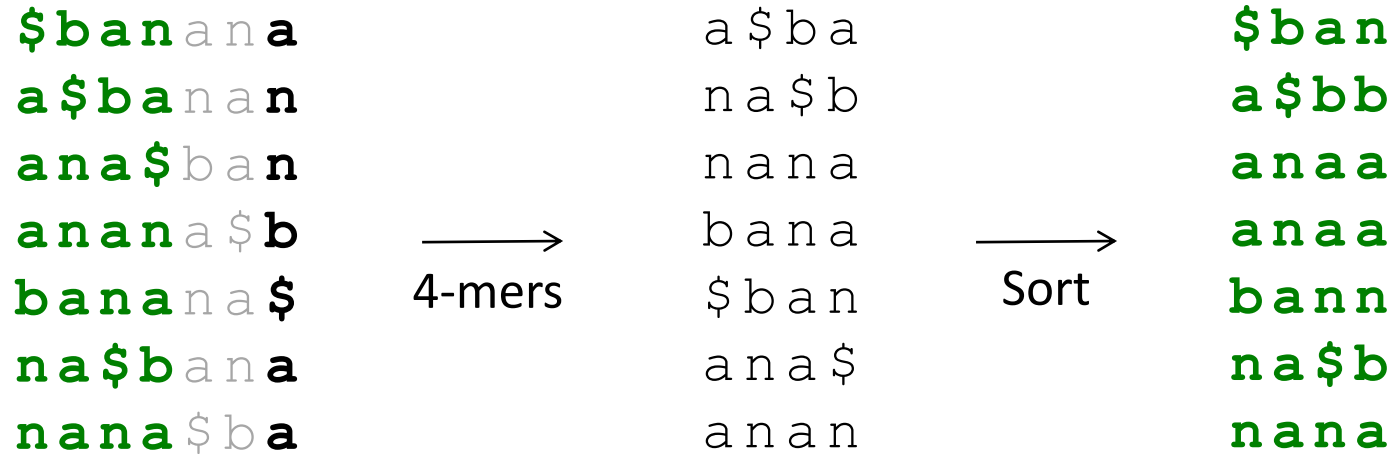
- We now know 4-mer composition of the circular string banana\$

Reconstructing banana



- We now know 4-mer composition of the circular string banana\$
- Sorting gives us the first 4 columns of the matrix.

Reconstructing banana



- We now know 4-mer composition of the circular string banana\$
- Sorting gives us the first 4 columns of the matrix.

Reconstructing banana

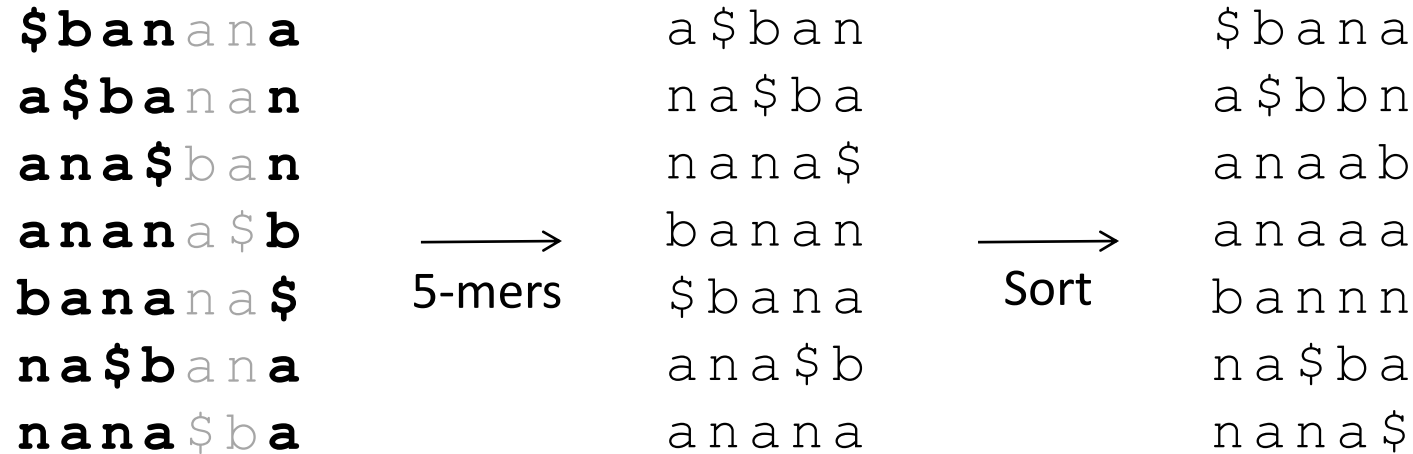
\$bana
a\$ban
ana\$b
anana\$b
banana\$
na\$ban
nana\$ba

Reconstructing banana

\$ banan a		a\$ban
a \$banan n		na\$b
ana \$ban		nana\$
anan a\$b	→	banan
ban ana\$	5-mers	\$bana
na \$bana		ana\$b
nana \$ba		anana

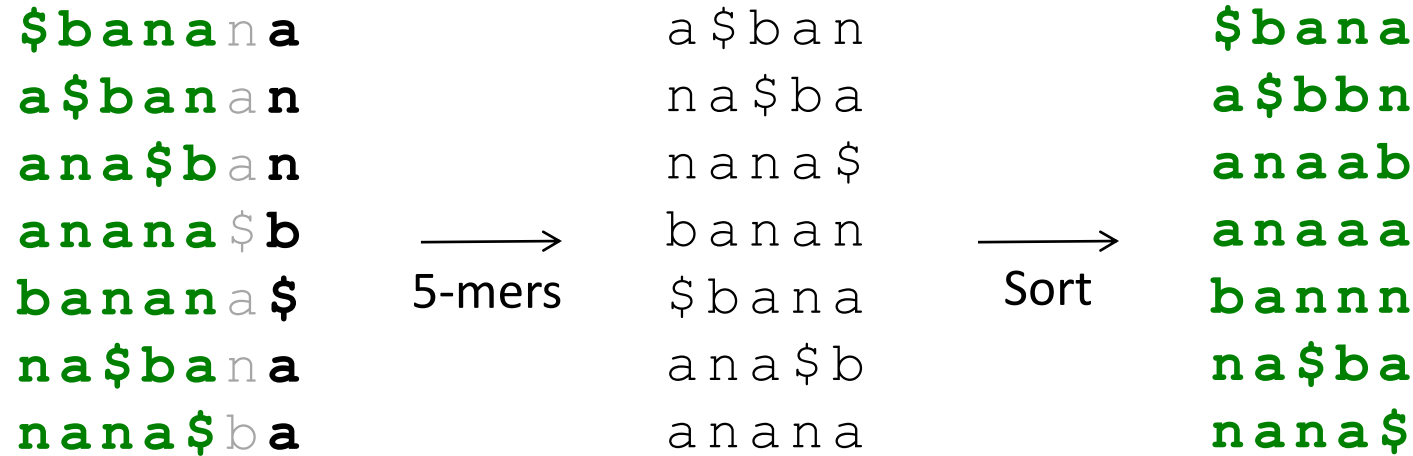
- We now know 5-mer composition of the circular string banana\$

Reconstructing banana



- We now know 5-mer composition of the circular string banana\$
- Sorting gives us the first 5 columns of the matrix.

Reconstructing banana



- We now know 5-mer composition of the circular string banana\$
- Sorting gives us the first 5 columns of the matrix.

Reconstructing banana

\$bana n a
a \$ban a n
ana \$b a n
anana \$b
banan a \$
na \$ba n a
nana \$b a

Reconstructing banana

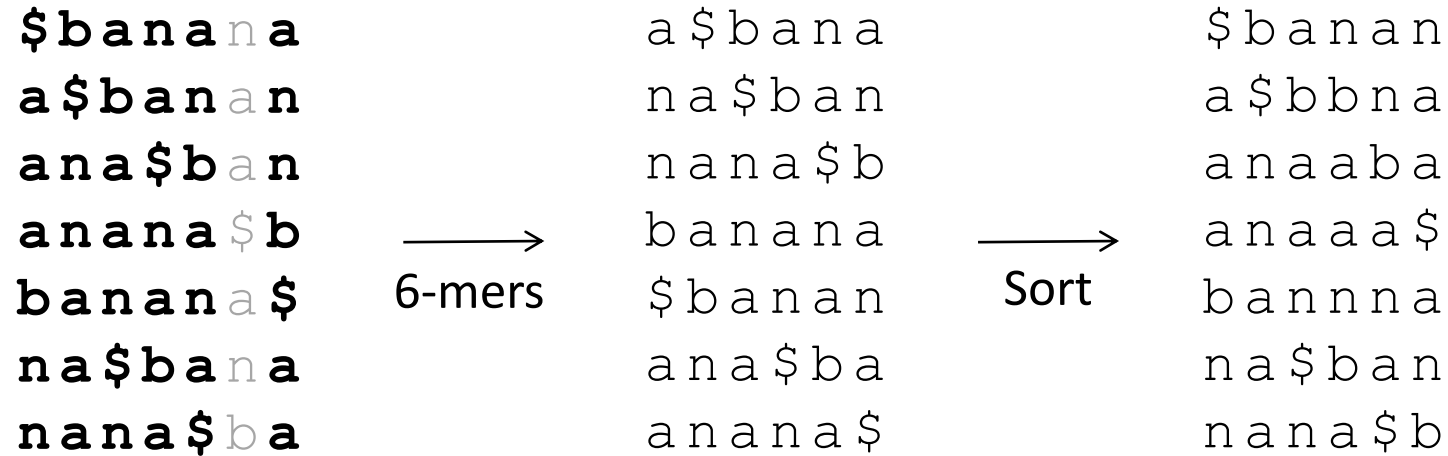
\$bana	n	a	
a\$bana	n		
ana\$b	a	n	
anana\$b			
banana	\$		
na\$b	a	n	a
nana	\$	b	a

→ 6-mers

a\$bana
na\$bana
nana\$b
banana
\$banan
ana\$ba
anana\$

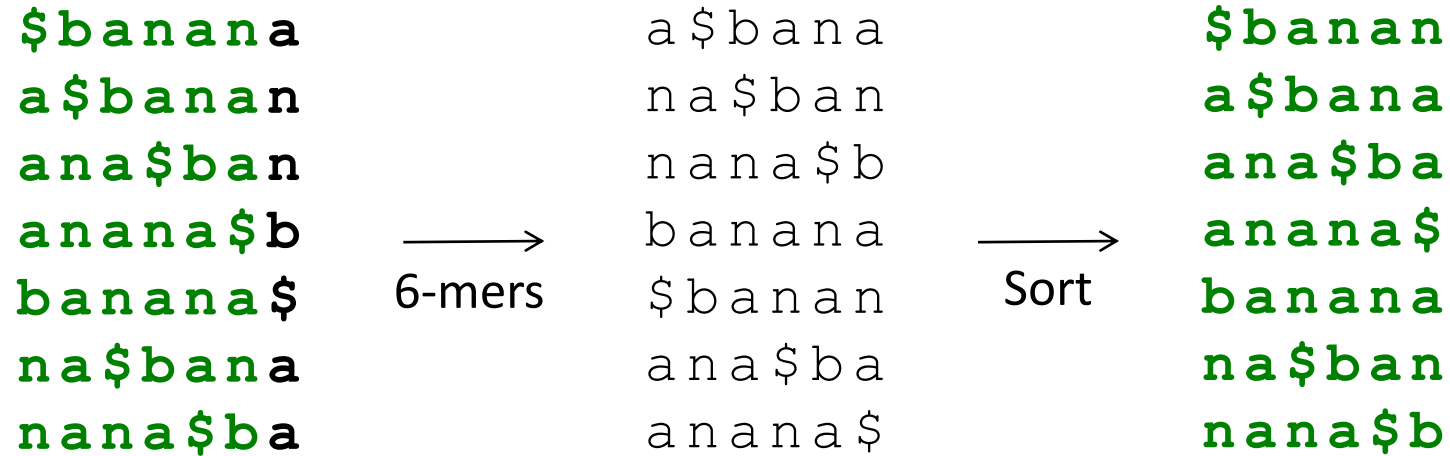
- We now know 6-mer composition of the circular string banana\$

Reconstructing banana



- We now know 6-mer composition of the circular string banana\$
- Sorting gives us the first 6 columns of the matrix.

Reconstructing banana



- We now know 6-mer composition of the circular string banana\$
- Sorting gives us the first 6 columns of the matrix.

Reconstructing banana

\$banana
a\$banan
ana\$ban
anana\$b
banana\$
na\$bana
nana\$ba

- We now know the entire matrix!

Reconstructing banana

\$banana

a \$ b a n a n

a n a \$ b a n

a n a n a \$ b

b a n a n a \$

n a \$ b a n a

n a n a \$ b a

- We now know the entire matrix!
- Symbols in the first row (after \$) spell **banana**.

More Memory Issues

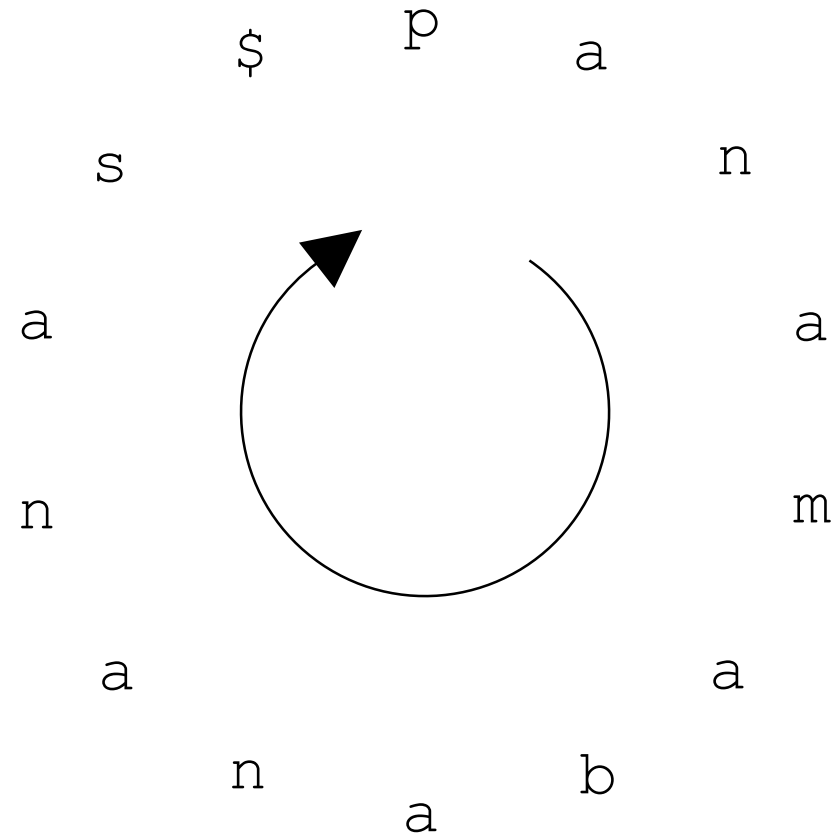
- Reconstructing *Text* from $BWT(Text)$ required us to store $|Text|$ cyclic rotations of $|Text|$.

```
$b a n a n a  
a $b a n a n  
a n a $b a n  
a n a n a $b  
b a n a n a $  
n a $b a n a  
n a n a $b a
```

- Can we invert $BWT(Text)$ with less space and without $|Text|$ rounds of sorting?

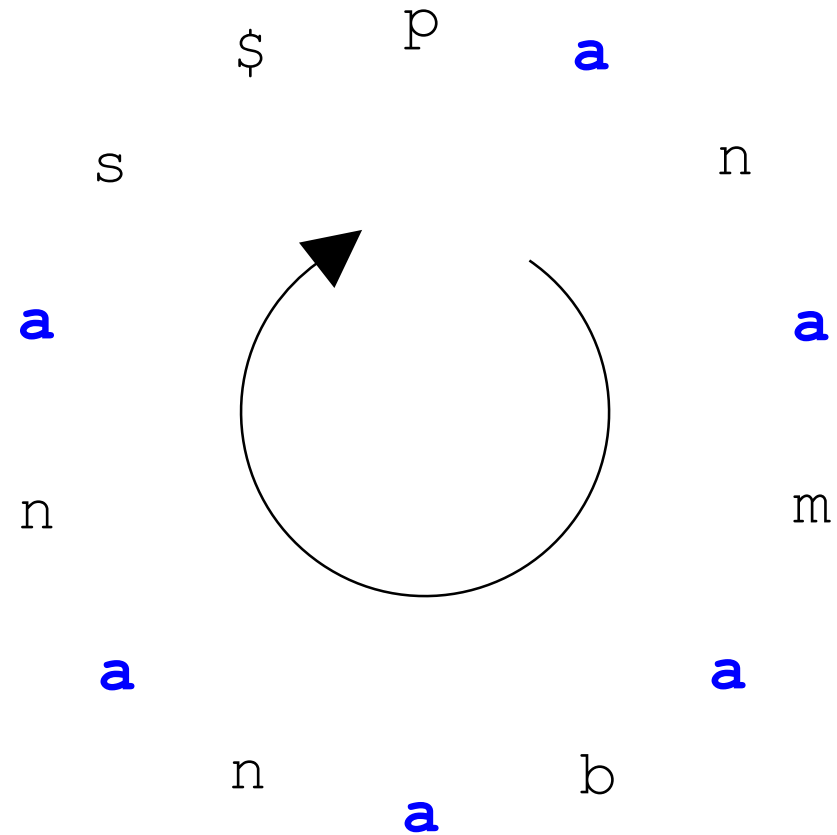
A Strange Observation

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$panamab
anas\$panamaban
as\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabannana



A Strange Observation

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$panamab
anas\$panamaban
as\$panamabanan
bananas\$panam**a**
mabananas\$pan**a**
namabananas\$p**a**
nanas\$panamab**a**
nas\$panamaban**a**
panamabananas\$
s\$panamabanan**a**



A Strange Observation

Where

is first

“a”

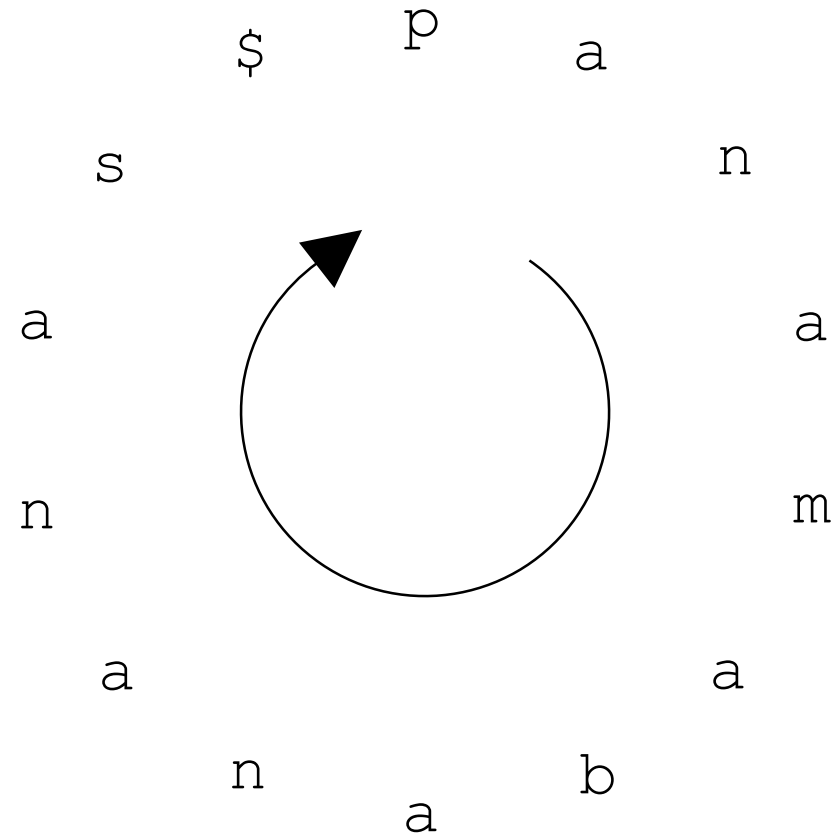
hiding

inside

the

circle?

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$panamab
anas\$panamaban
as\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabanana



A Strange Observation

Where

is first

"a"

hiding

inside

the

circle?

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$panamab
anas\$panamaban
as\$panamabanan
bananas\$panam(a
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabanana

Where

is first

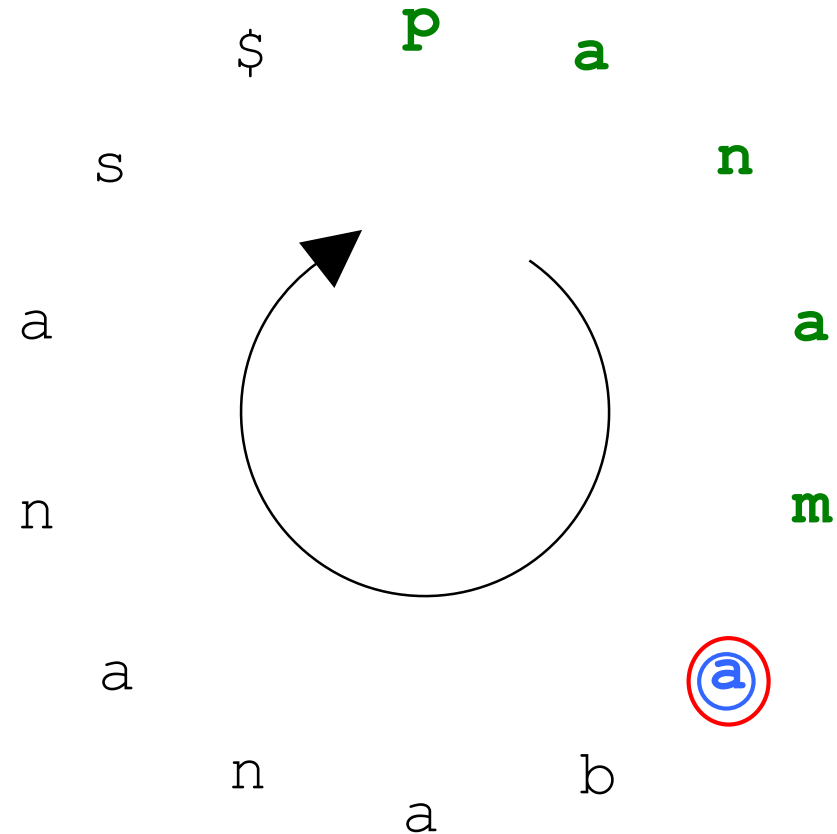
"a"

hiding

inside

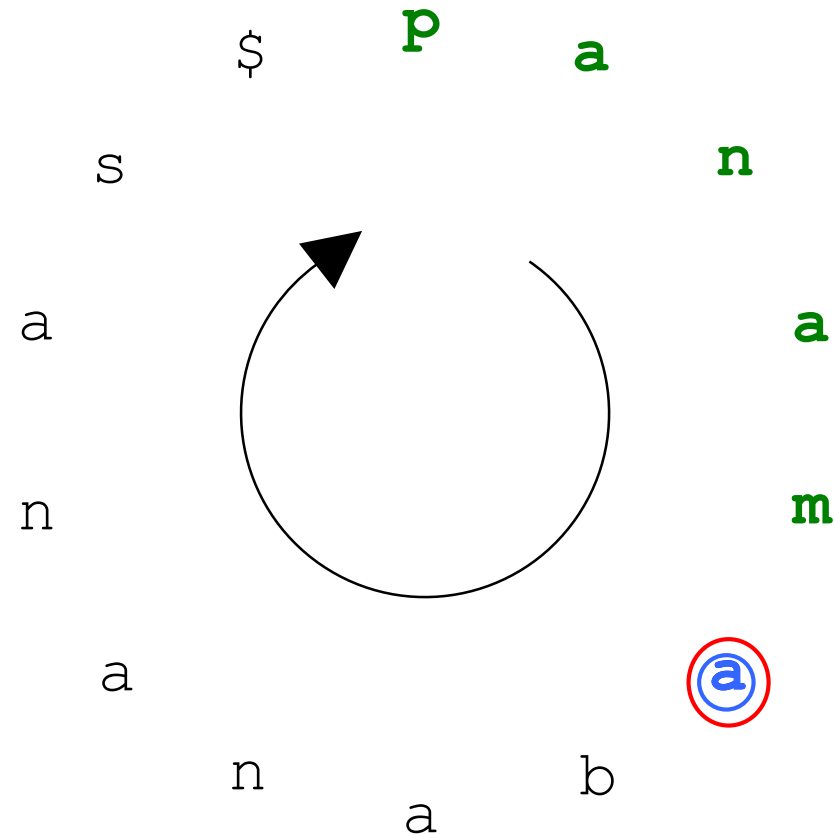
the

circle?



