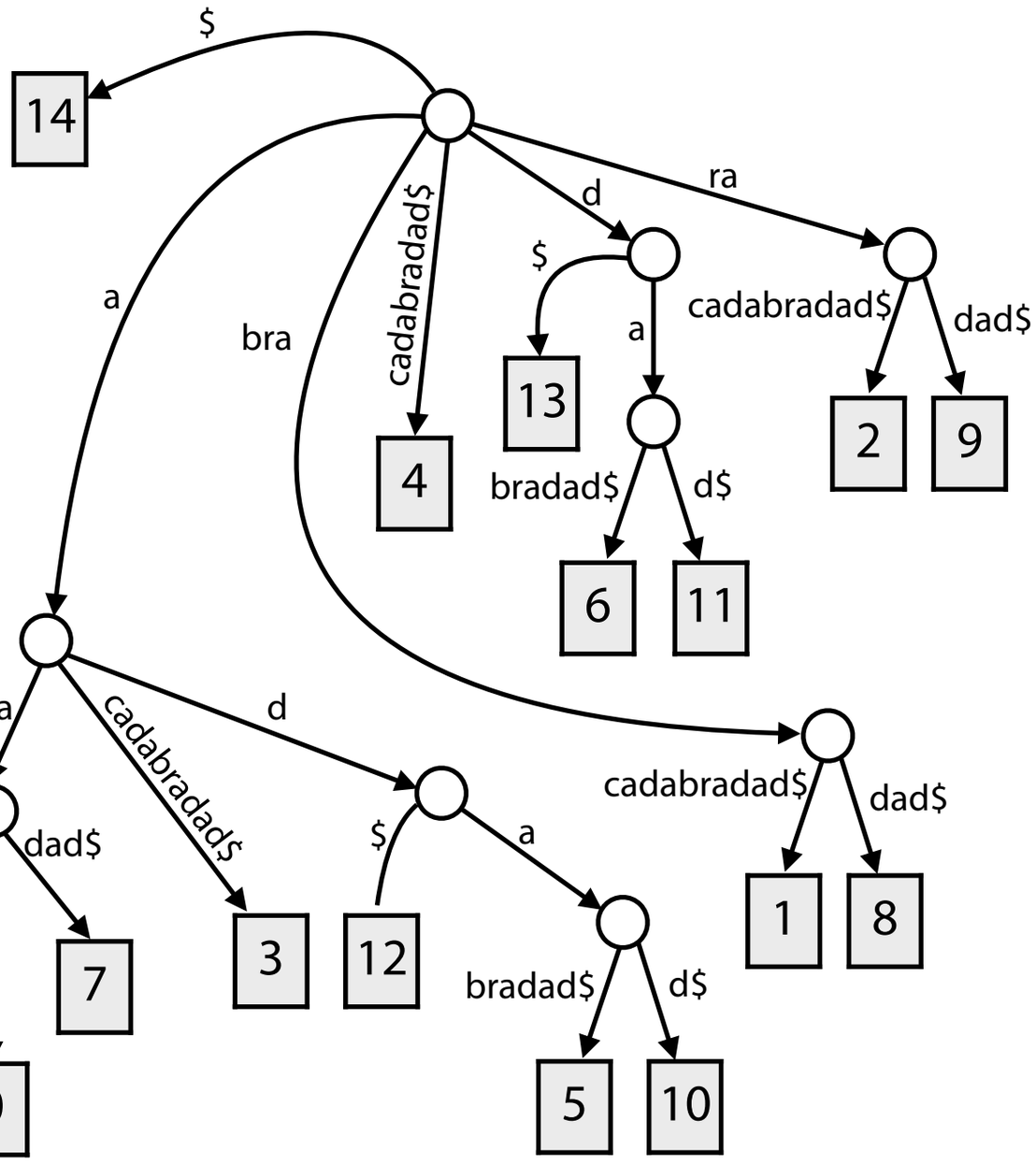
Suffix Arrays: the suffix tree is hiding

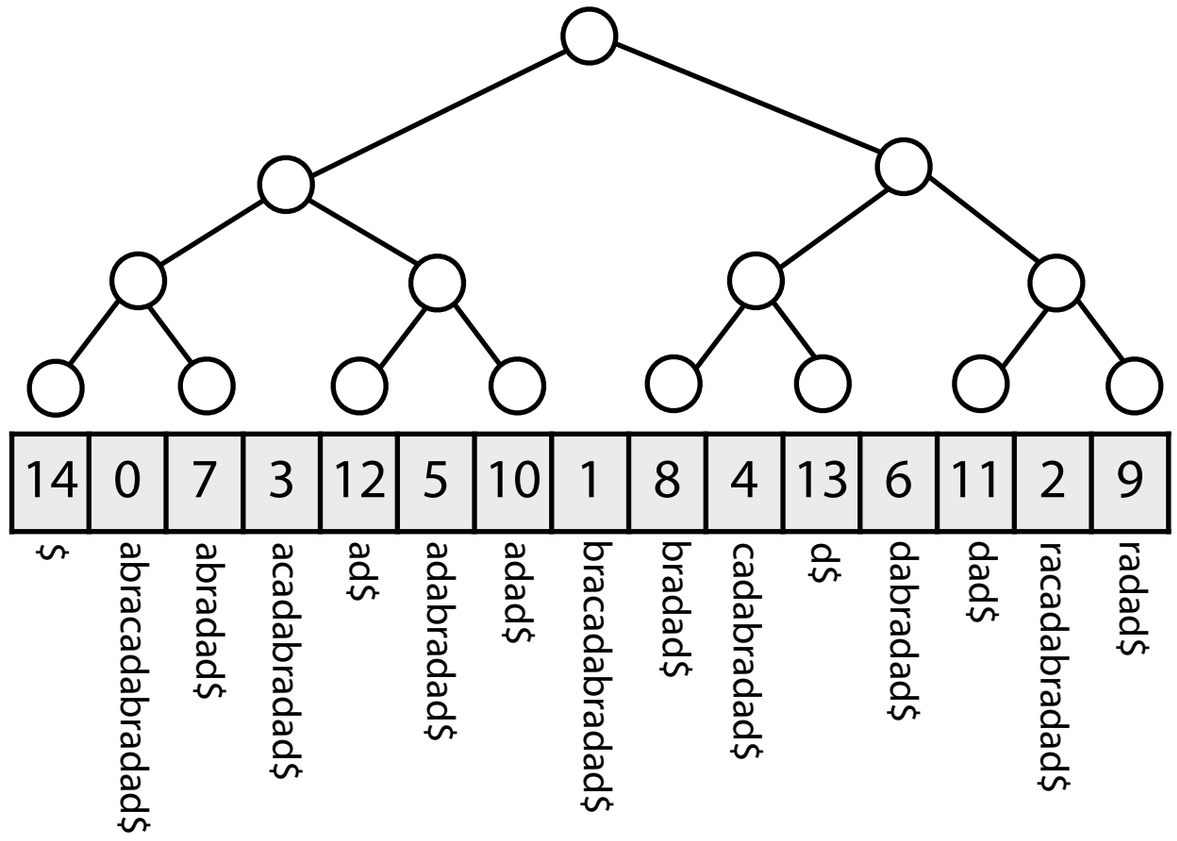
Ben Langmead

Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).

Suffix array



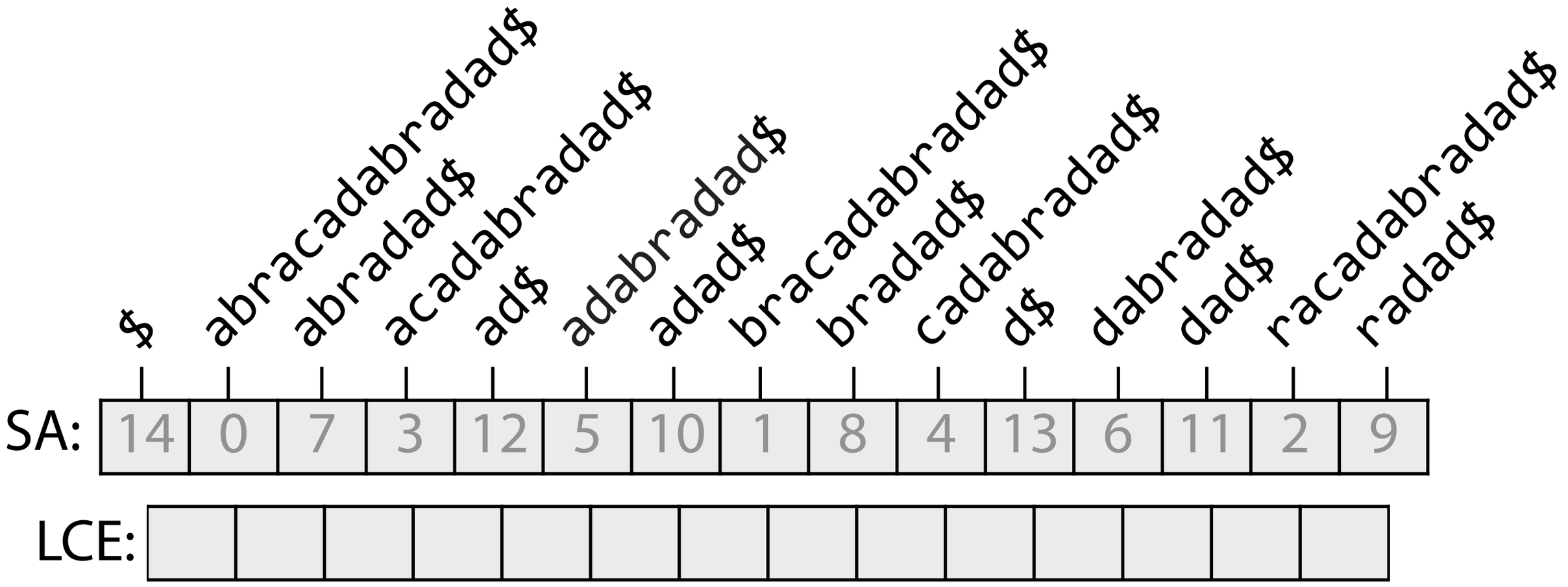
Can build the suffix **array** from suffix **tree**