They Are Hiding at the Same Position!

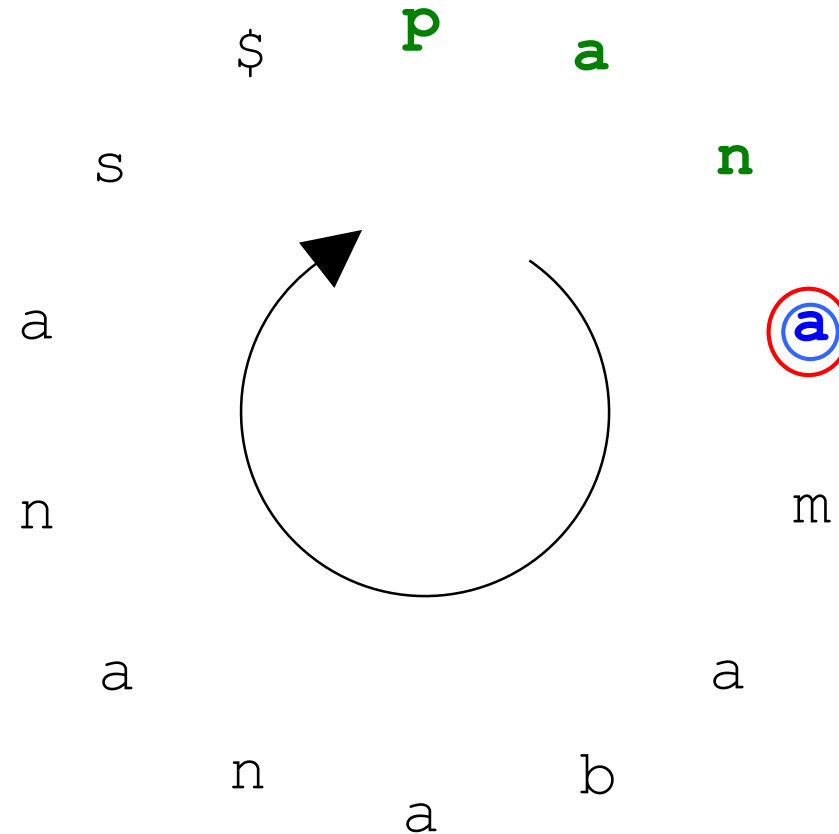
\$panamabananas
a bananas \$panam
amabananas \$pan
anamabananas \$p
ananas \$panamab
anas \$panamaban
as \$panamaban
bananas \$panam
mabananas \$pana
namabananas \$pa
nanas \$panamaba
nas \$panamabana
panamabananas \$
s \$panamabana



1st a in *FirstColumn* and 1st a in *LastColumn*
are hiding at the same position along the cycle!

Another Strange Observation

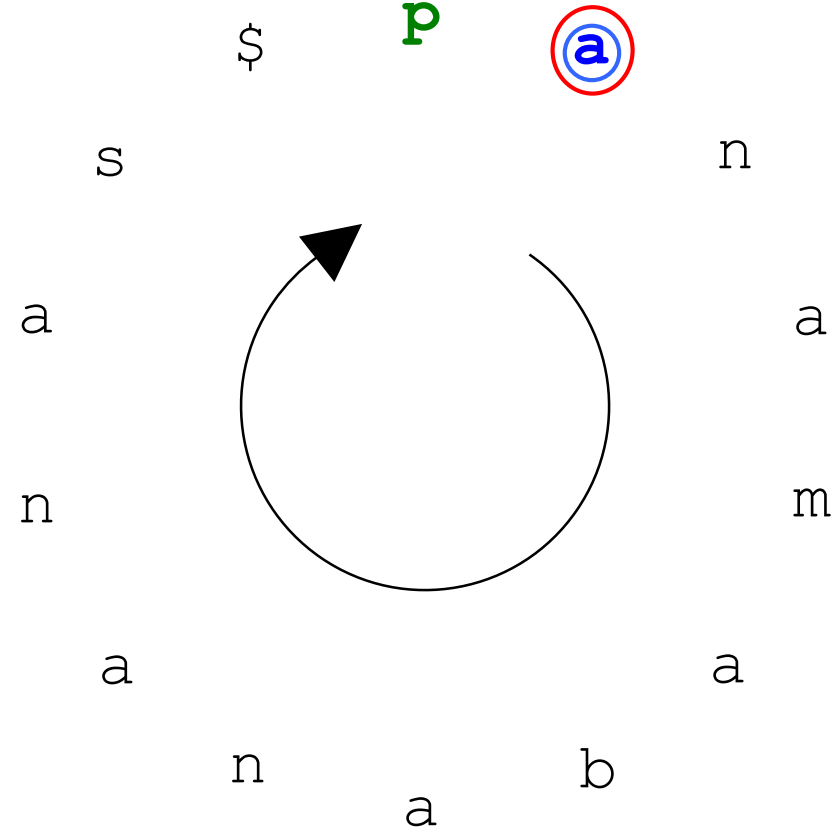
\$panamabananas
abananas\$panam
amabananas\$**pan**
anamabananas\$p
ananas\$panamab
anas\$panamaban
as\$panamabanan
bananas\$panama
mabananas\$**pana****a**
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabana



2nd **a** in *FirstColumn* and 2nd **a** in *LastColumn*
are hiding at the same position along the cycle!

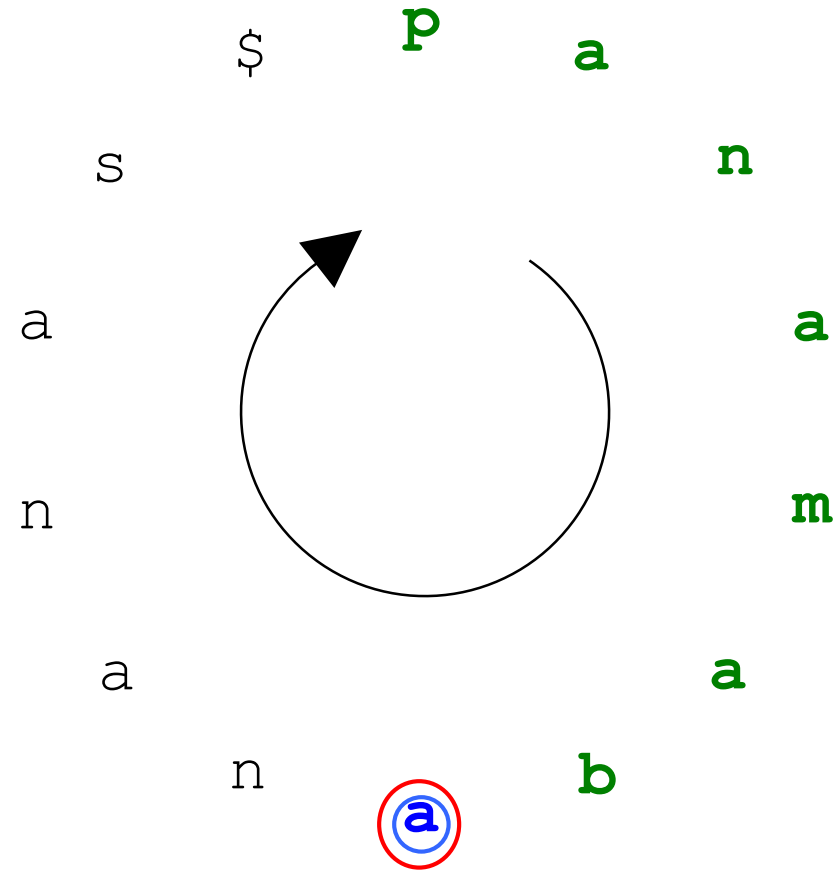
Another Strange Observation

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$**p**
ananas\$panamab
anas\$panamaban
as\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$**p****a**
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabannana



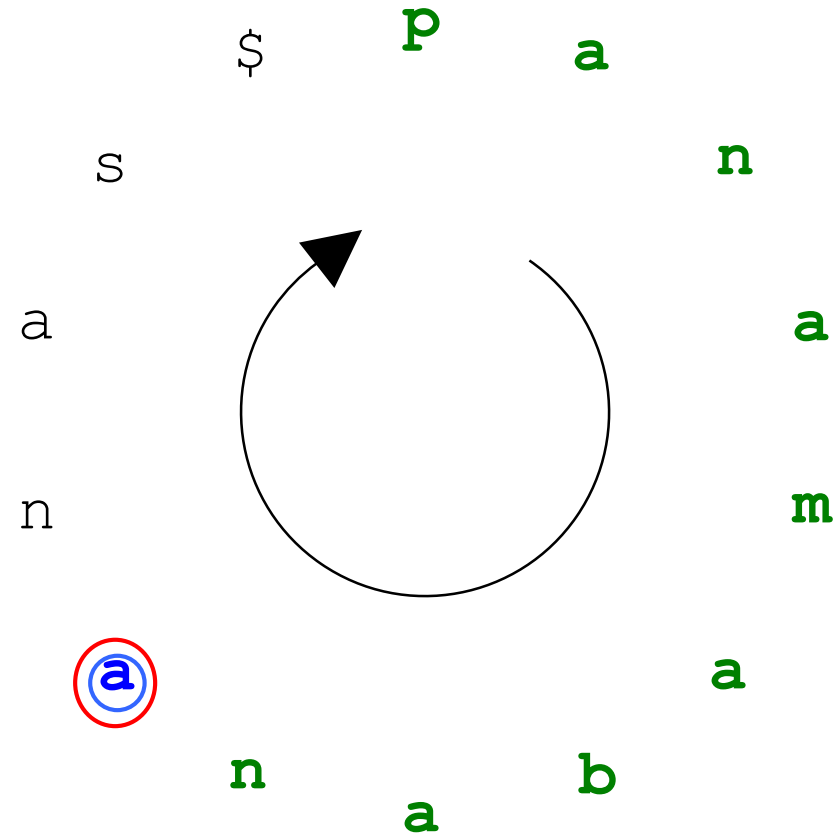
Another Strange Observation

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$**panamab**
anas\$panamaban
as\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$**panamaba**
nas\$panamabana
panamabananas\$
s\$panamabannana



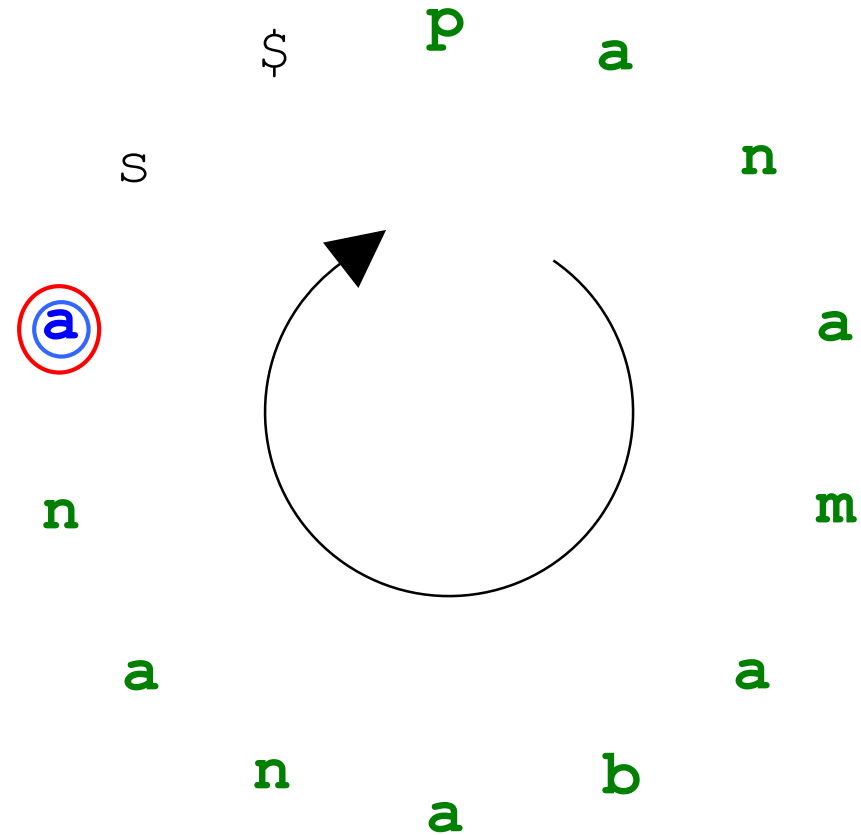
Another Strange Observation

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$panamab
anas\$**panamaban**
as\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$**panamabana**
panamabananas\$
s\$panamabannana



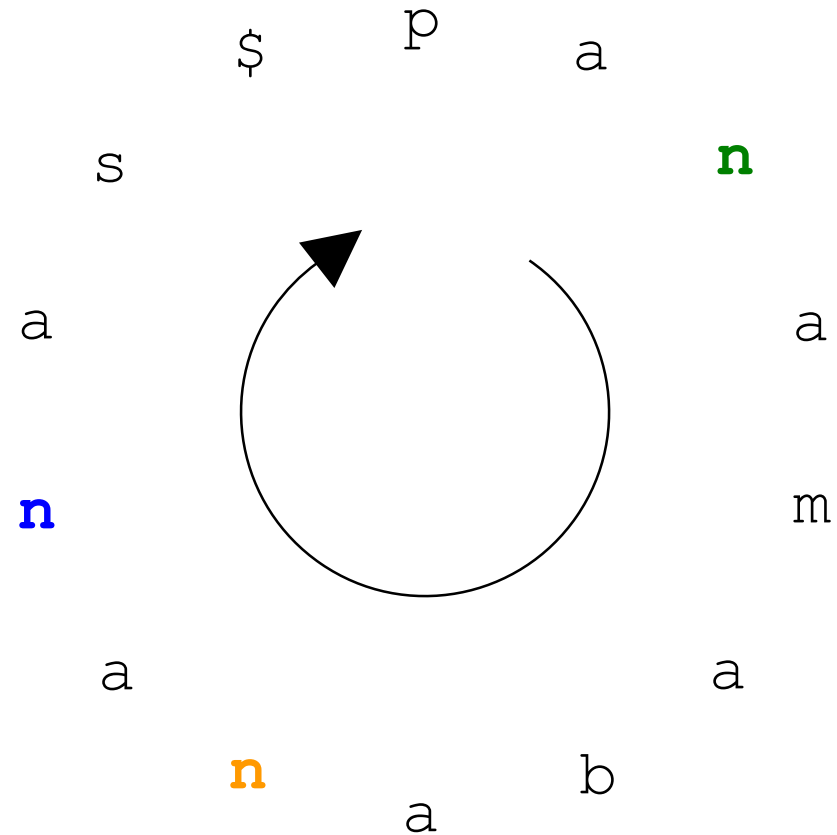
Another Strange Observation

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$panamab
anas\$panamaban
as\$**p****a****n****a****m****a****b****a****n****a****n**
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$**p****a****n****a****m****a****b****a****n****a****n****a**



Another Strange Observation

\$panamabananas
abananas\$panam
amabananas\$pan
anamabananas\$p
ananas\$panamab
anas\$panamaba
as\$panamabana
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabanana



Is It True in General?

```
$panamabananas  
1 abananas$panam  
2 amabananas$pan  
3 anabananas$p  
4 ananas$panamab  
5 anas$panamaban  
6 as$panamabanan  
bananas$panama  
mabananas$pana  
namabananas$pa  
nanas$panamaba  
nas$panamabana  
panamabananas$  
s$panamabana
```

These strings are sorted

Is It True in General?

\$panamabananas
1 a bananas\$panam
2 amabananas\$pan
3 anabananas\$p
4 ananas\$panamab
5 anas\$panamaban
6 as\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabannan

Chop off a

bananas\$panam
mabananas\$pan
namabananas\$p
nanas\$panamab
nas\$panamaban
s\$panamabanan

These strings are sorted

Is It True in General?

\$panamabananas
1 a bananas\$panam
2 amabananas\$pan
3 anabananas\$p
4 ananas\$panamab
5 as\$panamaban
6 as\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabannan

Chop off a

bananas\$panam
mabananas\$pan
namabananas\$p
nanas\$panamab
nas\$panamaban
s\$panamabanan

Still
sorted

These strings are sorted

Is It True in General?

\$panamabananas
1 **a**bananas\$panam
2 **a**mabananas\$pan
3 **a**namabananas\$p
4 **a**nanas\$panamab
5 **a**nas\$panamaban
6 **a**s\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabanana

These strings are sorted

Chop off **a**

bananas\$panam
mabananas\$pan
namabananas\$p
nanas\$panamab
nas\$panamaban
s\$panamabanan

Still
sorted

Add **a**
to end

bananas\$panam**a**
mabananas\$pan**a**
namabananas\$p**a**
nanas\$panamab**a**
nas\$panamaban**a**
s\$panamabanan**a**

Is It True in General?

\$panamabananas
1 **a**bananas\$panam
2 **a**mabananas\$pan
3 **a**namabananas\$p
4 **a**nanas\$panamab
5 **a**nas\$panamaban
6 **a**s\$panamabanan
bananas\$panama
mabananas\$pana
namabananas\$pa
nanas\$panamaba
nas\$panamabana
panamabananas\$
s\$panamabannan

These strings are sorted

Chop off **a**

bananas\$panam
mabananas\$pan
namabananas\$p
nanas\$panamab
nas\$panamaban
s\$panamabanan

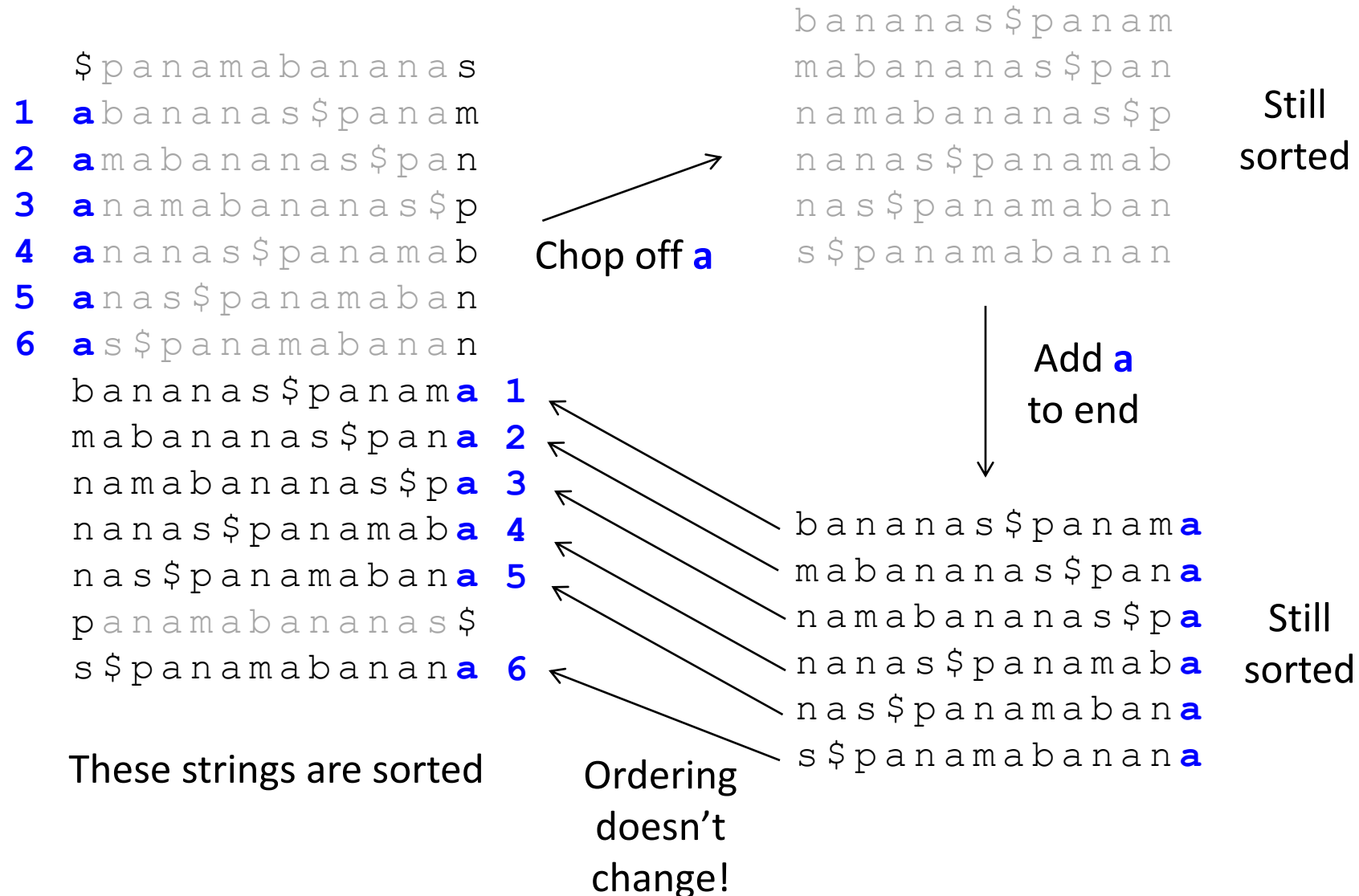
Still
sorted

Add **a**
to end

bananas\$panam**a**
mabananas\$pan**a**
namabananas\$p**a**
nanas\$panamab**a**
nas\$panamaban**a**
s\$panamabanan**a**

Still
sorted

Is It True in General?



First-Last Property

- the k -th occurrence of *symbol* in ***FirstColumn***
- and the k -th occurrence of *symbol* in ***LastColumn***
- correspond to appearance of *symbol* at the same position in *Text*.