Both encode trees

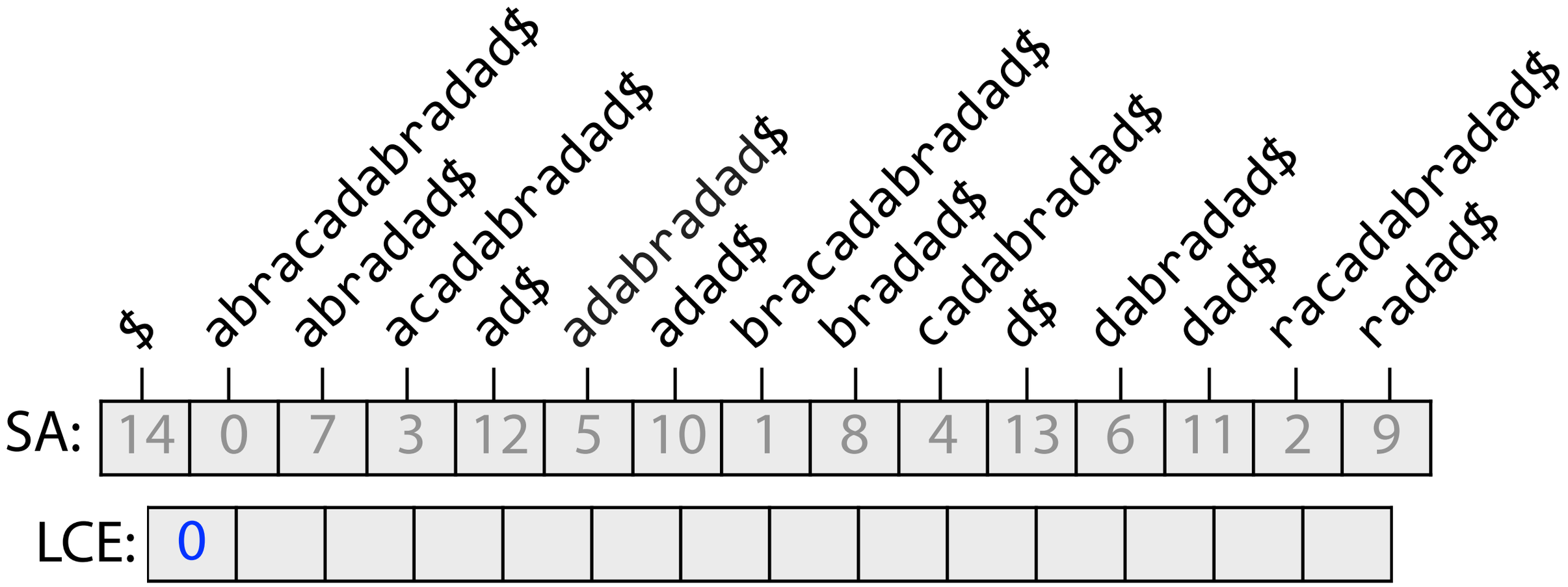
Further: the suffix tree can be **recovered from** the suffix array

Suffix array



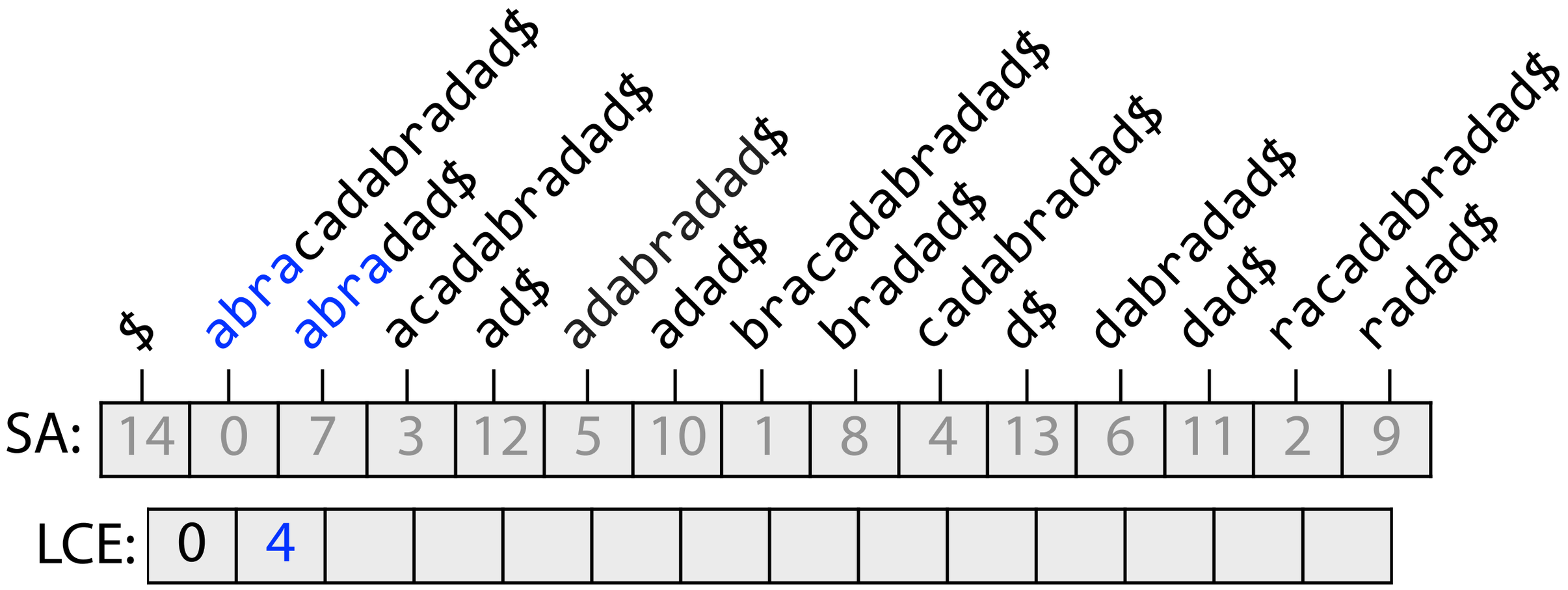
Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array



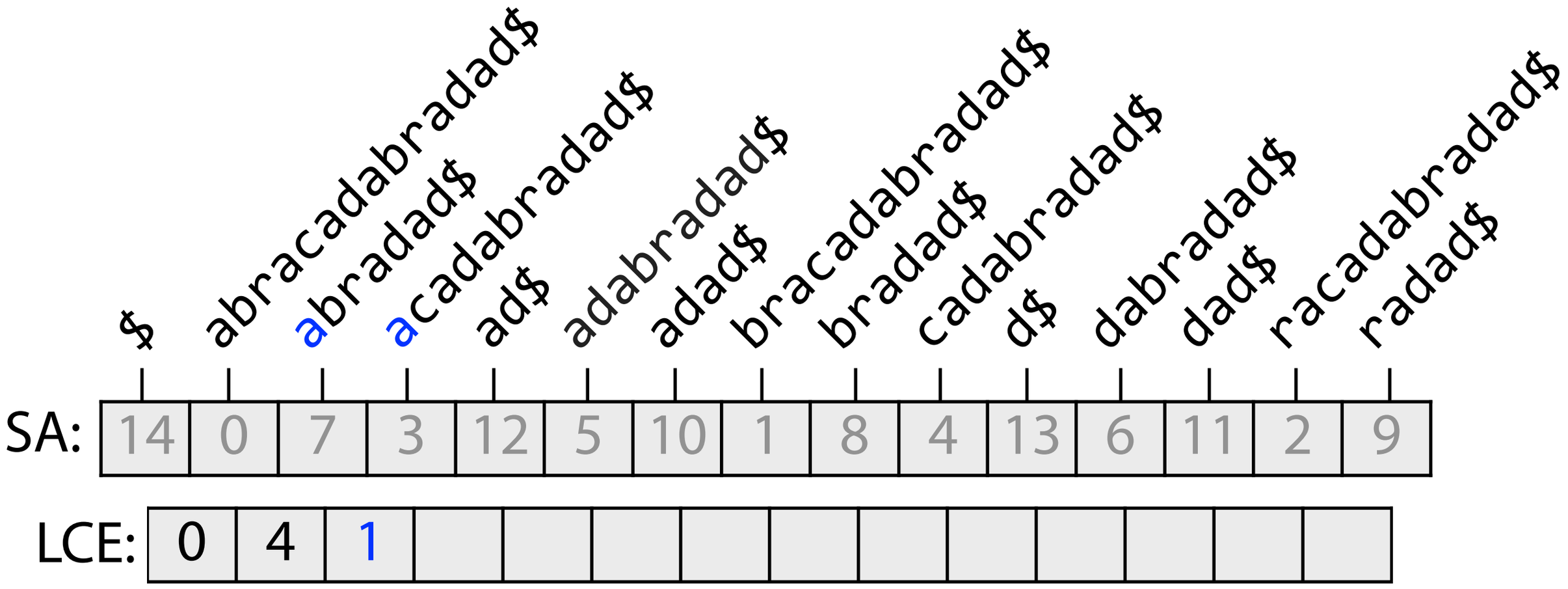
Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array