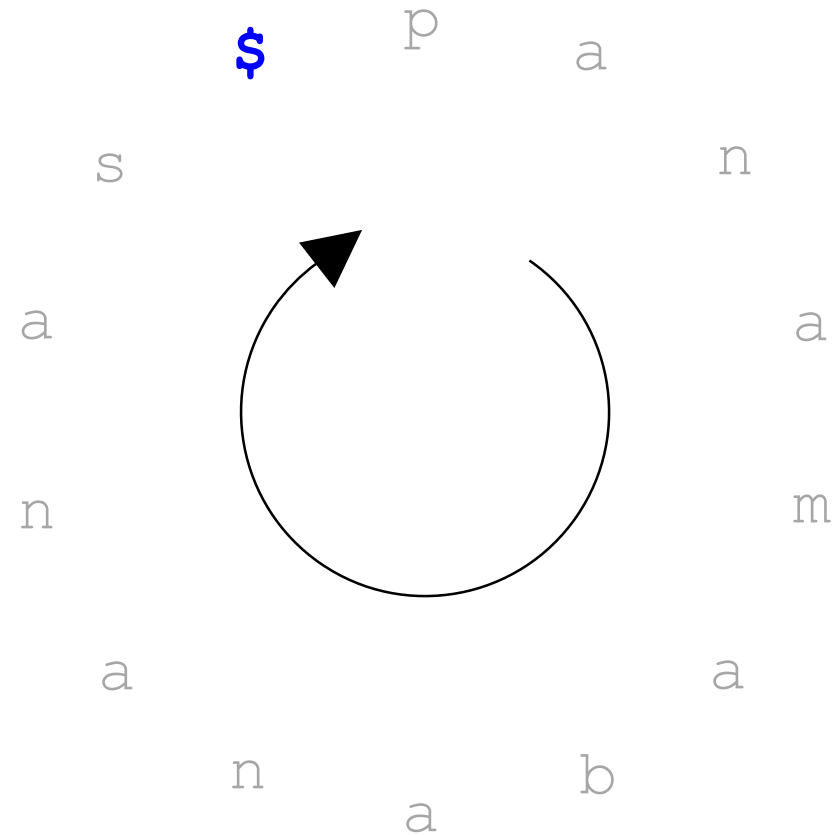
•

$p_1 a_3 n_1 a_2 m_1 a_1 b_1 a_4 n_2 a_5 n_3 a_6 s_1 \$_1$

$\$_1$ panamabanana s_1
 a_1 bananas\$panam m_1
 a_2 mabananas\$pan n_1
 a_3 namabananas\$ p_1
 a_4 nanas\$panamab b_1
 a_5 nas\$panamaban n_2
 a_6 s\$panamabanan n_3
 b_1 ananas\$panama a_1
 m_1 abananas\$pana a_2
 n_1 amabananas\$pa a_3
 n_2 anas\$panamaba a_4
 n_3 as\$panamabana a_5
 p_1 anamabananas $\$_1$
 s_1 \$panamabanana a_6

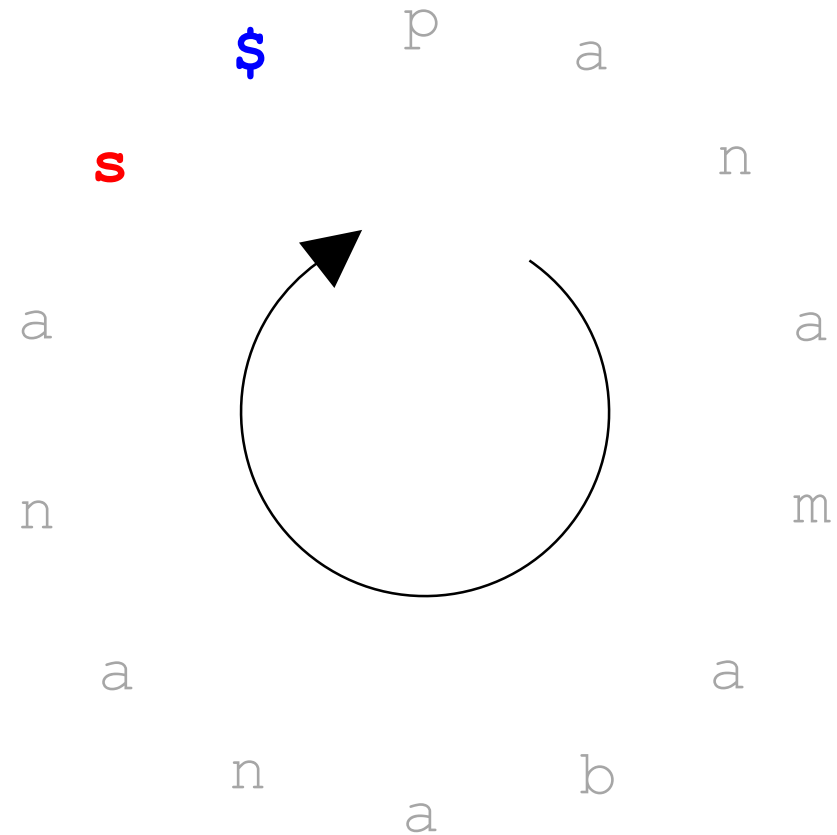
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



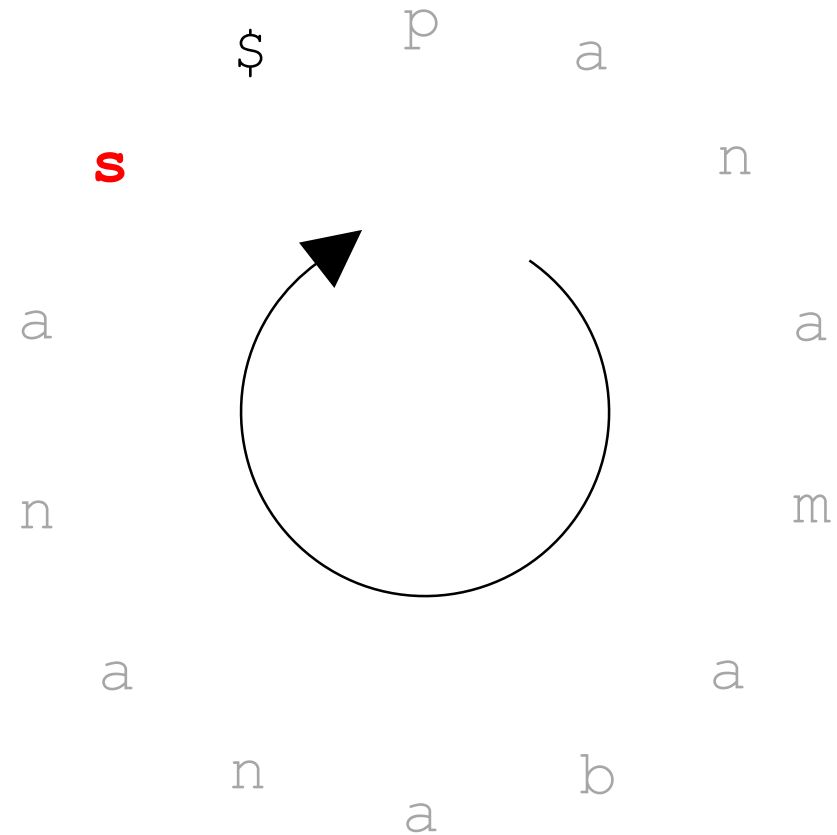
Inverting BWT Again

\$₁panamabanana**s**₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



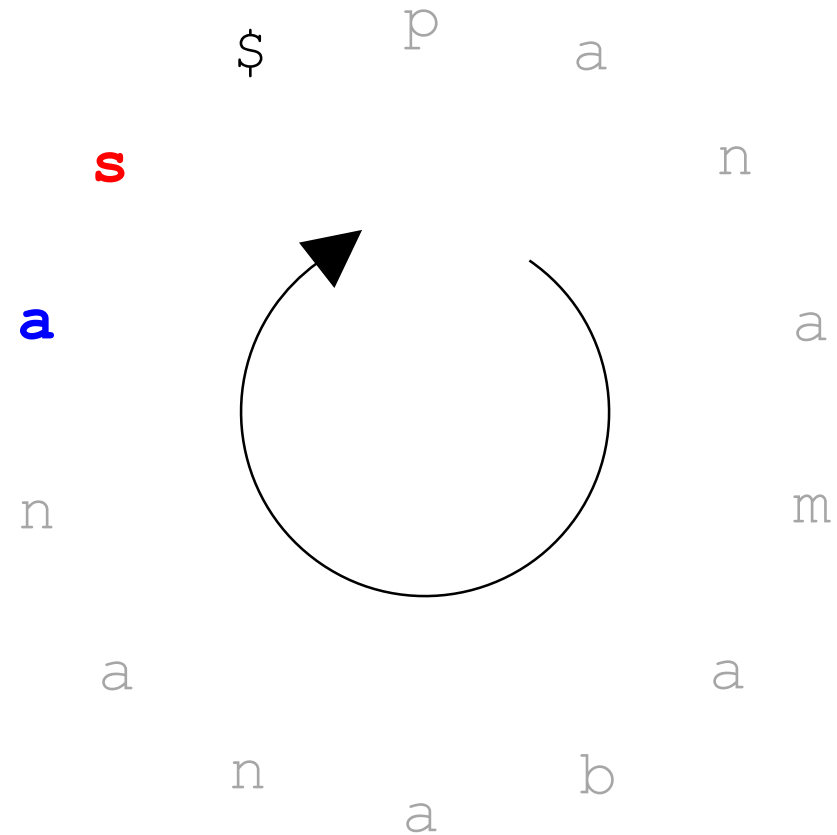
Inverting BWT Again

\$₁panamabanana **s**₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamabanana₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



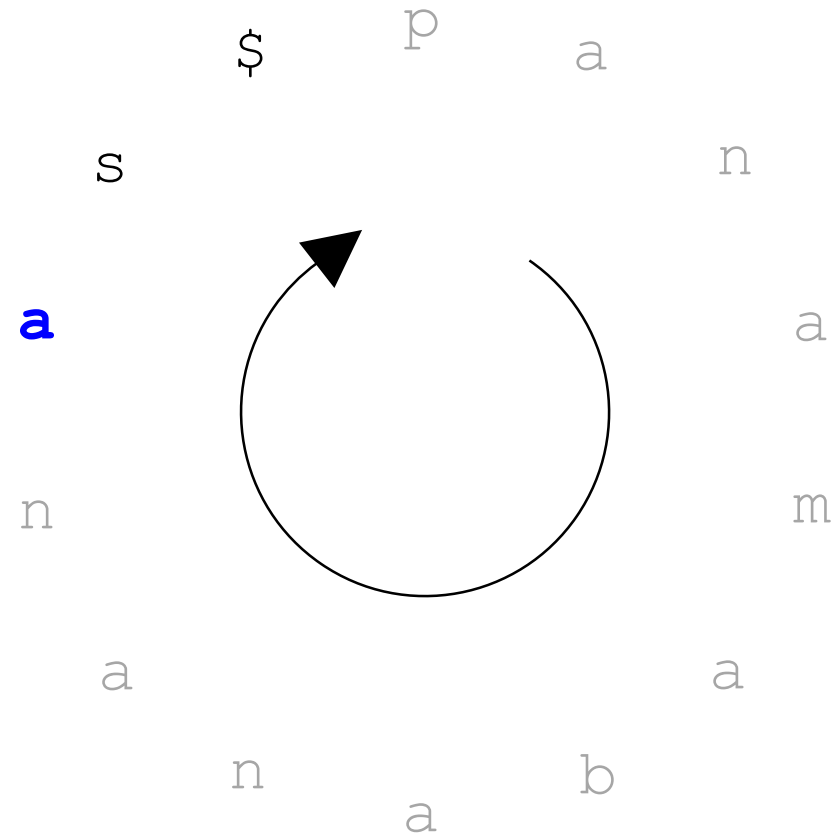
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamaban**a**₆



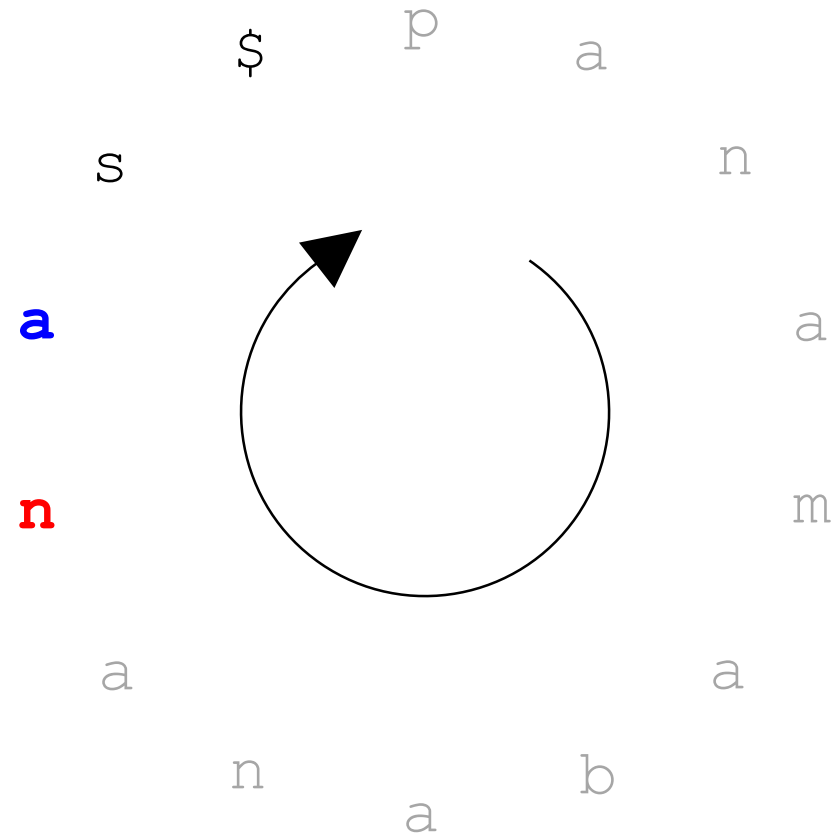
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamaban**a**₆



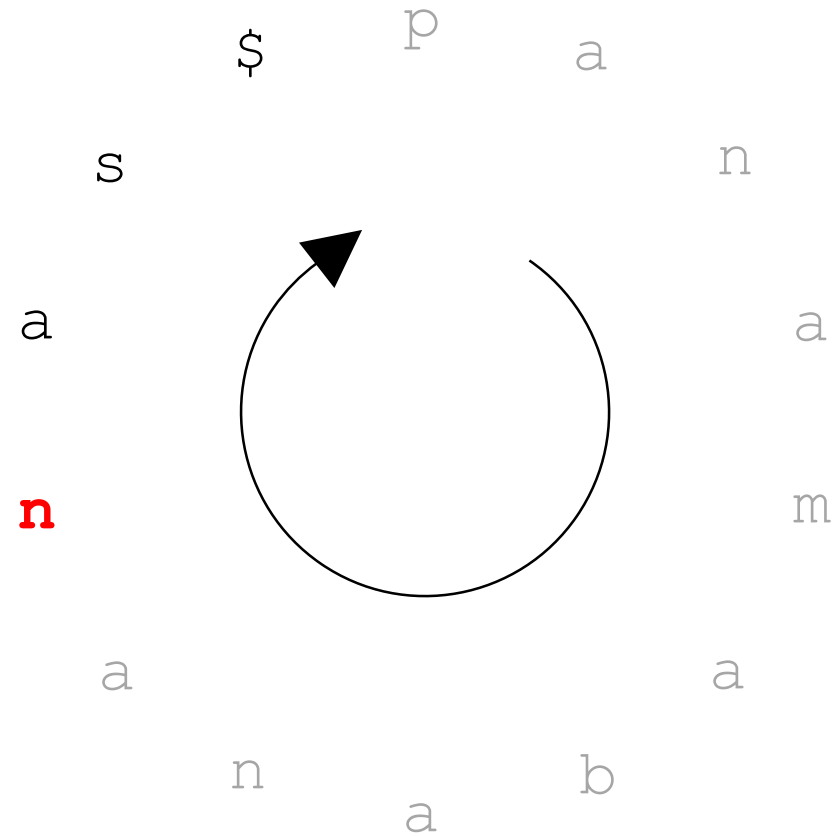
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamabana**n**₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanaa₆



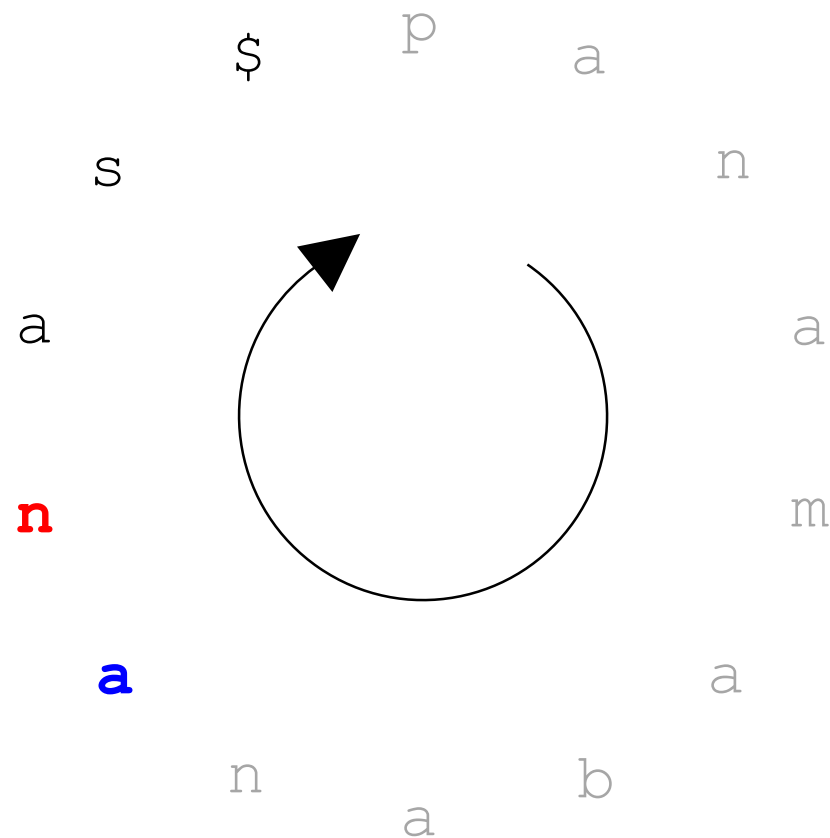
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamabana**n**₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



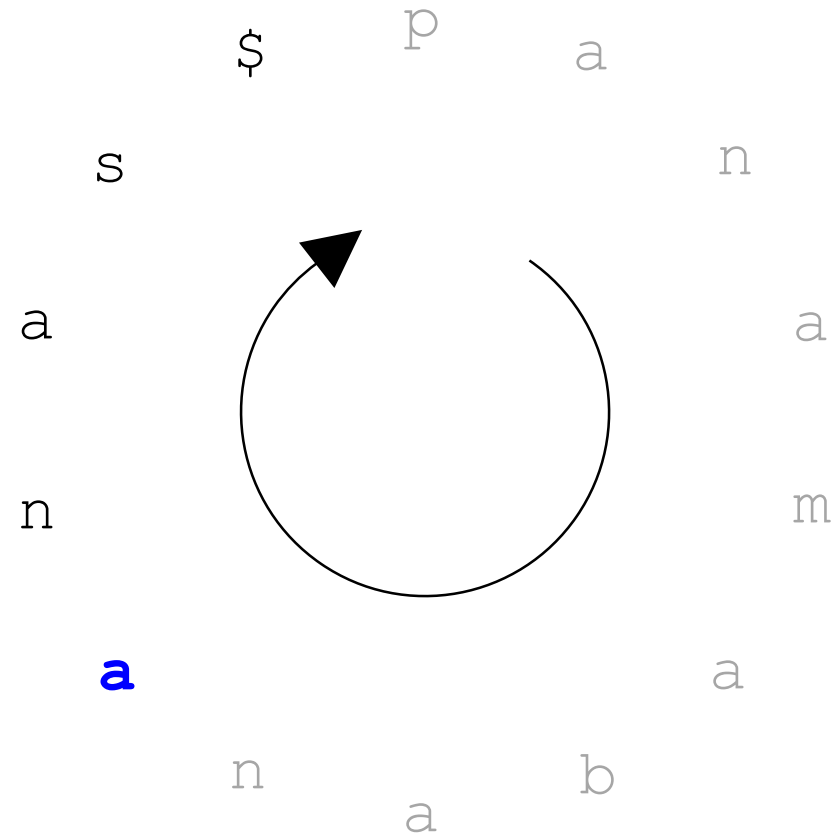
Inverting BWT Again

\$₁ panamabananas₁
 a₁ bananas \$panam₁
 a₂ mabananas \$pan₁
 a₃ namabananas \$p₁
 a₄ nanas \$panamab₁
 a₅ nas \$panamaban₂
 a₆ s \$panamabanan₃
 b₁ ananas \$panama₁
 m₁ abananas \$pana₂
 n₁ amabananas \$pa₃
 n₂ anas \$panamaba₄
n₃ as \$panamaban**a**₅
 p₁ anamabananas \$₁
 s₁ \$panamabanan a₆



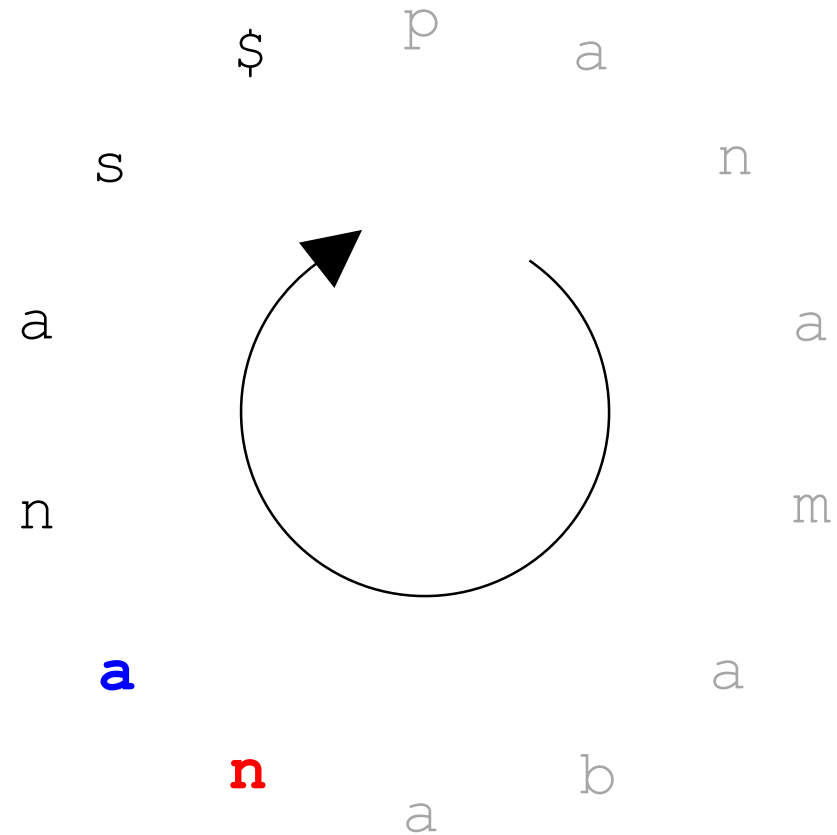
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamaban**a**₅
p₁anamabananas\$₁
s₁\$panamabanana₆