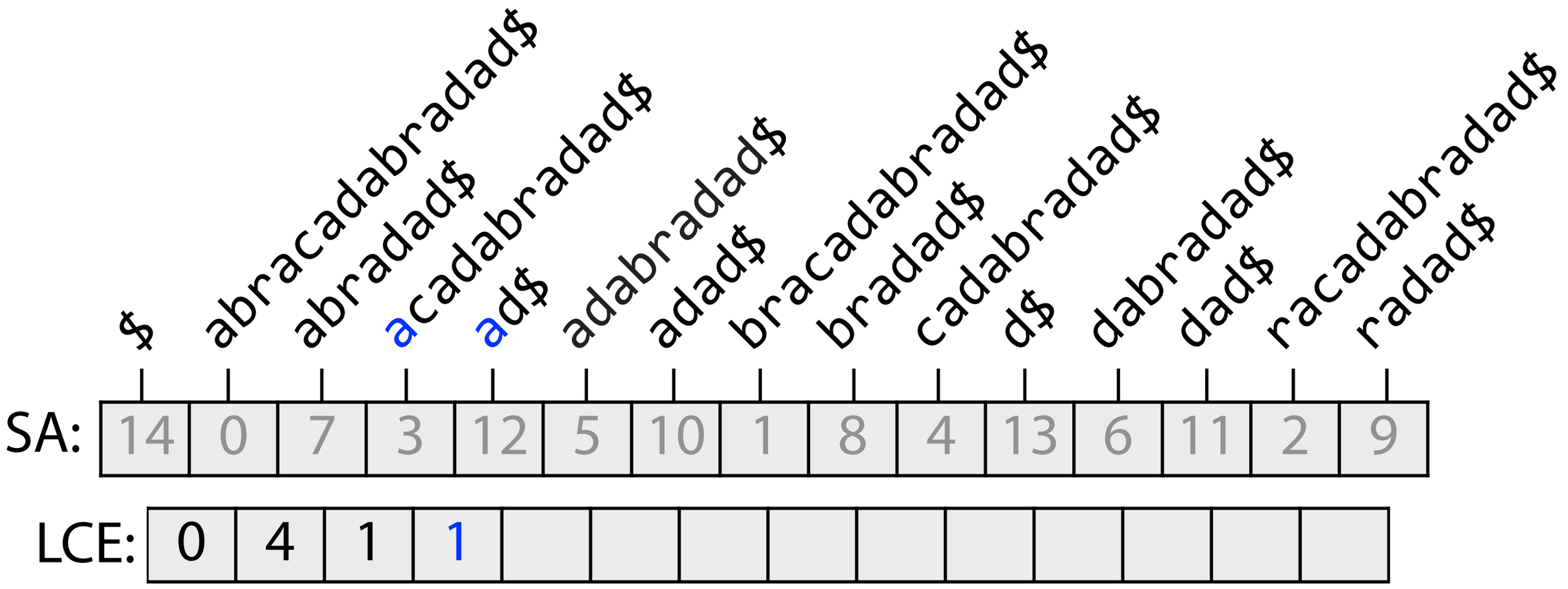
Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array