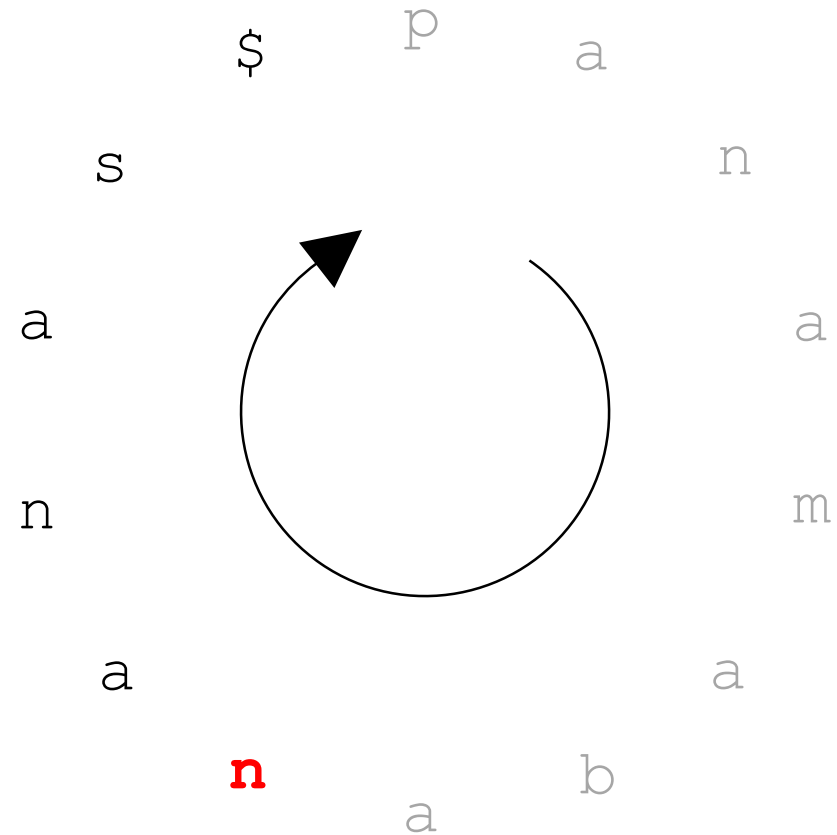
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaba**n**₂
a₆s\$panamabanana₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



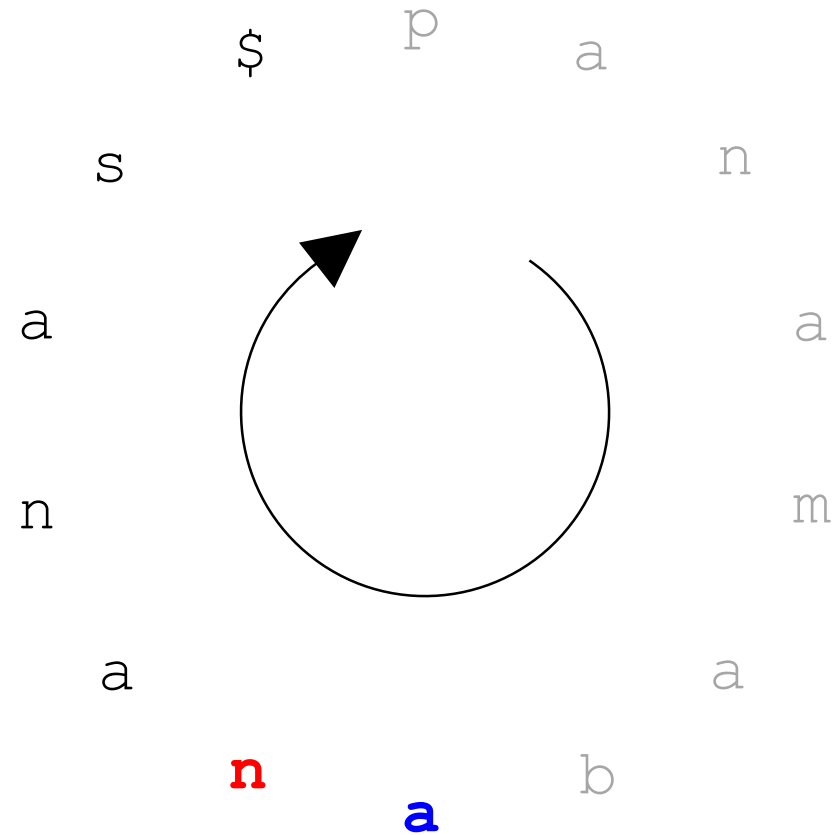
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaba**n**₂
a₆s\$panamabanan₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

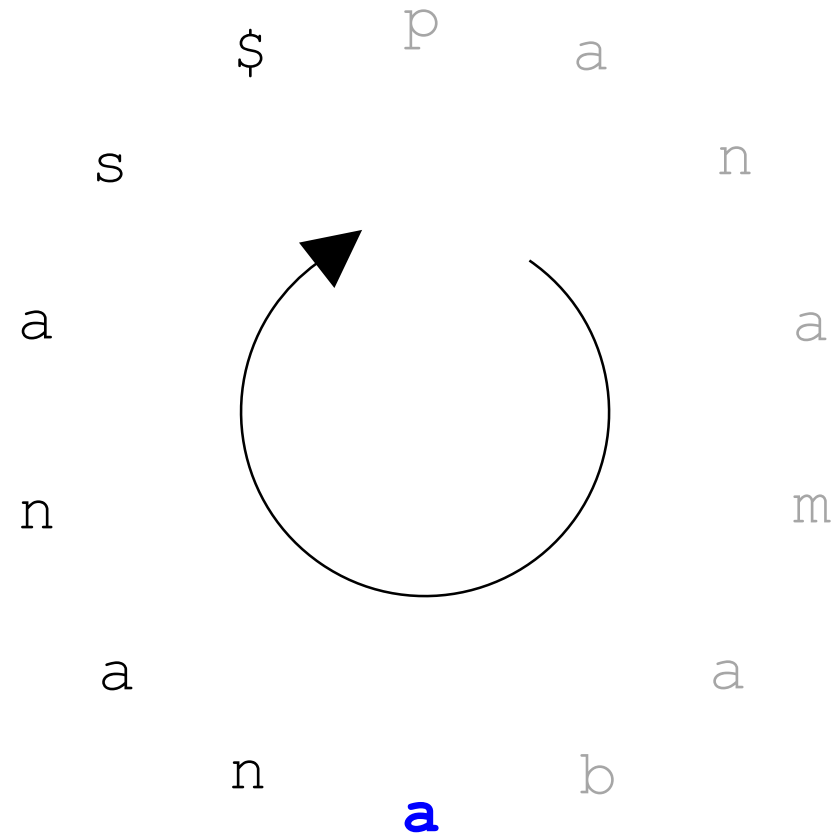


Inverting BWT Again

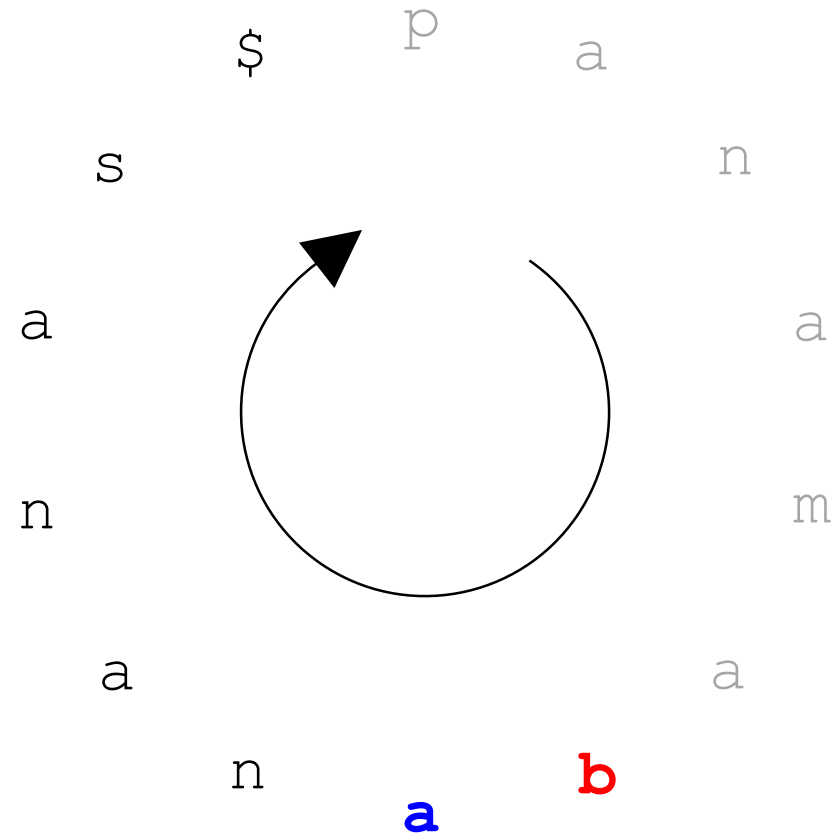
\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamab**a**₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



$\$$ ₁ p a n a m a b a n a n a s s ₁
a₁ b a n a n a s $\$$ p a n a m m ₁
a₂ m a b a n a n a s $\$$ p a n n ₁
a₃ n a m a b a n a n a s $\$$ p p ₁
a₄ n a n a s $\$$ p a n a m a b b ₁
a₅ n a s $\$$ p a n a m a b a n n ₂
a₆ s $\$$ p a n a m a b a n a n n ₃
b₁ a n a n a s $\$$ p a n a m a a ₁
m₁ a b a n a n a s $\$$ p a n a a ₂
n₁ a m a b a n a n a s $\$$ p a a ₃
n₂ a n a s $\$$ p a n a m a b **a**₄
n₃ a s $\$$ p a n a m a b a n a a ₅
p₁ a n a m a b a n a n a s $\$$ s ₁
s₁ $\$$ p a n a m a b a n a n a a ₆

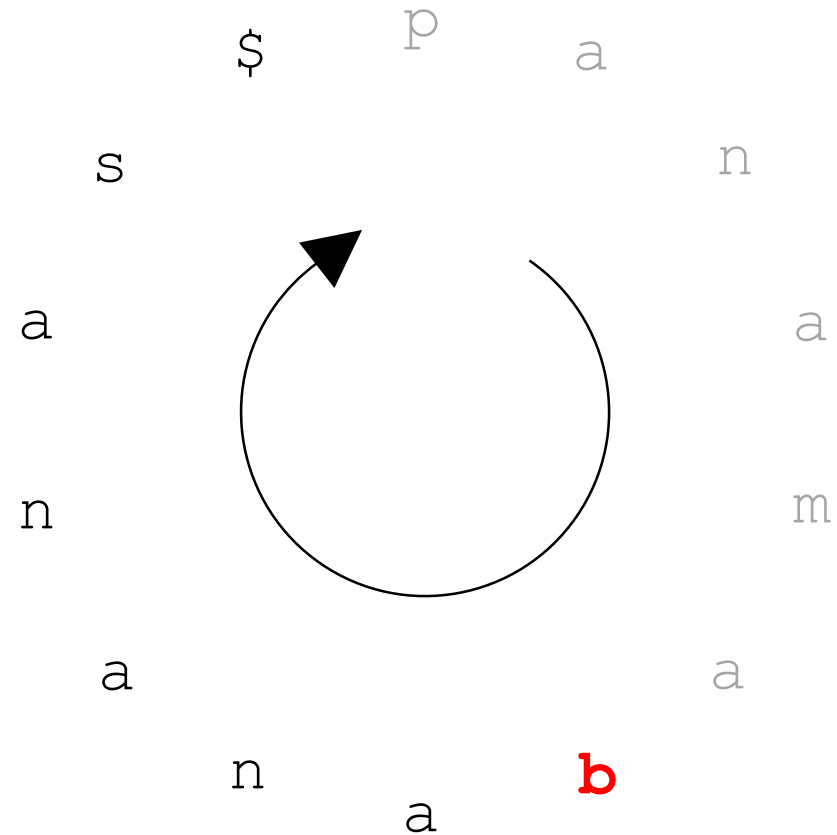


$\$$ ₁ p a n a m a b a n a n a s s ₁
a₁ b a n a n a s $\$$ p a n a m m ₁
a₂ m a b a n a n a s $\$$ p a n n ₁
a₃ n a m a b a n a n a s $\$$ p p ₁
a₄ n a n a s $\$$ p a n a m a **b**₁
a₅ n a s $\$$ p a n a m a b a n n ₂
a₆ s $\$$ p a n a m a b a n a n n ₃
b₁ a n a n a s $\$$ p a n a m a a ₁
m₁ a b a n a n a s $\$$ p a n a a ₂
n₁ a m a b a n a n a s $\$$ p a a ₃
n₂ a n a s $\$$ p a n a m a b a a ₄
n₃ a s $\$$ p a n a m a b a n a a ₅
p₁ a n a m a b a n a n a s $\$$ s ₁
s₁ $\$$ p a n a m a b a n a n a a ₆



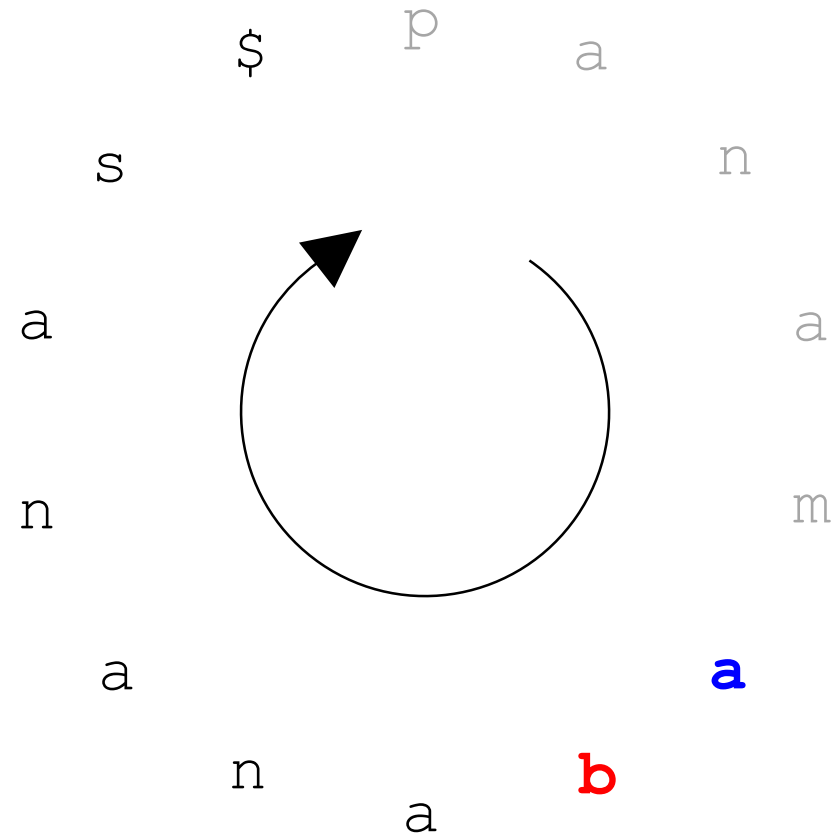
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panama**b**₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



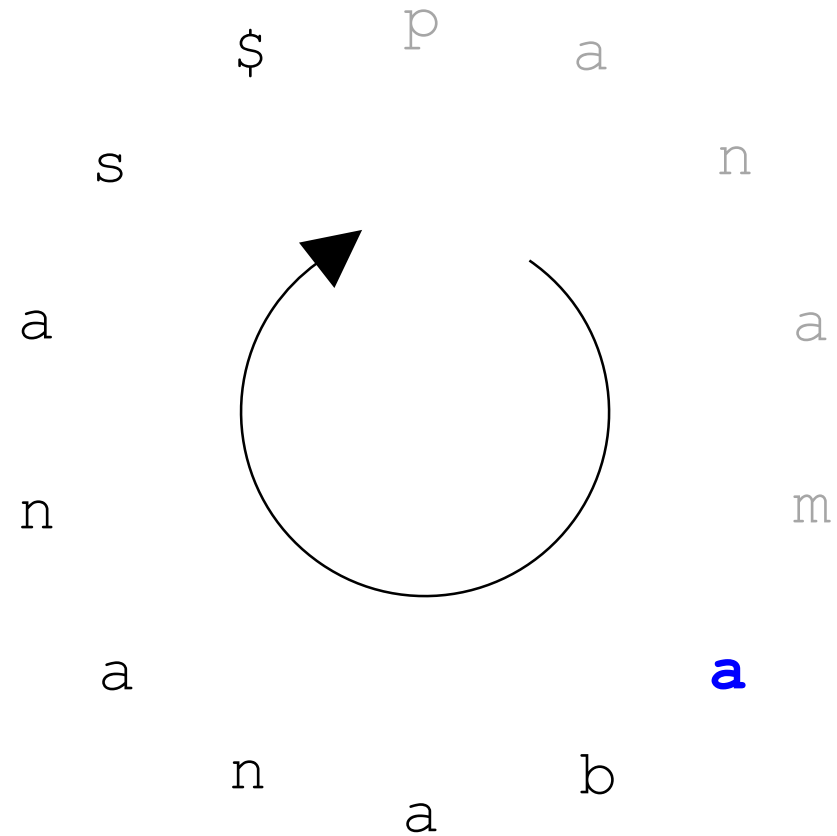
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panam**a**₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



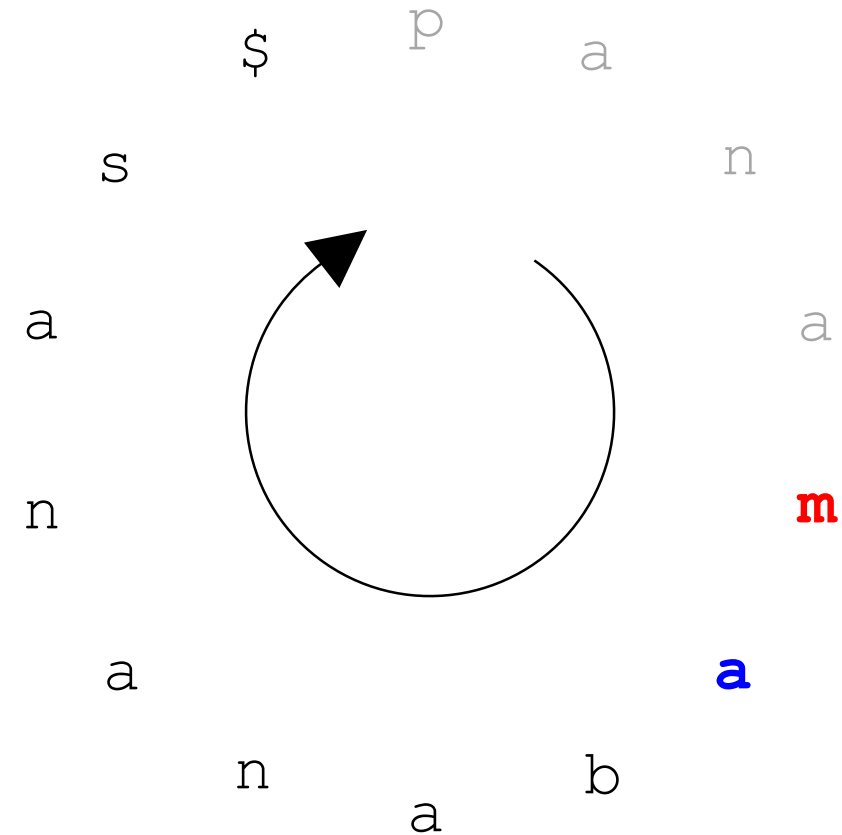
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panam**a**₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



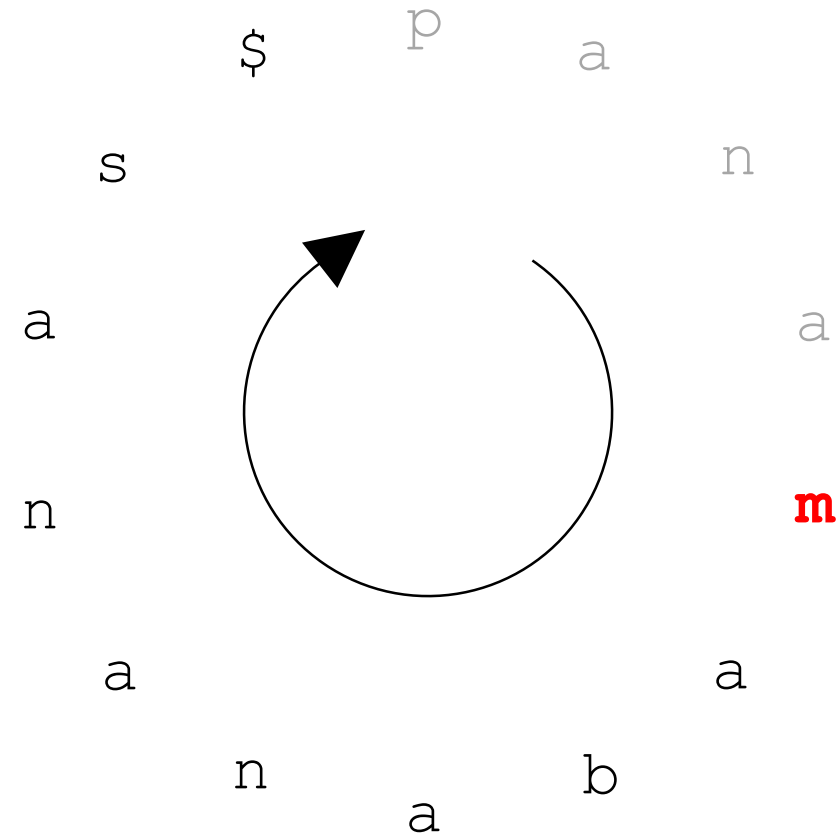
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$pana**m**₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



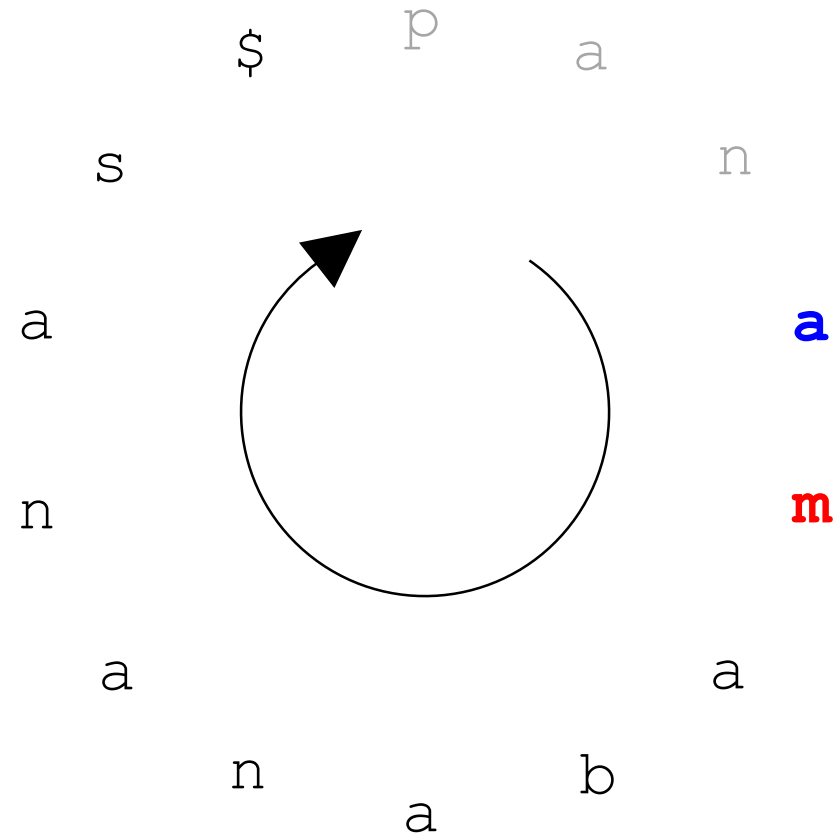
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$pana**m**₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



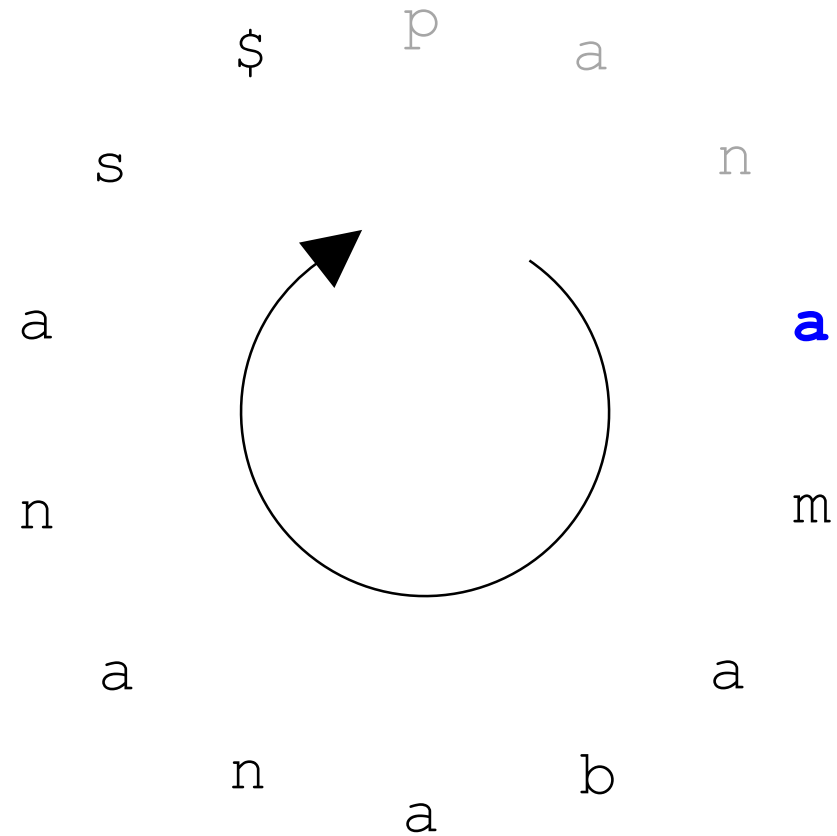
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pan**a**₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



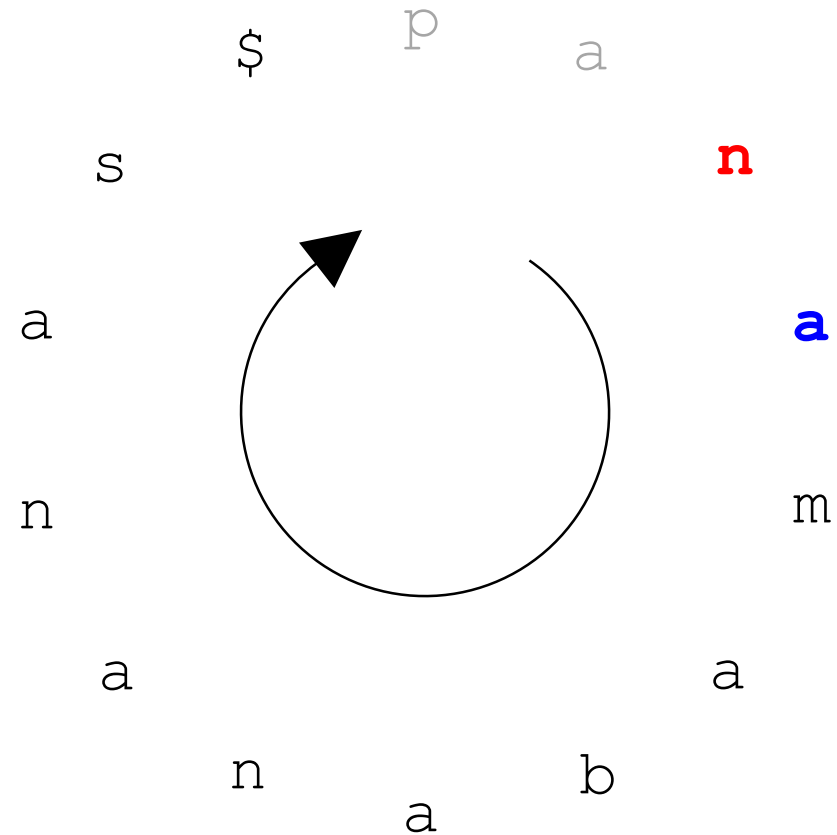
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamabanan₃
b₁ananas\$panama₁
m₁abananas\$pan**a**₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



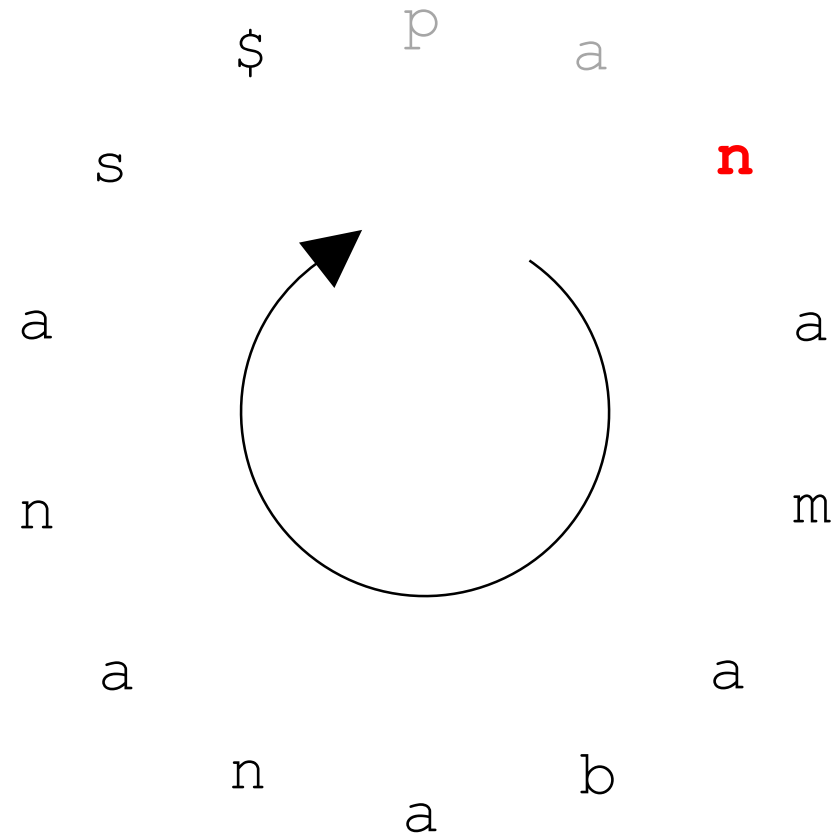
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pa**n**₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



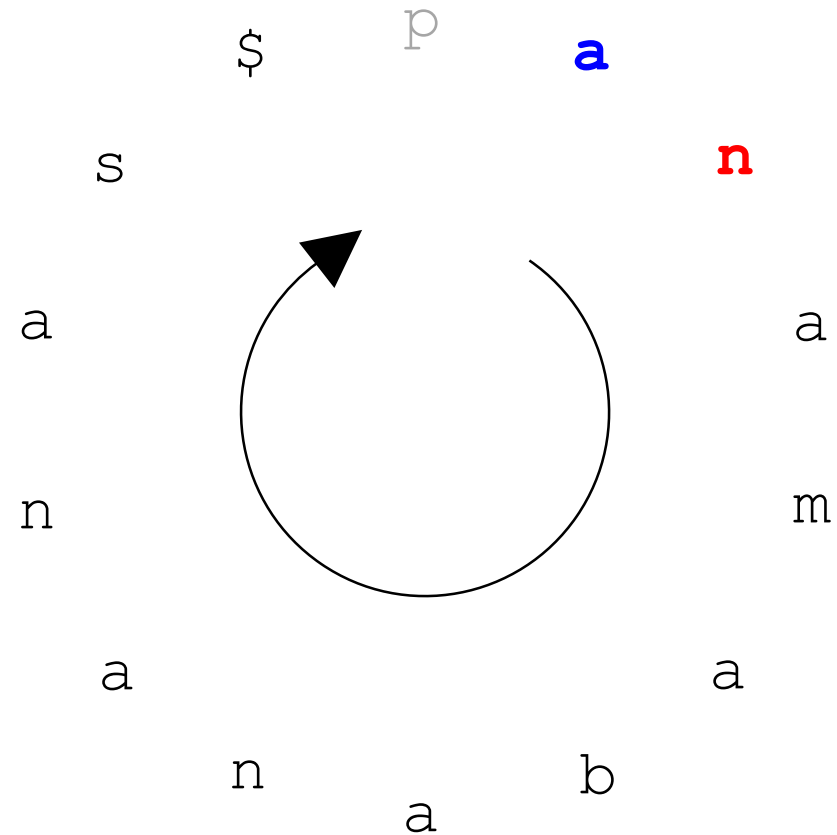
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan**n**₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



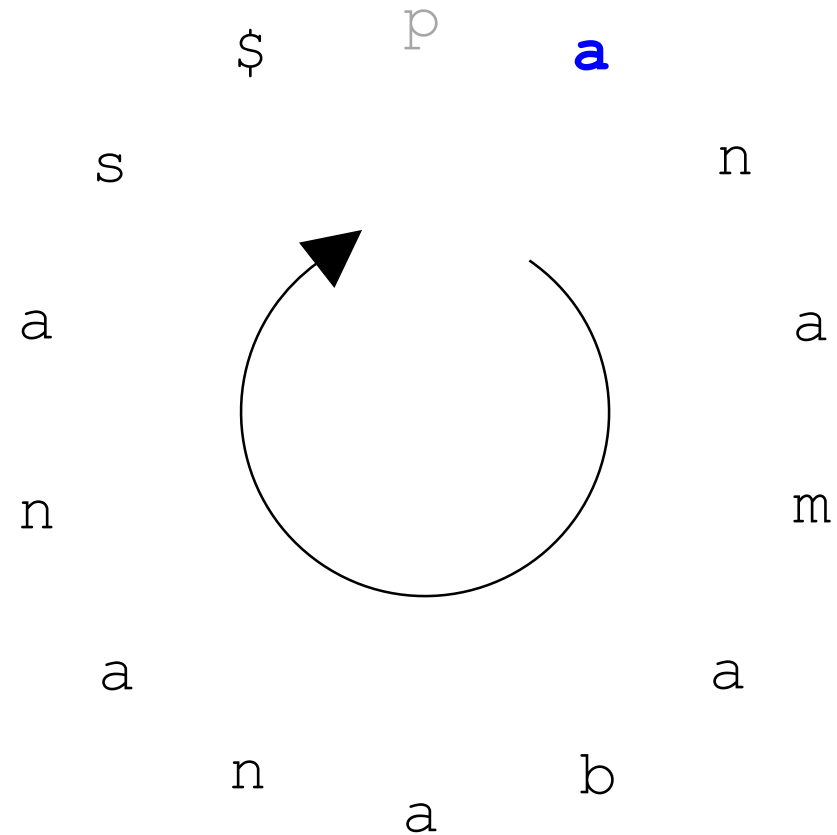
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$p**a**₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



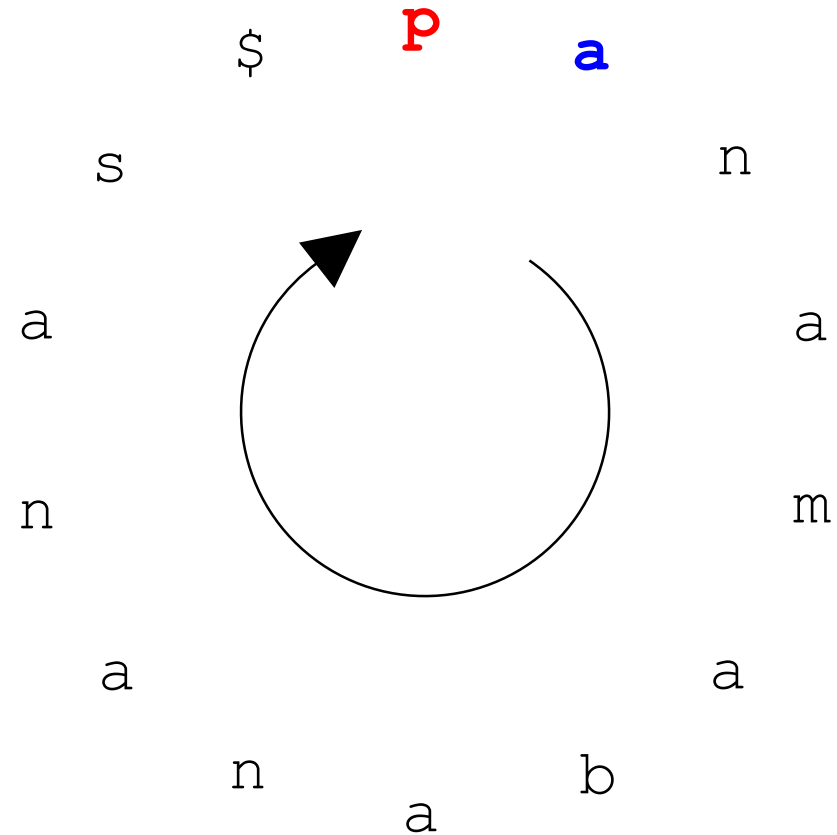
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamabanan₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$p₁**a**₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



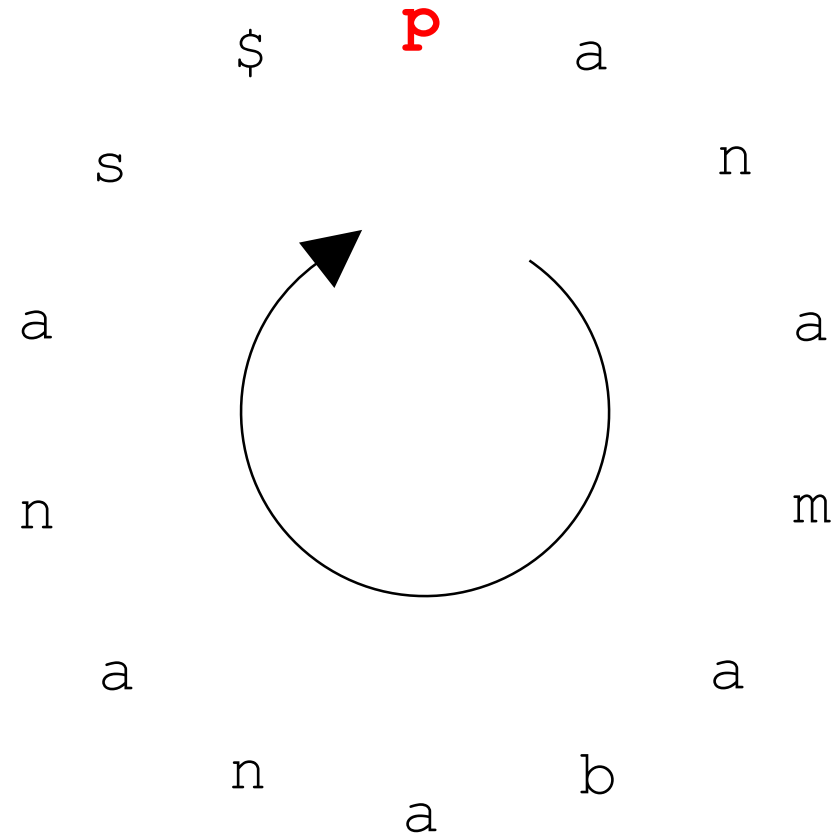
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$**p**₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆



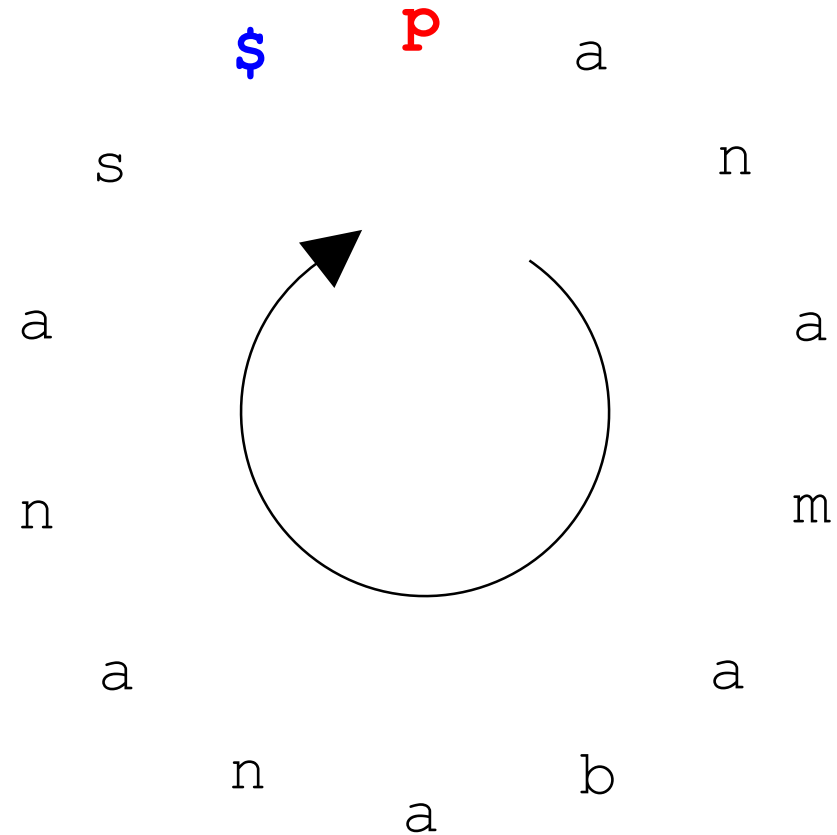
Inverting BWT Again

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$**p₁**
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

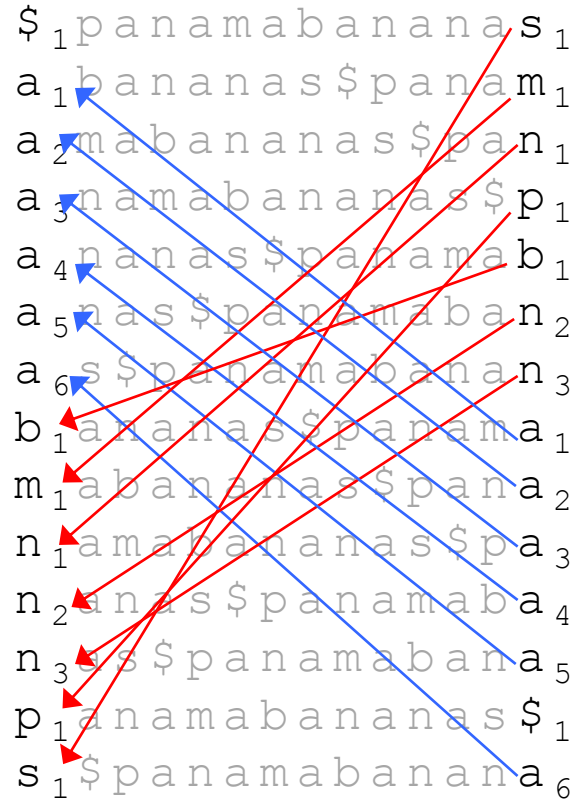


We Are Done!

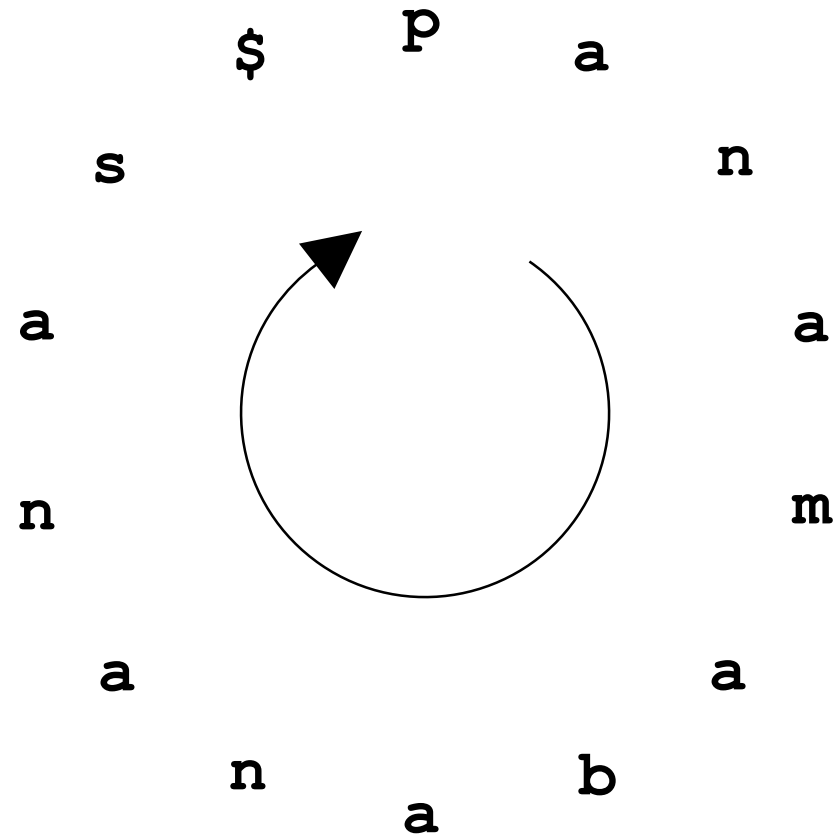
\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas**\$**₁
s₁\$panamabanana₆



This Was Fast!



- Memory: $2|Text|$
- Time: $O(|Text|)$



Outline

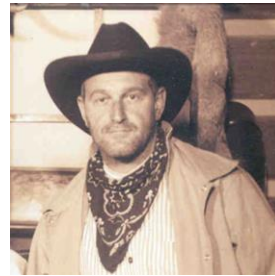
- Burrows-Wheeler Transform
- Inverting Burrows-Wheeler Transform
- **Using BWT for Pattern Matching**
- Suffix Arrays
- Approximate Pattern Matching

Back to Pattern Matching

- Suffix Tree Pattern Matching:
 - Runtime: $O(|Text| + |Patterns|)$
 - Memory: $20 \cdot |Text|$

For human genome:

- $|Text| \approx 3 \cdot 10^9$
- Can we use $BWT(Text)$ to design a more memory efficient linear-time algorithm for Multiple Pattern Matching?