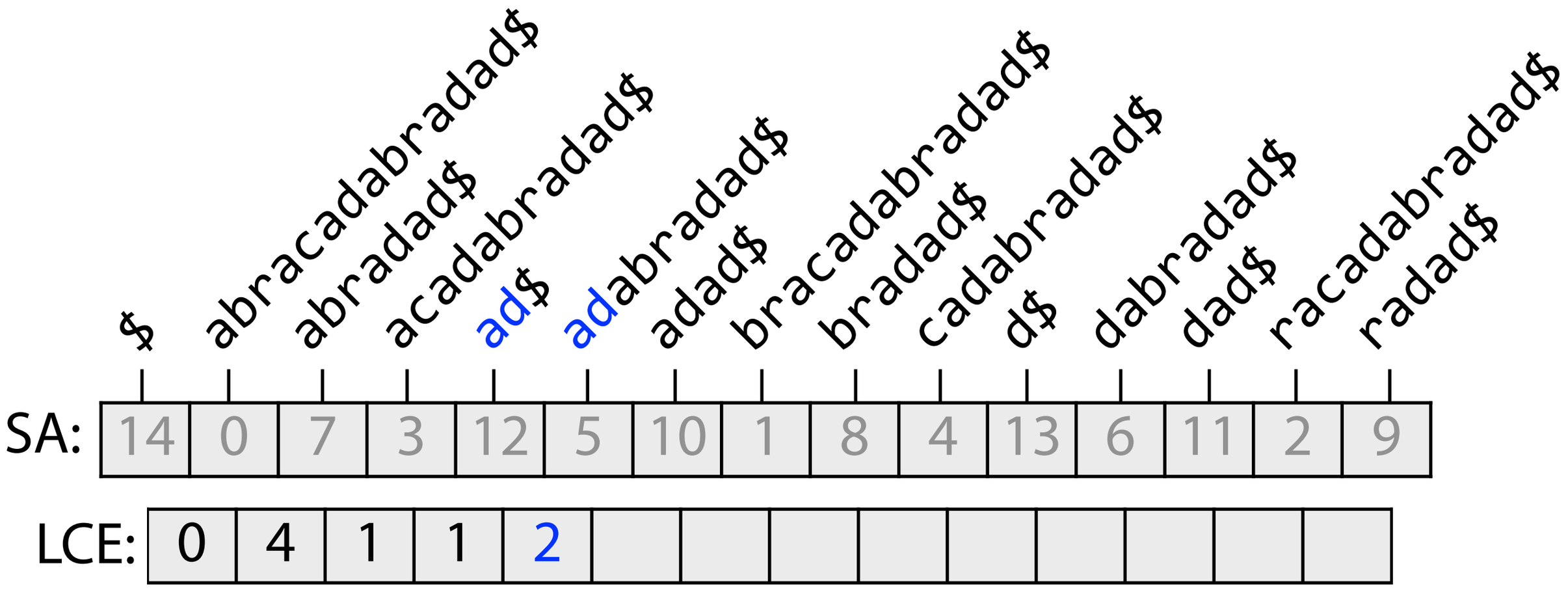
Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array



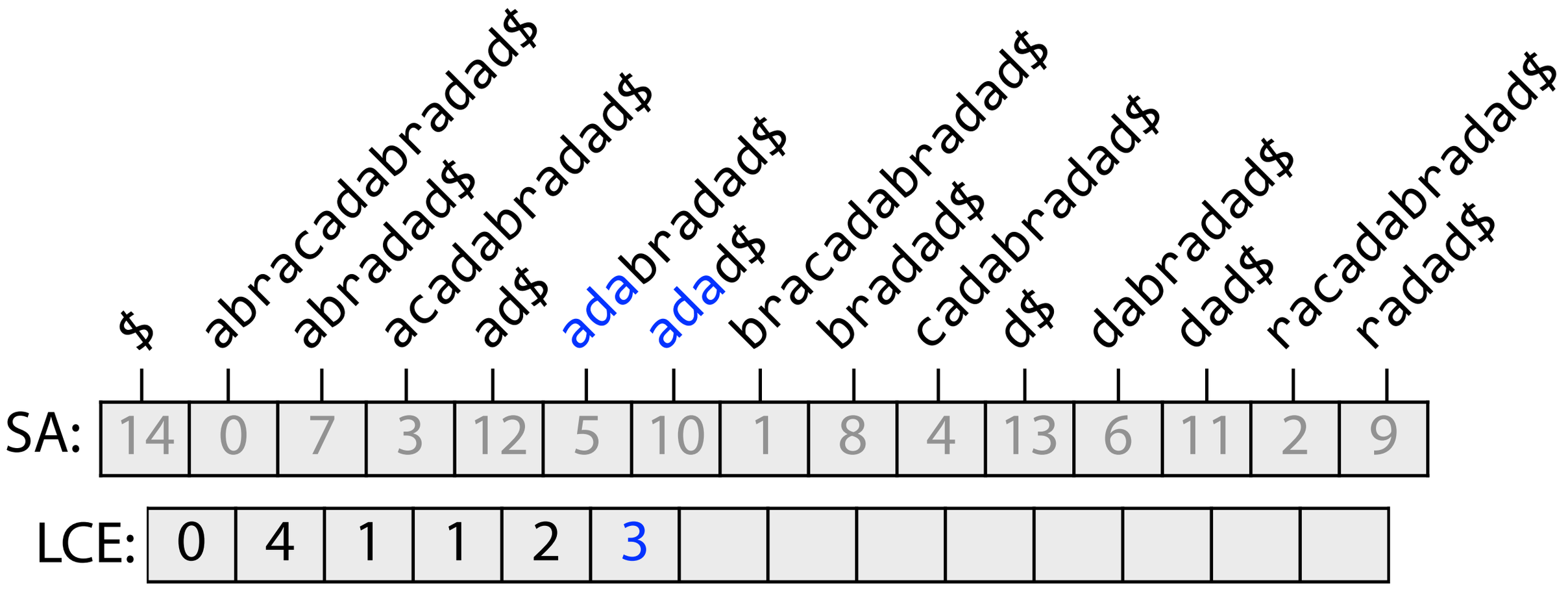
Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array