Finding Pattern Matches Using BWT

- Searching for **ana** in **p****ana**mab**anana**s

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamaban3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanana6
```

Lets Start by Matching the Last Symbol (**a**)

- Searching for an **a** in panamabananas

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

Matching the Last Two Symbols (**na**)

- Searching for a**na** in panamabananas

\$₁panamabananas₁
a₁bananas\$pana**m**₁
a₂mabananas\$pa**n**₁
a₃namabananas\$**p**₁
a₄nanas\$panama**b**₁
a₅nas\$panamaba**n**₂
a₆s\$panamabana**n**₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabana₆

Three Matches of **na** Found!

- Searching for a**na** in panamabananas

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pa**n**₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaba**n**₂
a₆s\$panamabana**n**₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabana₆

Three matches of **na** are found in the string **panamabananas**, indicated by green arrows pointing to the **a** and **n** characters in the substrings **a₂mabananas\$pa**n**₁**, **a₅nas\$panamaba**n**₂**, and **a₆s\$panamabana**n**₃**.

Three Matches of **na** Found!

- Searching for a**na** in panamabananas

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pa**n**₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaba**n**₂
a₆s\$panamabana**n**₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabana₆

Three Matches of **na** Found!

- Searching for a**na** in panamabananas

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

Matching **ana**

- Searching for **ana** in panamabananas

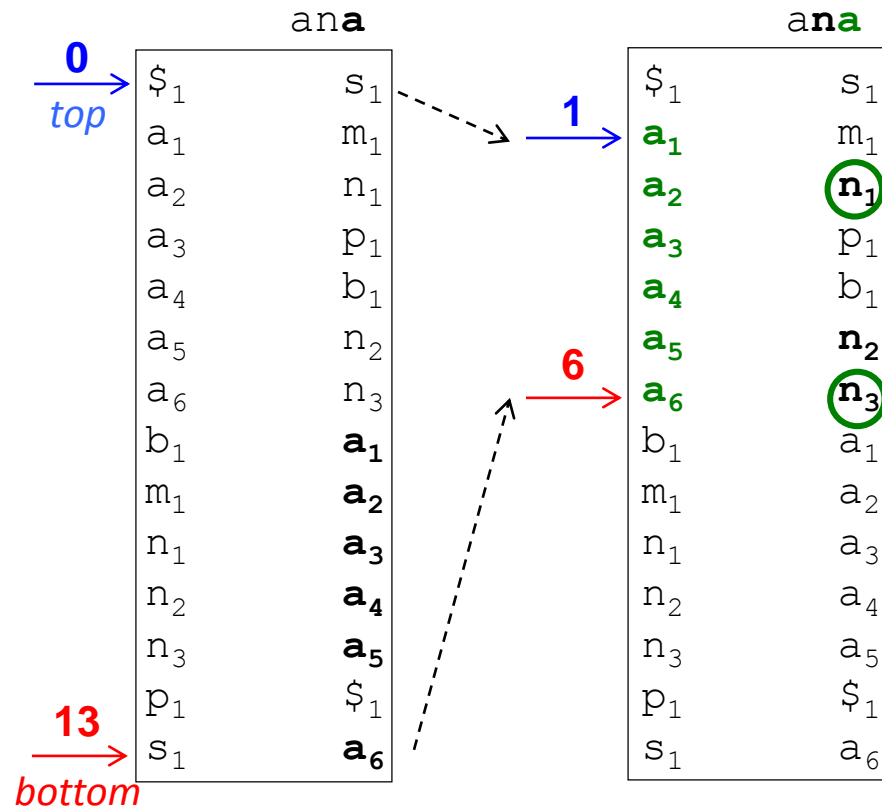
\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$p₁**a₃**
n₂anas\$panamab₁**a₄**
n₃as\$panamaban₂**a₅**
p₁anamabananas\$₁
s₁\$panamabanana₆

Three Matches of **ana** Found!

- Searching for **ana** in panamabananas

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

Searching for **ana** using *top* and *bottom* pointers



topIndex \leftarrow first position of symbol among positions from *top* to *bottom* in *LastColumn*

bottomIndex \leftarrow last position of symbol among positions from *top* to *bottom* in *LastColumn*

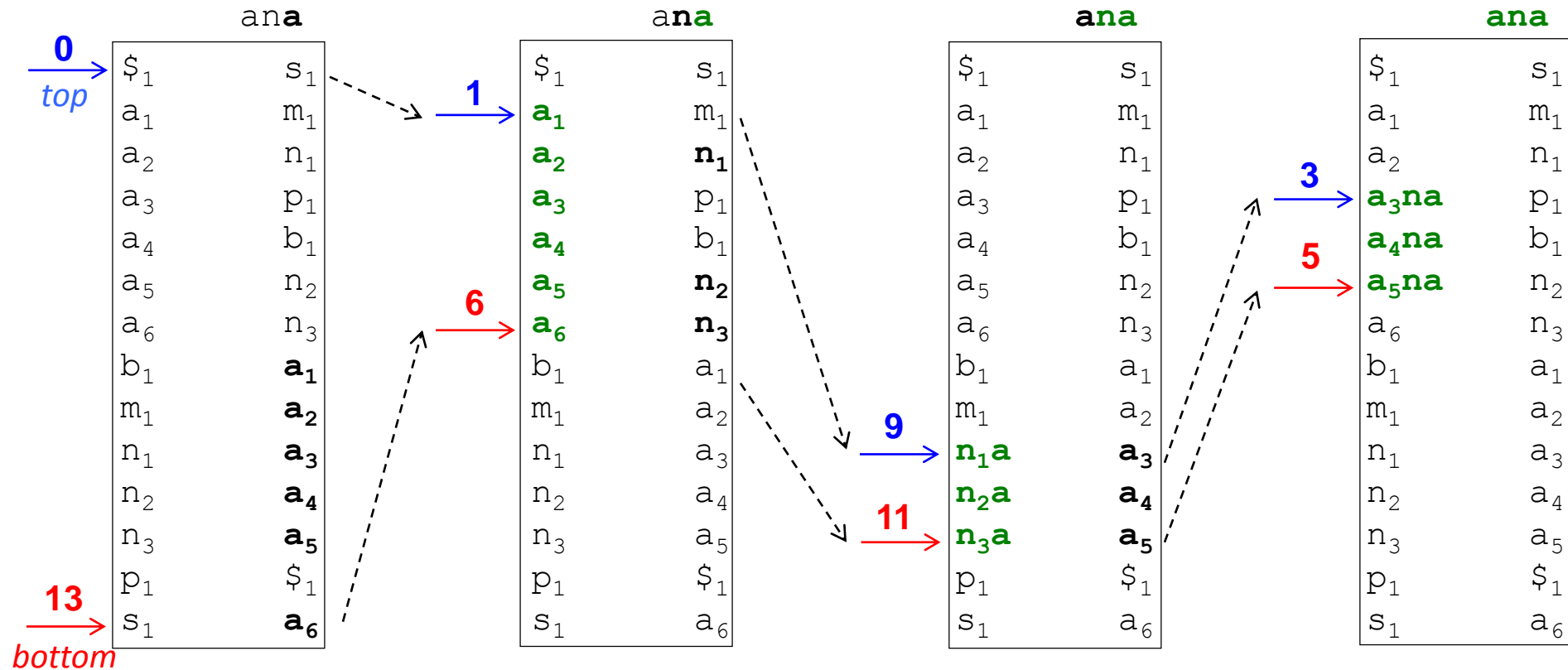
BWMatching

```
BWMATCHING(FirstColumn, LastColumn, Pattern, LASTTOFIRST)  
  top  $\leftarrow$  0  
  bottom  $\leftarrow$  |LastColumn| - 1  
  while top  $\leq$  bottom  
    if Pattern is nonempty  
      symbol  $\leftarrow$  last letter in Pattern  
      remove last letter from Pattern  
      if positions from top to bottom in LastColumn contain symbol  
        topIndex  $\leftarrow$  first position of symbol among positions from top to bottom  
          in LastColumn  
        bottomIndex  $\leftarrow$  last position of symbol among positions from top to  
          bottom in LastColumn  
        top  $\leftarrow$  LASTTOFIRST(topIndex)  
        bottom  $\leftarrow$  LASTTOFIRST(bottomIndex)  
      else  
        return 0  
    else  
      return bottom - top + 1
```

Given a symbol at position *index* in *LastColumn*,
LastToFirst(*index*) defines the position of this symbol in *FirstColumn*

BWMatching is slow:

it analyzes every symbol from *top* to *bottom* in each step!



if positions from *top* to *bottom* in *LastColumn* contain symbol
topIndex ← first position of symbol among positions from *top* to *bottom*
in *LastColumn*
bottomIndex ← last position of symbol among positions from *top* to
bottom in *LastColumn*

Introducing *Count* Array

<i>i</i>	<i>FirstColumn</i>	<i>LastColumn</i>	LASTTOFIRST(<i>i</i>)	COUNT						
				\$	a	b	m	n	p	s
0	\$ ₁	s ₁	13	0	0	0	0	0	0	0
1	a ₁	m ₁	8	0	0	0	0	0	0	1
2	a ₂	n ₁	9	0	0	0	1	0	0	1
3	a ₃	p ₁	12	0	0	0	1	1	0	1
4	a ₄	b ₁	7	0	0	0	1	1	1	1
5	a ₅	n ₂	10	0	0	1	1	1	1	1
6	a ₆	n ₃	11	0	0	1	1	2	1	1
7	b ₁	a ₁	1	0	0	1	1	3	1	1
8	m ₁	a ₂	2	0	1	1	1	3	1	1
9	n ₁	a ₃	3	0	2	1	1	3	1	1
10	n ₂	a ₄	4	0	3	1	1	3	1	1
11	n ₃	a ₅	5	0	4	1	1	3	1	1
12	p ₁	\$ ₁	0	0	5	1	1	3	1	1
13	s ₁	a ₆	6	1	5	1	1	3	1	1
				1	6	1	1	3	1	1

*Count*_{*symbol*}(*i*, *LastColumn*):

#occurrences of *symbol* in the first *i* positions of *LastColumn*

BetterBWMatching

```
BETTERBWMATCHING(FIRSTOCCURRENCE, LastColumn, Pattern, COUNT)
  top  $\leftarrow$  0
  bottom  $\leftarrow$  |LastColumn| - 1
  while top  $\leq$  bottom
    if Pattern is nonempty
      symbol  $\leftarrow$  last letter in Pattern
      remove last letter from Pattern
      top  $\leftarrow$  FIRSTOCCURRENCE(symbol) + COUNTsymbol(top, LastColumn)
      bottom  $\leftarrow$  FIRSTOCCURRENCE(symbol) + COUNTsymbol(bottom + 1,
        LastColumn) - 1
    else
      return bottom - top + 1
  return
```



Where Are the Matches?

- We know that **ana** occurs 3 times, but where does **ana** appear in *Text*???

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

Outline

- Burrows-Wheeler Transform
- Inverting Burrows-Wheeler Transform
- Using BWT for Pattern Matching
- **Suffix Arrays**
- Approximate Pattern Matching

Where Are the Matches?

- **Suffix array** holds starting position of each suffix

```
$1panamabananas1
a1bananas$panam1
a2mabananas$pan1
a3namabananas$p1
a4nanas$panamab1
a5nas$panamaban2
a6s$panamaban3
b1ananas$panama1
m1abananas$pana2
n1amabananas$pa3
n2anas$panamaba4
n3as$panamabana5
p1anamabananas$1
s1$panamabanana6
```


Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabananas\$

1 3	\$ ₁ panamabananas ₁
	a ₁ bananas\$panam ₁
	a ₂ mabananas\$pan ₁
	a ₃ namabananas\$p ₁
	a ₄ nanas\$panamab ₁
	a ₅ nas\$panamaban ₂
	a ₆ s\$panamabanan ₃
	b ₁ ananas\$panama ₁
	m ₁ abananas\$pana ₂
	n ₁ amabananas\$pa ₃
	n ₂ anas\$panamaba ₄
	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panam**ab**ananas\$

1	3	\$ ₁	panamab	ananas	s ₁
	5	a ₁	bananas	\$	panam ₁
		a ₂	mab	ananas	\$pan ₁
		a ₃	namab	ananas	\$p ₁
		a ₄	nanas	\$panamab	₁
		a ₅	nas	\$panamaban	₂
		a ₆	s	\$panamabanan	₃
		b ₁	ananas	\$panama	₁
		m ₁	ab	ananas	\$pana ₂
		n ₁	amab	ananas	\$pa ₃
		n ₂	anas	\$panamaba	₄
		n ₃	as	\$panamabana	₅
		p ₁	anamab	ananas	\$ ₁
		s ₁	\$panamab	anana	₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

pan**amabananas**\$

1	3	\$ ₁ panamabananas ₁
5		a ₁ bananas \$panam ₁
3		a ₂ mabananas \$pan ₁
		a ₃ namabananas\$pa ₁
		a ₄ nanas\$panamab ₁
		a ₅ nas\$panamaban ₂
		a ₆ s\$panamabanan ₃
		b ₁ ananas\$panama ₁
		m ₁ abananas\$pana ₂
		n ₁ amabananas\$pa ₃
		n ₂ anas\$panamaba ₄
		n ₃ as\$panamabana ₅
		p ₁ anamabananas\$ ₁
		s ₁ \$panamabannana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

`p`**anamabananas**`$`

1 3	<code>\$</code> ₁ <code>p</code> <code>a</code> <code>n</code> <code>a</code> <code>m</code> <code>a</code> <code>b</code> <code>a</code> <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> ₁
5	<code>a</code> ₁ <code>b</code> <code>a</code> <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> <code>a</code> <code>n</code> <code>a</code> <code>m</code> ₁
3	<code>a</code> ₂ <code>m</code> <code>a</code> <code>b</code> <code>a</code> <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> <code>a</code> <code>n</code> ₁
1	<code>a</code> ₃ <code>n</code> <code>a</code> <code>m</code> <code>a</code> <code>b</code> <code>a</code> <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> ₁
	<code>a</code> ₄ <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> <code>a</code> <code>n</code> <code>a</code> <code>m</code> <code>a</code> ₁
	<code>a</code> ₅ <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> <code>a</code> <code>n</code> <code>a</code> <code>m</code> <code>a</code> <code>b</code> ₂
	<code>a</code> ₆ <code>s</code> <code>\$</code> <code>p</code> <code>a</code> <code>n</code> <code>a</code> <code>m</code> <code>a</code> <code>b</code> <code>a</code> <code>n</code> ₃
	<code>b</code> ₁ <code>a</code> <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> <code>a</code> <code>n</code> <code>a</code> ₁
	<code>m</code> ₁ <code>a</code> <code>b</code> <code>a</code> <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> <code>a</code> ₂
	<code>n</code> ₁ <code>a</code> <code>m</code> <code>a</code> <code>b</code> <code>a</code> <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> ₃
	<code>n</code> ₂ <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> <code>p</code> <code>a</code> <code>n</code> <code>a</code> <code>m</code> <code>a</code> ₄
	<code>n</code> ₃ <code>a</code> <code>s</code> <code>\$</code> <code>p</code> <code>a</code> <code>n</code> <code>a</code> <code>m</code> <code>a</code> <code>b</code> ₅
	<code>p</code> ₁ <code>a</code> <code>n</code> <code>a</code> <code>m</code> <code>a</code> <code>b</code> <code>a</code> <code>n</code> <code>a</code> <code>n</code> <code>a</code> <code>s</code> <code>\$</code> ₁
	<code>s</code> ₁ <code>\$</code> <code>p</code> <code>a</code> <code>n</code> <code>a</code> <code>m</code> <code>a</code> <code>b</code> <code>a</code> <code>n</code> <code>a</code> ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamab**ananas**\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
	a ₅ nas\$panamaban ₂
	a ₆ s\$panamabanan ₃
	b ₁ ananas\$panama ₁
	m ₁ abananas\$pana ₂
	n ₁ amabananas\$pa ₃
	n ₂ anas\$panamaba ₄
	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamaban**anas**\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
	a ₆ s\$panamabanan ₃
	b ₁ ananas\$panama ₁
	m ₁ abananas\$pana ₂
	n ₁ amabananas\$pa ₃
	n ₂ anas\$panamaba ₄
	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$_ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabananas\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
1 1	a ₆ s\$panamaban ₃
	b ₁ ananas\$panama ₁
	m ₁ abananas\$pana ₂
	n ₁ amabananas\$pa ₃
	n ₂ anas\$panamaba ₄
	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$_ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panama**bananas**\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
1 1	a ₆ s\$panamaban ₃
6	b ₁ ananas\$panama ₁
	m ₁ abananas\$pana ₂
	n ₁ amabananas\$pa ₃
	n ₂ anas\$panamaba ₄
	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$_ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panam**abananas**\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
1 1	a ₆ s\$panamaban ₃
6	b ₁ ananas\$panama ₁
4	m ₁ abananas\$pana ₂
	n ₁ amabananas\$pa ₃
	n ₂ anas\$panamaba ₄
	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$_ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabananas\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
1 1	a ₆ s\$panamaban ₃
6	b ₁ ananas\$panama ₁
4	m ₁ abananas\$pana ₂
2	n ₁ amabananas\$pa ₃
	n ₂ anas\$panamaba ₄
	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamab**ananas**\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
1 1	a ₆ s\$panamabanan ₃
6	b ₁ ananas\$panama ₁
4	m ₁ abananas\$pana ₂
2	n ₁ amabananas\$pa ₃
8	n ₂ anas\$panamaba ₄
	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabanan**as**\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
1 1	a ₆ s\$panamabanan ₃
6	b ₁ ananas\$panama ₁
4	m ₁ abananas\$pana ₂
2	n ₁ amabananas\$pa ₃
8	n ₂ anas\$panamaba ₄
1 0	n ₃ as\$panamabana ₅
	p ₁ anamabananas\$ ₁
	s ₁ \$panamabannana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabananas\$

13	\$₁ panamabananas ₁
5	a₁ bananas\$panam ₁
3	a₂ mabananas\$pan ₁
1	a₃ namabananas\$p ₁
7	a₄ nanas\$panamab ₁
9	a₅ nas\$panamaban ₂
11	a₆ s\$panamaban ₃
6	b₁ ananas\$panama ₁
4	m₁ abananas\$pana ₂
2	n₁ amabananas\$pa ₃
8	n₂ anas\$panamaba ₄
10	n₃ as\$panamabana ₅
0	p₁ anamabananas\$ ₁
	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

panamabanana**s**\$

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
1 1	a ₆ s\$panamaban ₃
6	b ₁ ananas\$panama ₁
4	m ₁ abananas\$pana ₂
2	n ₁ amabananas\$pa ₃
8	n ₂ anas\$panamaba ₄
1 0	n ₃ as\$panamabana ₅
0	p ₁ anamabananas\$_ ₁
1 2	s ₁ \$panamabanana ₆

Suffix Array

- **Suffix array:** holds starting position of each suffix beginning a row.

1 3	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a ₃ namabananas\$p ₁
7	a ₄ nanas\$panamab ₁
9	a ₅ nas\$panamaban ₂
1 1	a ₆ s\$panamaban ₃
6	b ₁ ananas\$panama ₁
4	m ₁ abananas\$pana ₂
2	n ₁ amabananas\$pa ₃
8	n ₂ anas\$panamaba ₄
1 0	n ₃ as\$panamabana ₅
0	p ₁ anamabananas\$ ₁
1 2	s ₁ \$panamabanana ₆

Using the Suffix Array to Find Matches

- Thus, **ana** occurs at positions **1, 7, 9**:

- `panamabananas$`

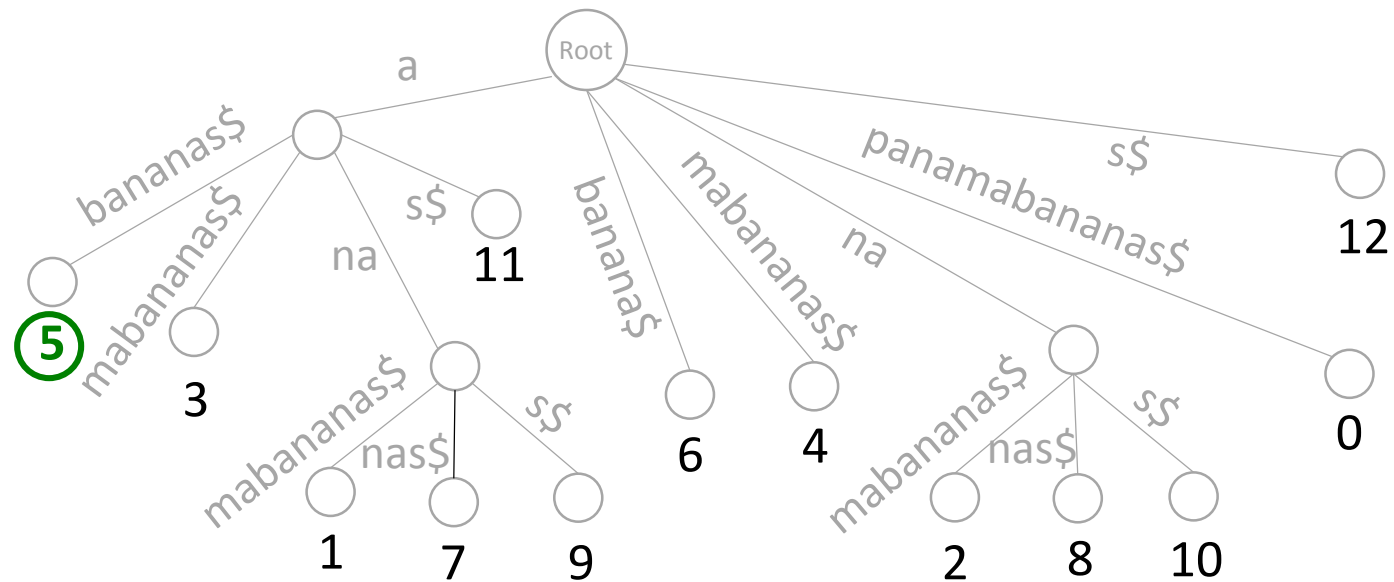


13	\$ ₁ panamabananas ₁
5	a ₁ bananas\$panam ₁
3	a ₂ mabananas\$pan ₁
1	a₃na mabananas\$p ₁
7	a₄na nas\$panamab ₁
9	a₅na s\$panamaban ₂
11	a ₆ s\$panamaban ₃
6	b ₁ ananas\$panama ₁
4	m ₁ abananas\$pana ₂
2	n ₁ amabananas\$pa ₃
8	n ₂ anas\$panamaba ₄
10	n ₃ as\$panamabana ₅
0	p ₁ anamabananas\$ ₁
12	s ₁ \$panamabanana ₆

Naïve algorithm for constructing suffix array (sorting all suffixes of *Text*)

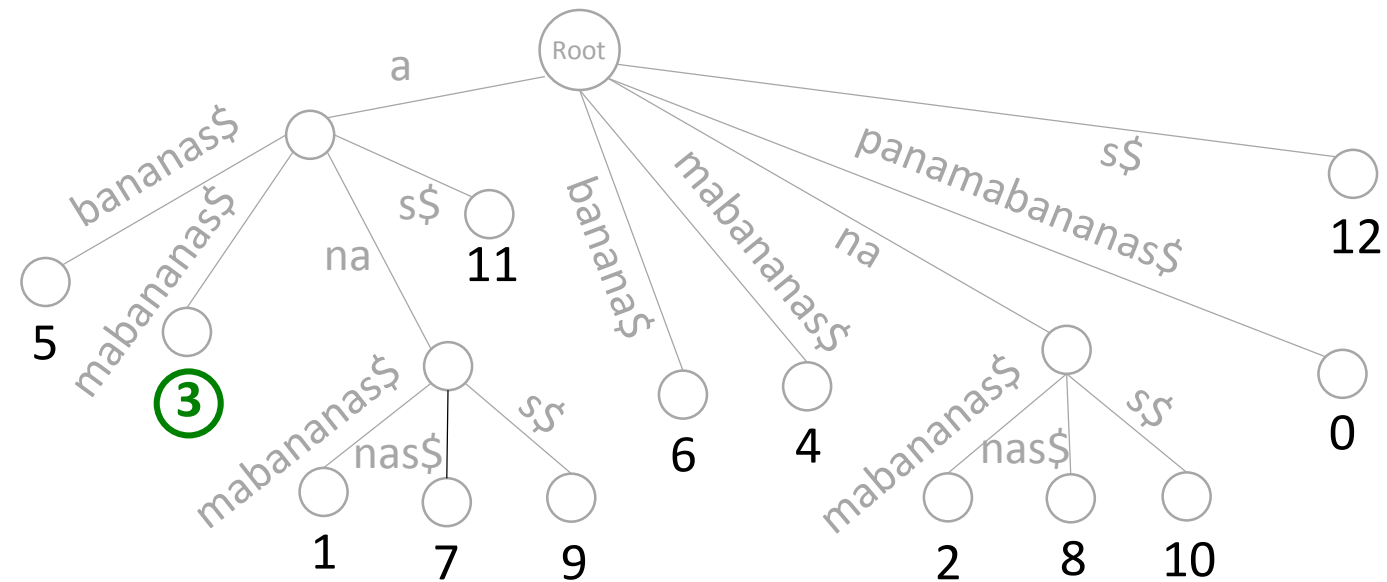
$O(|Text| \cdot \log |Text|)$ comparisons

From Suffix Tree to Suffix Array: Depth-First Traversal



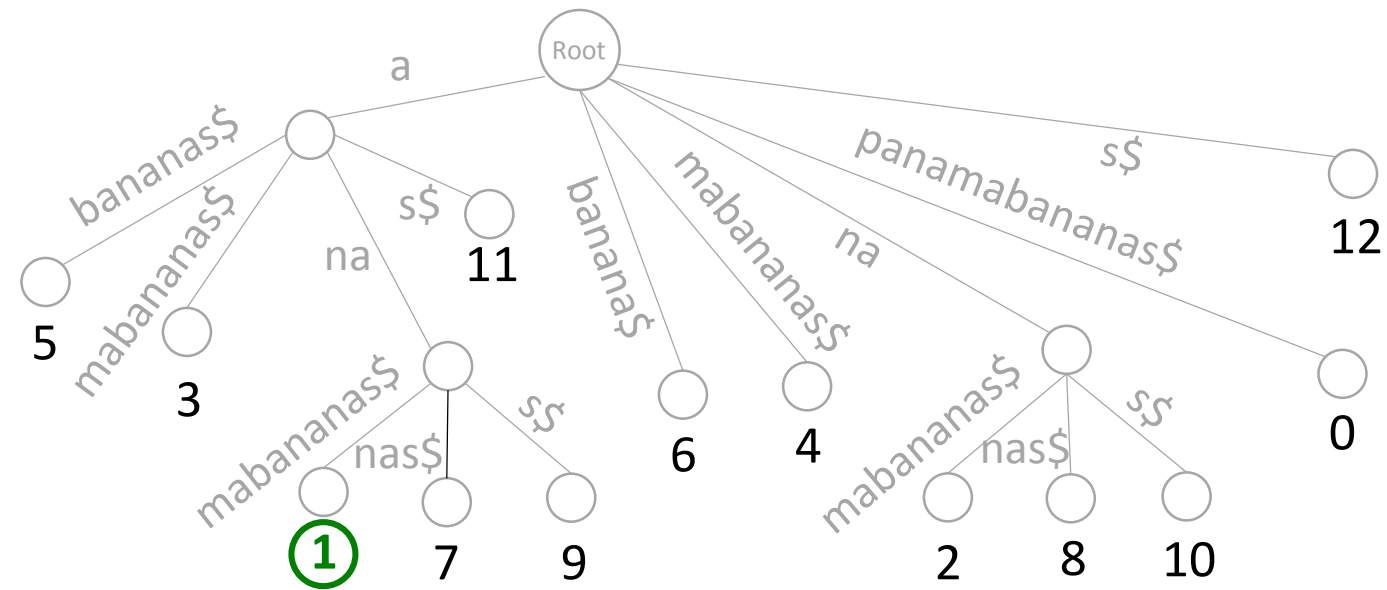
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array: Depth-First Traversal



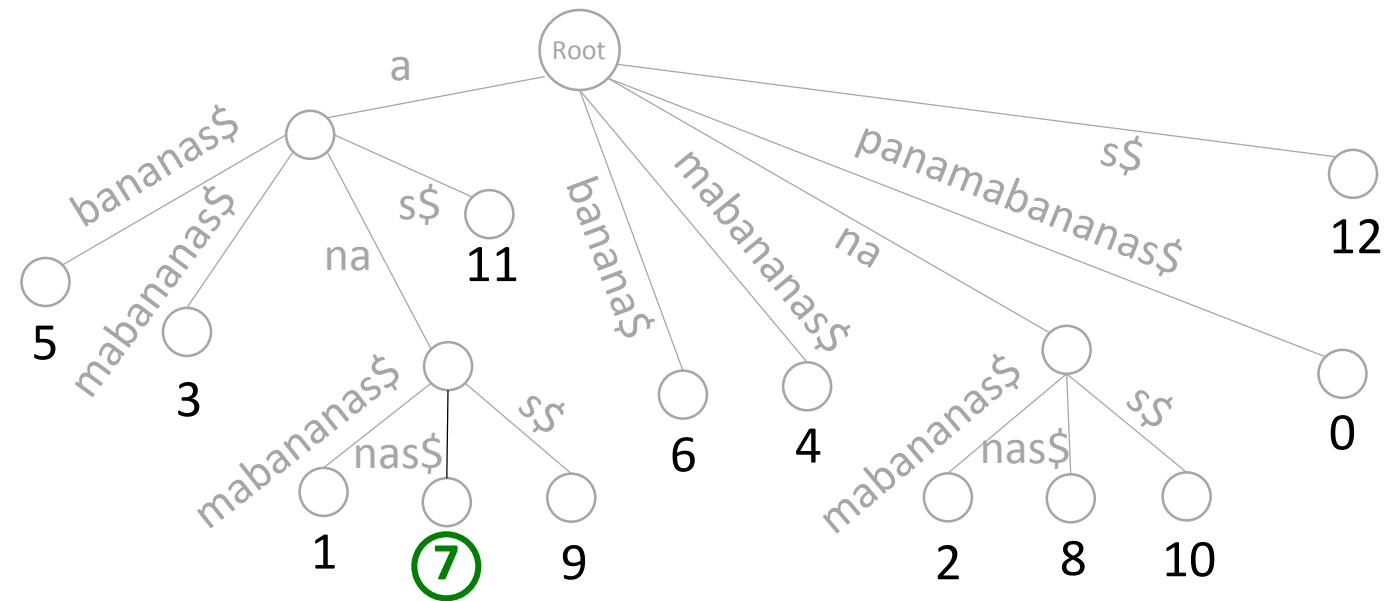
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array: Depth-First Traversal



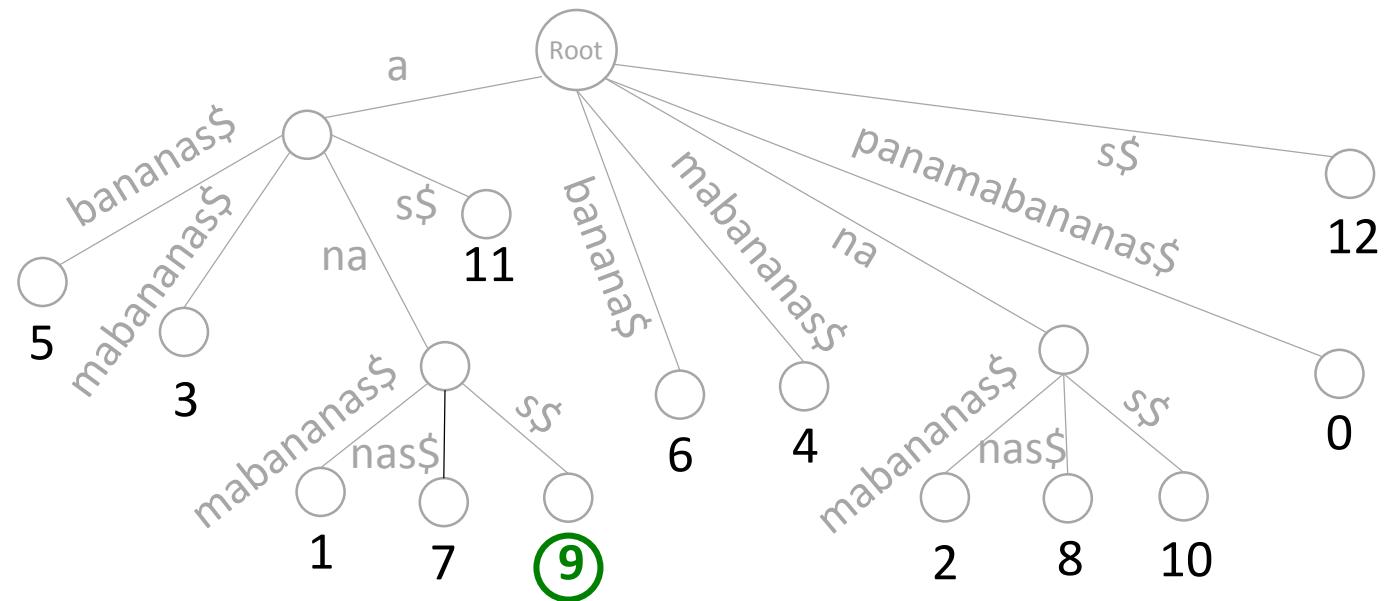
[13 5 3 **1** 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array



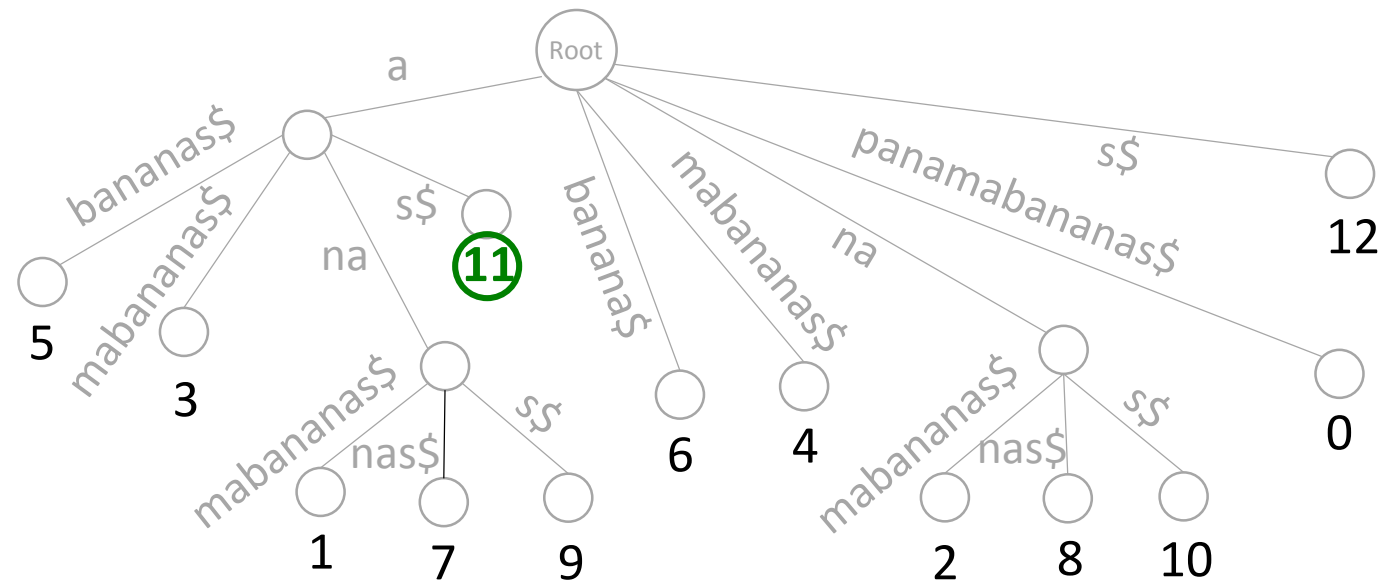
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array



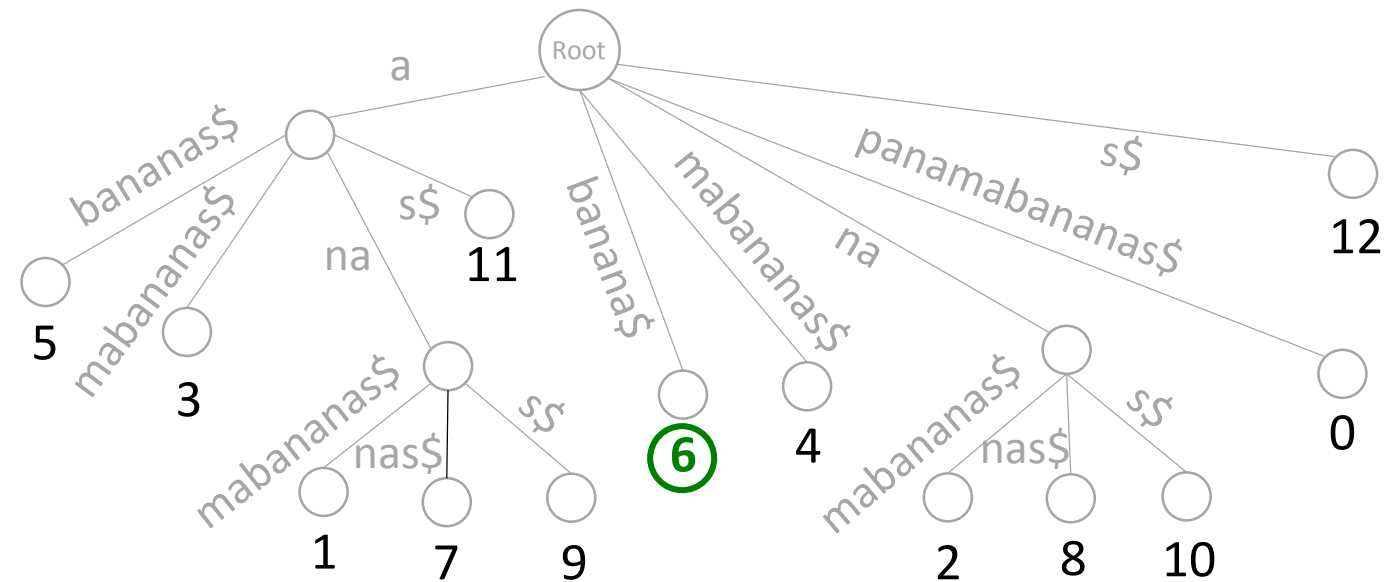
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array



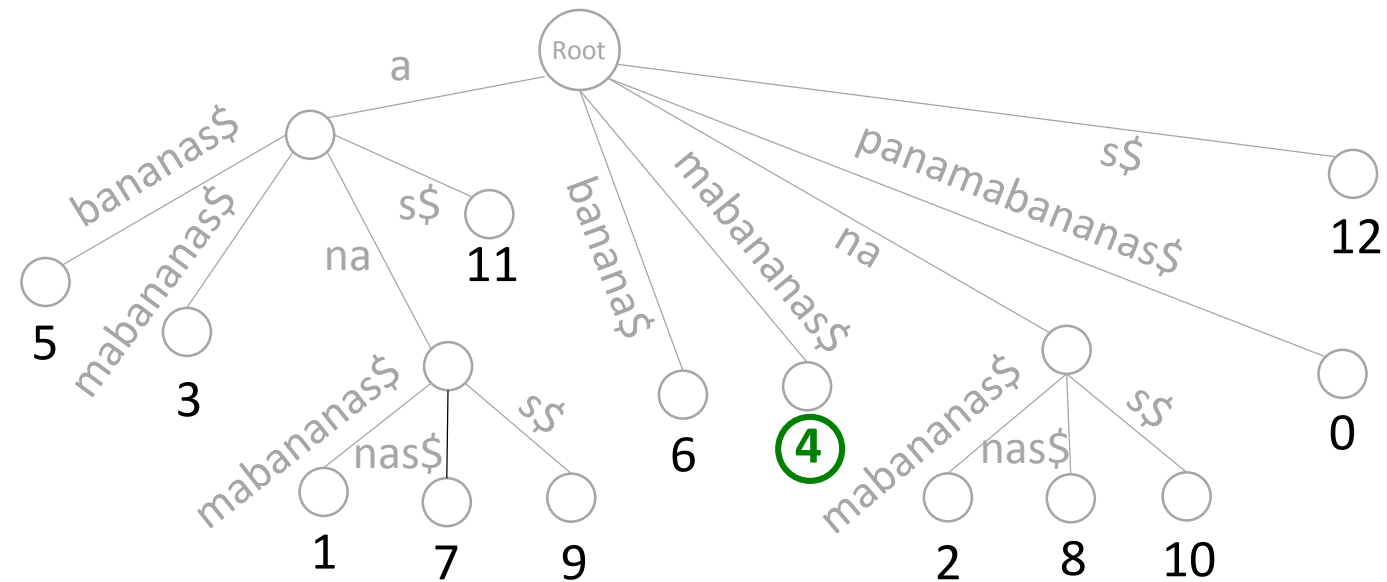
[13 5 3 1 7 9 **11** 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array



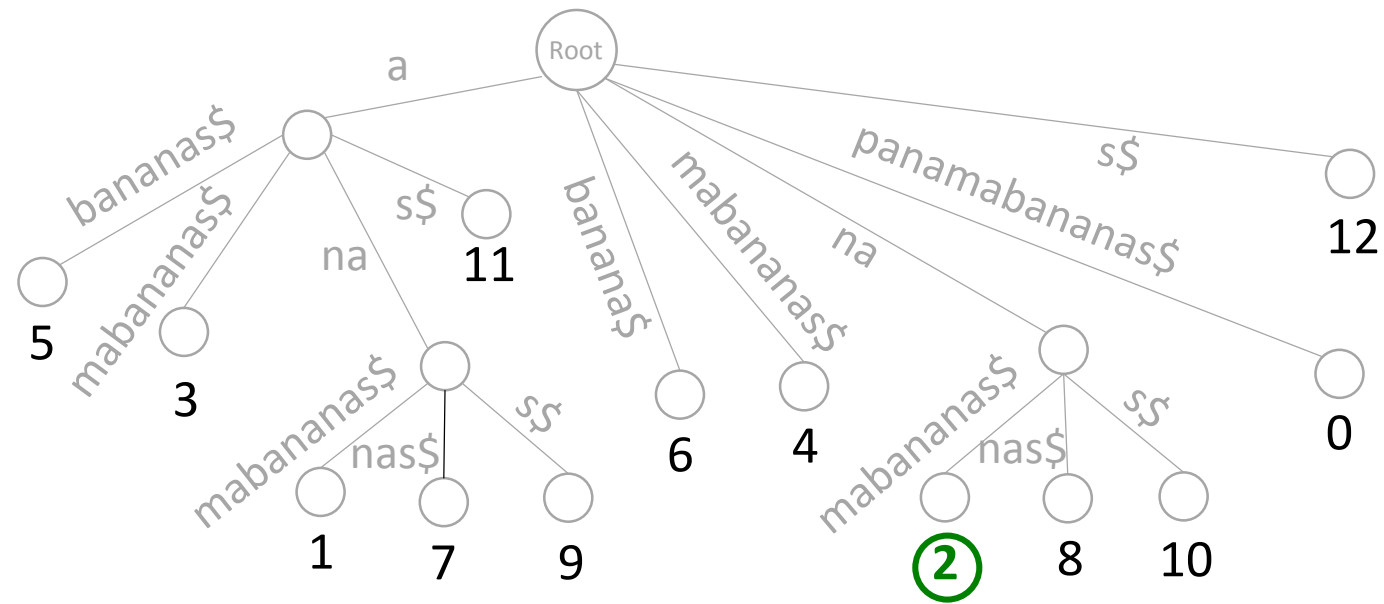
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array



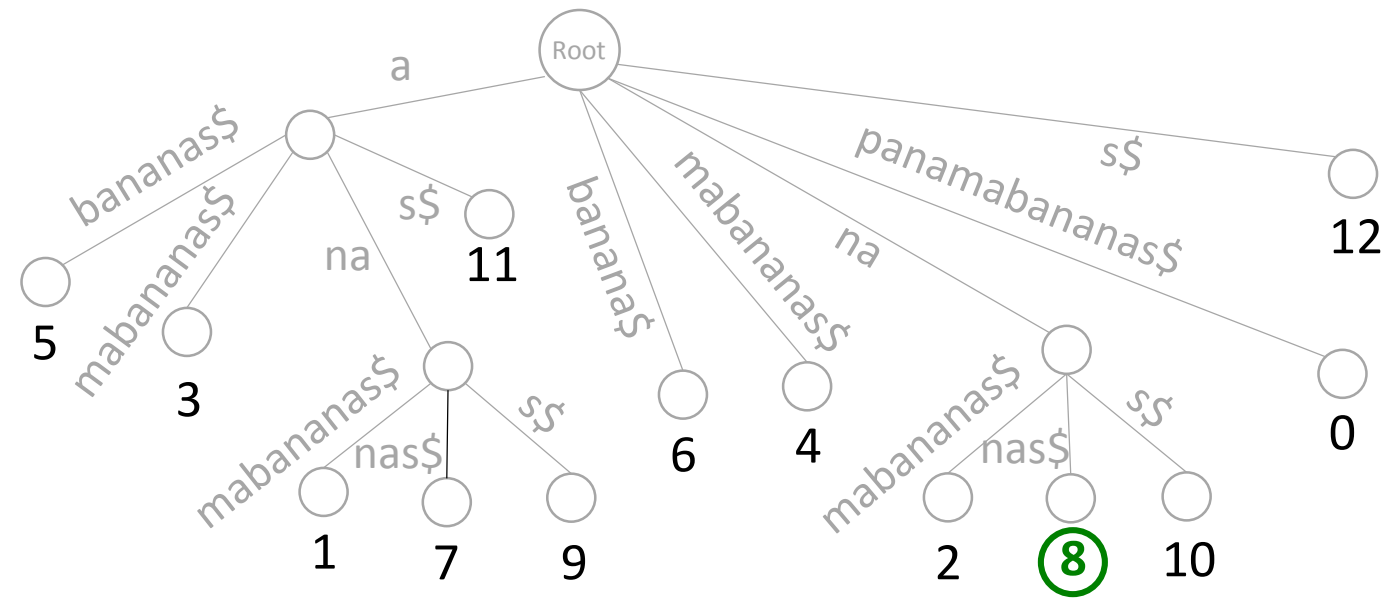
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array



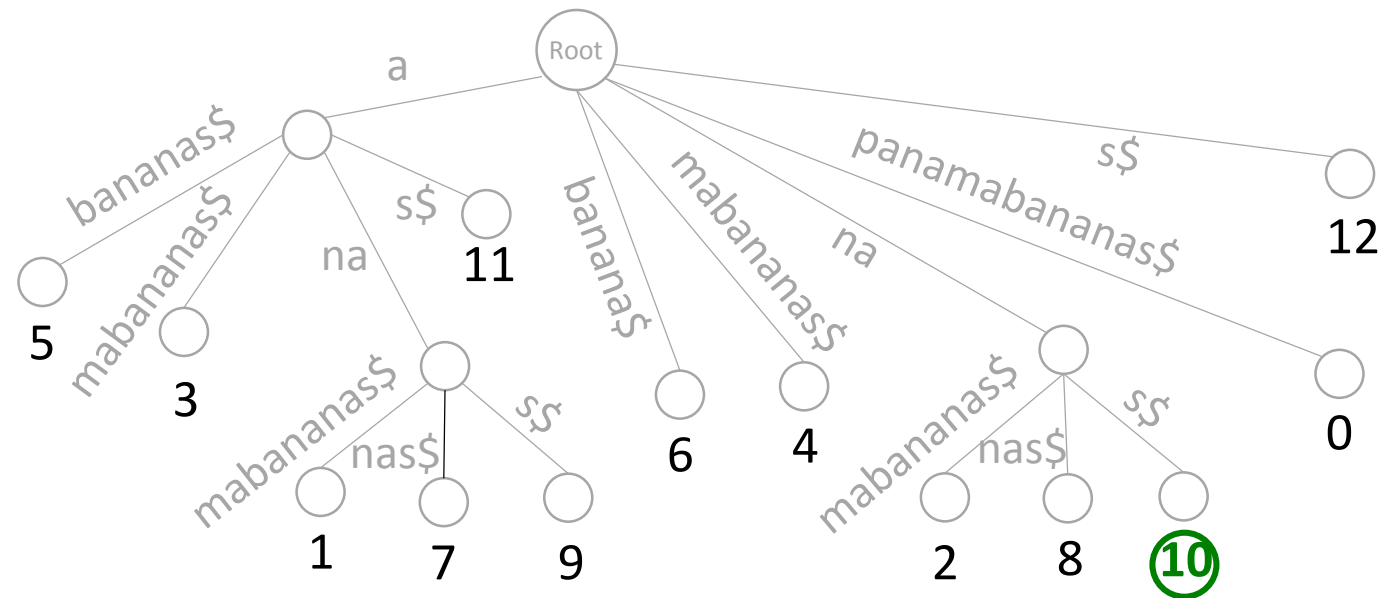
[13 5 3 1 7 9 11 6 4 **2** 8 10 0 12]

From Suffix Tree to Suffix Array



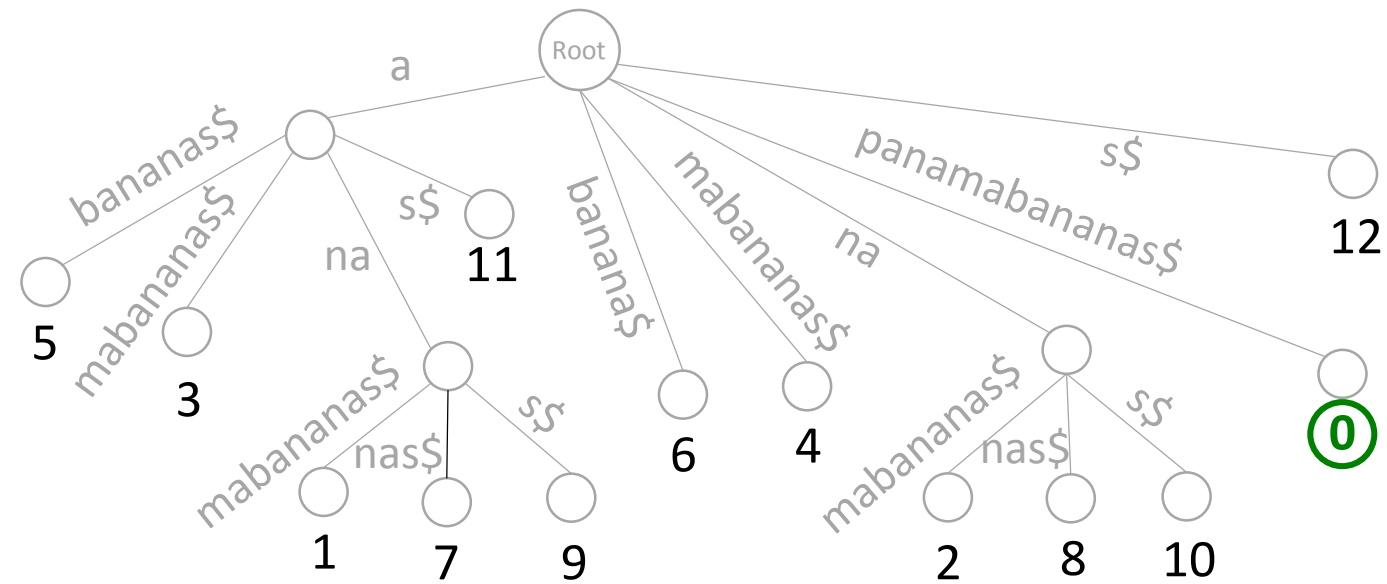
[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array



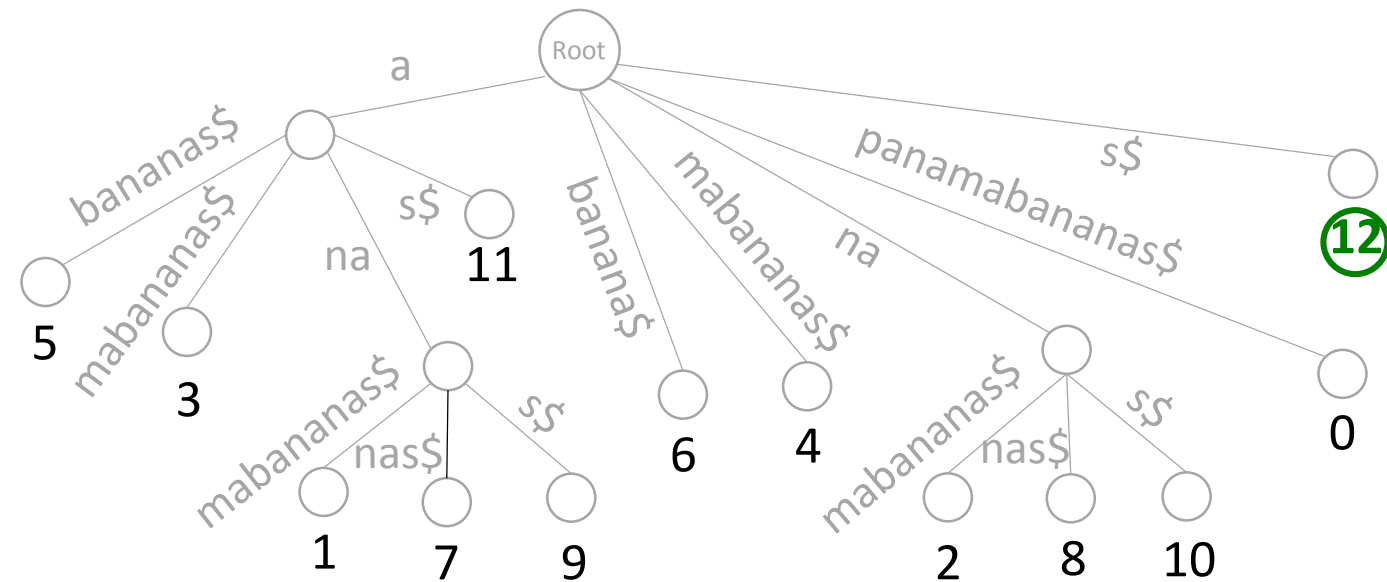
[13 5 3 1 7 9 11 6 4 2 8 **10** 0 12]

From Suffix Tree to Suffix Array



[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

From Suffix Tree to Suffix Array

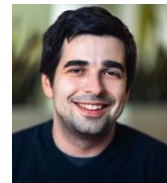


[13 5 3 1 7 9 11 6 4 2 8 10 0 12]

Constructing Suffix Array

- Depth-first traversal of suffix tree
 - $O(|Text|)$ time and $\sim 20 \cdot |Text|$ space
- Manber-Myers algorithm (1990):
 - $O(|Text|)$ time and $\sim 4 \cdot |Text|$ space
- But memory footprint is still large for human genome!

We will learn how to quickly construct suffix array
without relying on suffix tree later in this course



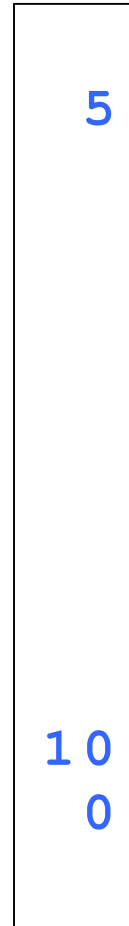
Reducing Memory Footprint for Suffix Array

- Can we store only a fraction of the suffix array but still do fast pattern matching?

1	3
	5
	3
	1
	7
	9
1	1
	6
	4
	2
	8
1	0
	0
1	2

Reducing Memory Footprint for Suffix Array

- Can we store only a fraction of the suffix array but still do fast pattern matching?
- Partial suffix array $\text{SuffixArray}_K(\text{Text})$ only contains values that are multiples of some integer K



Using the Suffix Array to Find Matches

	suffix array
s_1 panamabananas s_1	1 3
a_1 bananas\$panam a_1	5
a_2 mabananas\$pan a_2	3
a_3na mabananas\$pa a_3	1
a_4na nas\$panamab a_4	7
a_5na s\$panamaban a_5	9
a_6 s\$panamaban a_6	1 1
b_1 ananas\$panama b_1	6
m_1 abananas\$pana m_1	4
n_1 amabananas\$pa n_1	2
n_2 anas\$panamaba n_2	8
n_3 as\$panamabana n_3	1 0
p_1 anamabananas\$ p_1	0
s_1 \$panamabana s_1	1 2

Using the Partial Suffix Array to Find Matches

partial
suffix
array

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$pa₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

5

10

0

Using the Partial Suffix Array to Find Matches

partial
suffix
array

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panamab₁
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

5

Where are these **ana** prefixes located in *Text*???

10

0

Focus on **a₄na**

partial
suffix
array

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄nanas\$panama**b₁**
a₅nas\$panamaban₂
a₆s\$panamaban₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

Where is **a₄na**?

5

10

0

Focus on b_1 ana

partial
suffix
array

\$₁panamabananas₁
a₁bananas\$panam₁
a₂mabananas\$pan₁
a₃namabananas\$p₁
a₄**n**anas\$panama**b**₁
a₅**n**a\$s\$panamaban₂
a₆s\$panamaban₃
b₁**a****n**aanas\$panama**a**₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

5

Where is b_1 ana?

10

0

Focus on a_1bana

partial
suffix
array

\$₁panamabananas₁
 a_1 b ananas\$panam₁
a₂mabananas\$pan₁
 a_3 namabananas\$p₁
 a_4 nanas\$panama b_1
 a_5 nas\$panamaban₂
a₆s\$panamaban₃
 b_1 a nanas\$panam a_1
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabanana₆

Where is a_1bana ?

5

10

0

Partial suffix array reveals position of **a₁bana**

\$₁ panamabananas₁
a₁ bananas\$panam₁
 a₂ mabananas\$pan₁
a₃ namabananas\$pa₁
a₄ nanas\$panamab₁
a₅ nas\$panamaban₂
 a₆ s\$panamabanan₃
b₁ bananas\$panam**a₁**
 m₁ abananas\$pana₂
 n₁ amabananas\$pa₃
 n₂ anas\$panamaba₄
 n₃ as\$panamabana₅
 p₁ anamabananas\$₁
 s₁ \$panamabana₆

a₁bana is at position 5

a₄na is at position 7

b₁ana is at position 6

partial
suffix
array

5

7

6

1 0

0

Outline

- Burrows-Wheeler Transform
- Inverting Burrows-Wheeler Transform
- Using BWT for Pattern Matching
- Suffix Arrays
- **Approximate Pattern Matching**

Returning to Search for Mutations

- **Approximate Pattern Matching Problem:**
 - **Input:** A string *Pattern*, a string *Text*, and an integer d .
 - **Output:** All positions in *Text* where the string *Pattern* appears as a substring with at most d mismatches.

Revealing Mutations by Analyzing **Billions** of Reads

- **Multiple Approximate Pattern Matching Problem**
 - **Input:** A **set** of strings *Patterns*, a string *Text*, and an integer d .
 - **Output:** All positions in *Text* where a string from *Patterns* appears as a substring with at most d mismatches.

BWT Saves the Day Again

- searching for ana in panamabananas

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamaban3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanana6
```