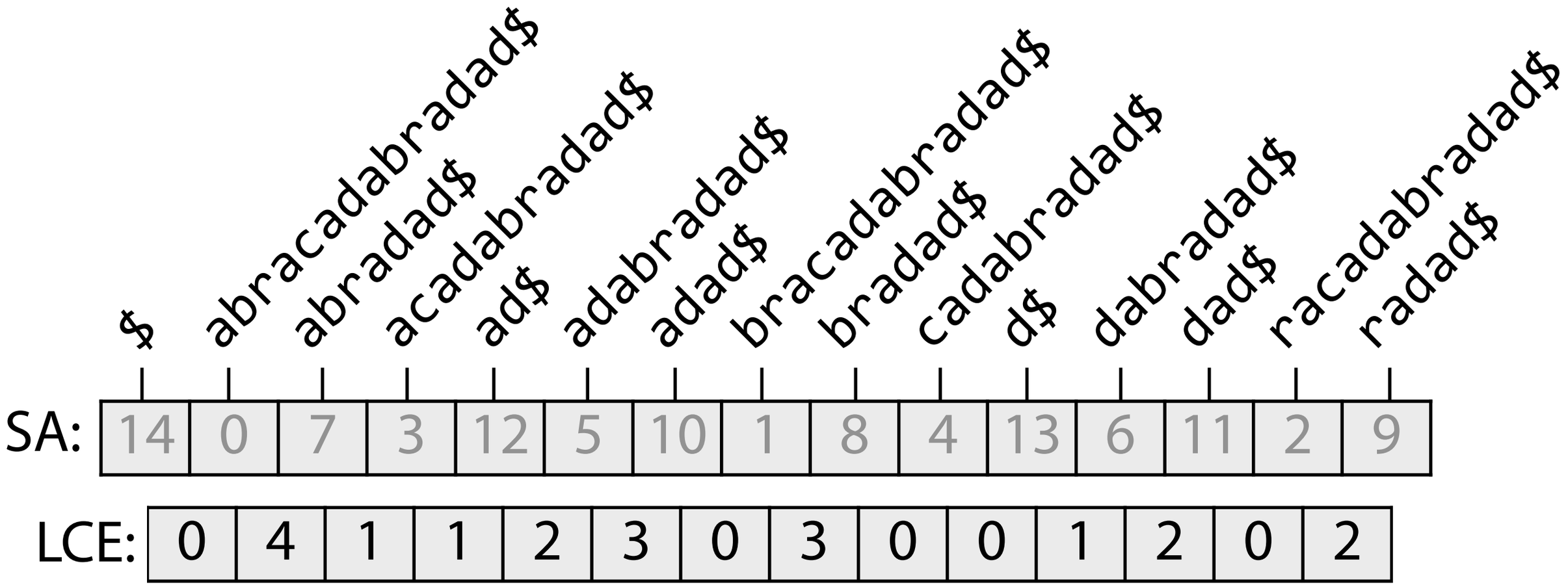
Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array



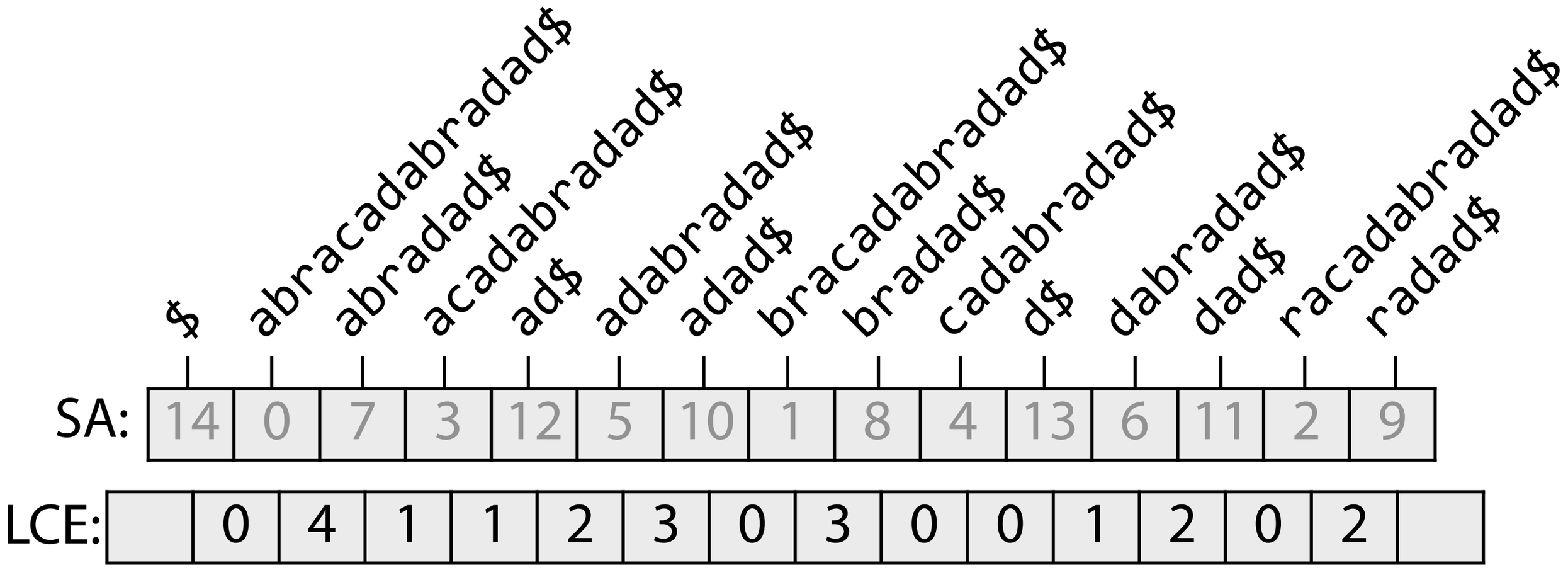
Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array