BWT Saves the Day Again

- searching for an **a** in panamabananas

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamaban3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanana6
```

BWT Saves the Day Again

- searching for **ana** in panamabananas

\$₁panamabananas₁
a₁bananas\$pana**m**₁
a₂mabananas\$pa**n**₁
a₃namabananas\$**p**₁
a₄nanas\$panama**b**₁
a₅nas\$panamaba**n**₂
a₆s\$panamabana**n**₃
b₁ananas\$panama₁
m₁abananas\$pana₂
n₁amabananas\$pa₃
n₂anas\$panamaba₄
n₃as\$panamabana₅
p₁anamabananas\$₁
s₁\$panamabana₆

Exact matching

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

To allow for 1 mismatch, we need to analyze the rows ending in red letters as well.

```
$1panamabananas1  
a1bananas$panam1  
a2mabananas$pan1  
a3namabananas$p1  
a4nanas$panamab1  
a5nas$panamaban2  
a6s$panamabanan3  
b1ananas$panama1  
m1abananas$pana2  
n1amabananas$pa3  
n2anas$panamaba4  
n3as$panamabana5  
p1anamabananas$1  
s1$panamabanaa6
```

Approximate matching
with at most 1 mismatch

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

To allow for 1 mismatch, we need to analyze the rows ending in red letters as well.

	# Mismatches
\$ ₁ panamabananas ₁	
a ₁ bananas\$pana m ₁	1
a ₂ mabananas\$pa n ₁	0
a ₃ namabananas\$ p ₁	1
a ₄ nanas\$panama b ₁	1
a ₅ nas\$panamaba n ₂	0
a ₆ s\$panamabana n ₃	0
b ₁ ananas\$panama ₁	
m ₁ abananas\$pana ₂	
n ₁ amabananas\$pa ₃	
n ₂ anas\$panamaba ₄	
n ₃ as\$panamabana ₅	
p ₁ anamabananas\$ ₁	
s ₁ \$panamabanana ₆	

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

Now we analyze all rows with at most 1 mismatch using the First-Last property.

	# Mismatches
\$ ₁ panamabananas ₁	
a ₁ bananas\$pana m ₁	1
a ₂ mabananas\$pa n ₁	0
a ₃ namabananas\$ p ₁	1
a ₄ nanas\$panama b ₁	1
a ₅ nas\$panamaba n ₂	0
a ₆ s\$panamabana n ₃	0
b ₁ ananas\$panama ₁	
m ₁ abananas\$pana ₂	
n ₁ amabananas\$pa ₃	
n ₂ anas\$panamaba ₄	
n ₃ as\$panamabana ₅	
p ₁ anamabananas\$ ₁	
s ₁ \$panamabanana ₆	

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

Now we analyze all rows with at most 1 mismatch using the First-Last property.

	# Mismatches
\$ ₁ panamabananas ₁	
a ₁ bananas\$pana m ₁	1
a ₂ mabananas\$pa n ₁	0
a ₃ namabananas\$ p ₁	1
a ₄ nanas\$panama b ₁	1
a ₅ nas\$panamaba n ₂	0
a ₆ s\$panamabana n ₃	0
b ₁ a nanas\$panama ₁	
m ₁ a bananas\$pana ₂	
n ₁ a mabananas\$pa ₃	
n ₂ a nas\$panamaba ₄	
n ₃ a s\$panamabana ₅	
p ₁ a namabananas\$ ₁	
s ₁ \$panamabana ₆	

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

Now we analyze all rows with at most 1 mismatch using the First-Last property.

	# Mismatches
\$ ₁ panamabananas ₁	
a ₁ bananas\$panam ₁	
a ₂ mabananas\$pan ₁	
a ₃ namabananas\$p ₁	
a ₄ nanas\$panamab ₁	
a ₅ nas\$panamaban ₂	
a ₆ s\$panamaban ₃	
b ₁ a nanas\$panama ₁	1
m ₁ a bananas\$pana ₂	1
n ₁ a mabananas\$pa ₃	0
n ₂ a nas\$panamaba ₄	0
n ₃ a s\$panamabana ₅	0
p ₁ a namabananas\$ ₁	1
s ₁ \$panamabanana ₆	

BWT Pattern Matching with 1 Mismatch

- searching for **ana** in panamabananas

	# Mismatches
\$ ₁ panamabananas ₁	
a ₁ bananas\$panam ₁	
a ₂ mabananas\$pan ₁	
a ₃ namabananas\$p ₁	
a ₄ nanas\$panamab ₁	
a ₅ nas\$panamaban ₂	
a ₆ s\$panamaban ₃	
b ₁ a nanas\$panam a ₁	1
m ₁ a bananas\$pan a ₂	1
n ₁ a mabananas\$p a ₃	0
n ₂ a nas\$panamab a ₄	0
n ₃ a s\$panamaban a ₅	0
p ₁ a namabananas\$ s ₁	2
s ₁ \$panamabanana ₆	

This row results in a 2nd mismatch (the **s**), so we discard it.

Five Approximate Matches Found!

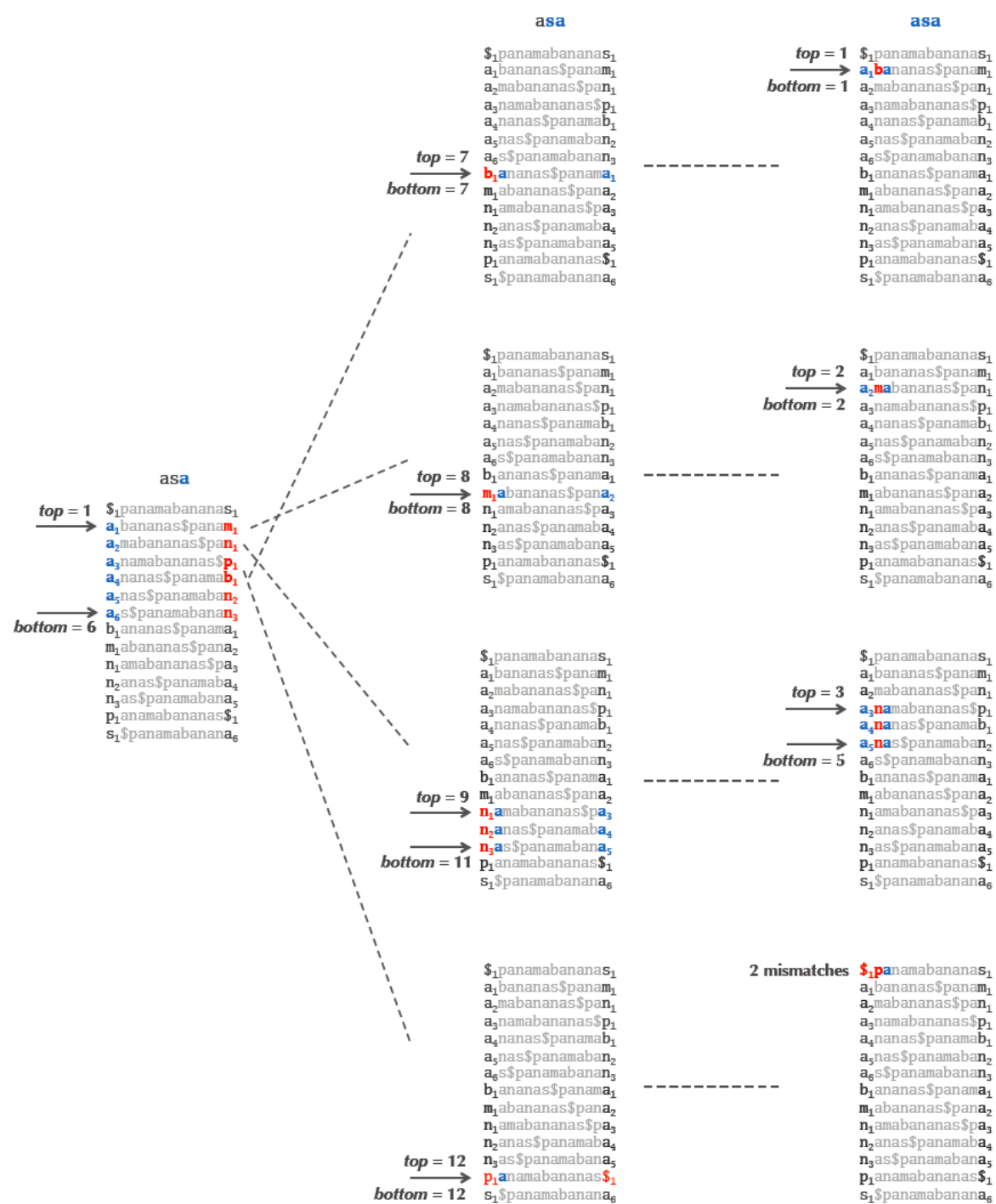
- searching for **ana** in panamabananas

	# Mismatches
\$ ₁ panamabananas ₁	
a ₁ b ananas\$panam ₁	1
a ₂ m abananas\$pan ₁	1
a ₃ n amabananas\$p ₁	0
a ₄ n anas\$panamab ₁	0
a ₅ n as\$panamaban ₂	0
a ₆ s\$panamaban ₃	
b ₁ ananas\$panam a ₁	
m ₁ abananas\$pan a ₂	
n ₁ amabananas\$p a ₃	
n ₂ anas\$panamab a ₄	
n ₃ as\$panamaban a ₅	
p ₁ anamabananas\$ s ₁	
s ₁ \$panamabanana ₆	

Where Are The Matches?

- searching for **ana** in panamabananas

Suffix Array	
\$ ₁ panamabananas ₁	
a ₁ b ananas\$panam ₁	5
a ₂ m abananas\$pan ₁	3
a ₃ n amabananas\$p ₁	1
a ₄ n anas\$panamab ₁	7
a ₅ n as\$panamaban ₂	9
a ₆ s\$panamaban ₃	
b ₁ ananas\$panama ₁	
m ₁ abananas\$pana ₂	
n ₁ amabananas\$pa ₃	
n ₂ anas\$panamaba ₄	
n ₃ as\$panamabana ₅	
p ₁ anamabananas\$ ₁	
s ₁ \$panamabanana ₆	



In reality, approximate pattern matching with BWT is more complex (we omitted various details)