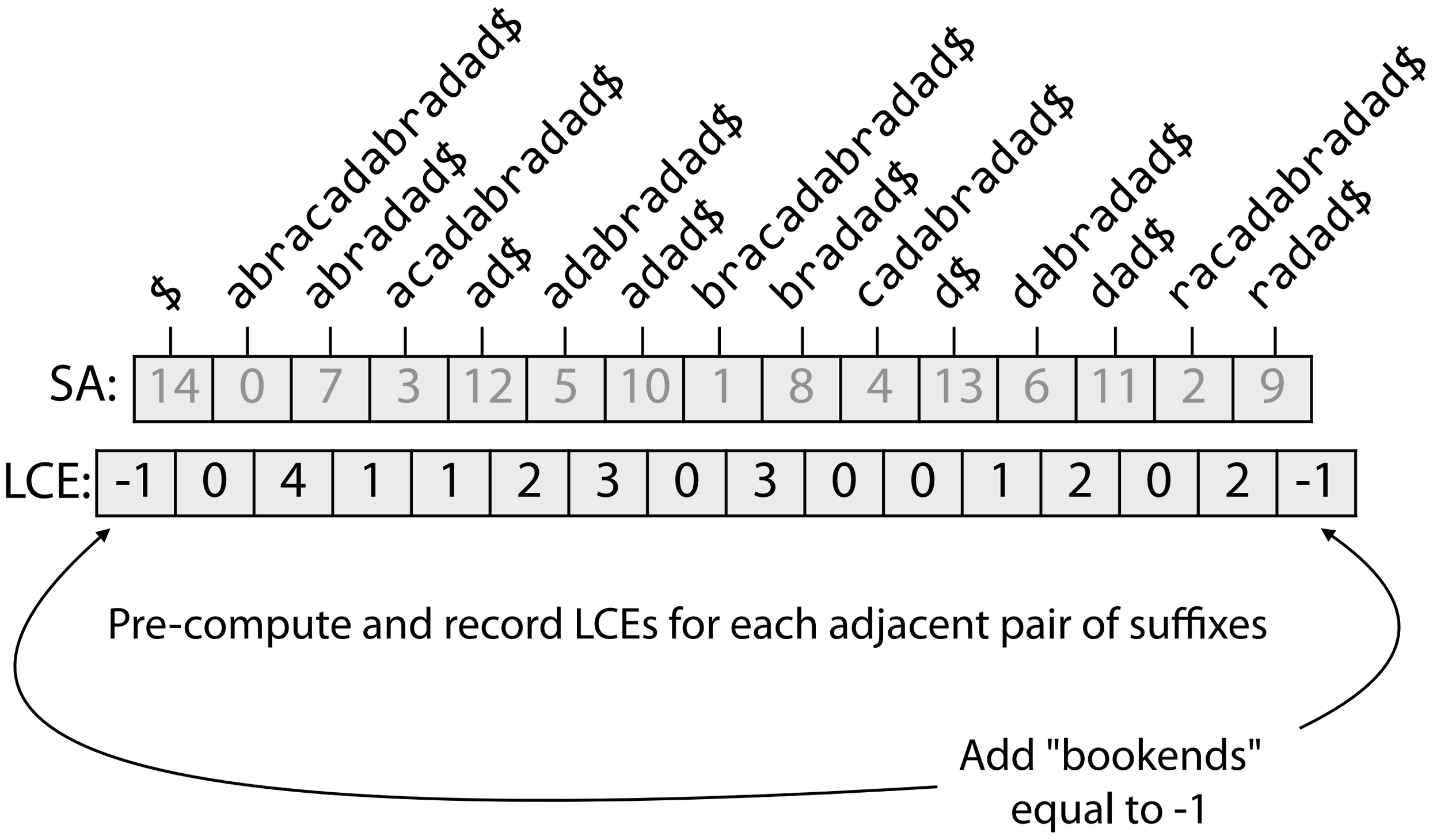
Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array

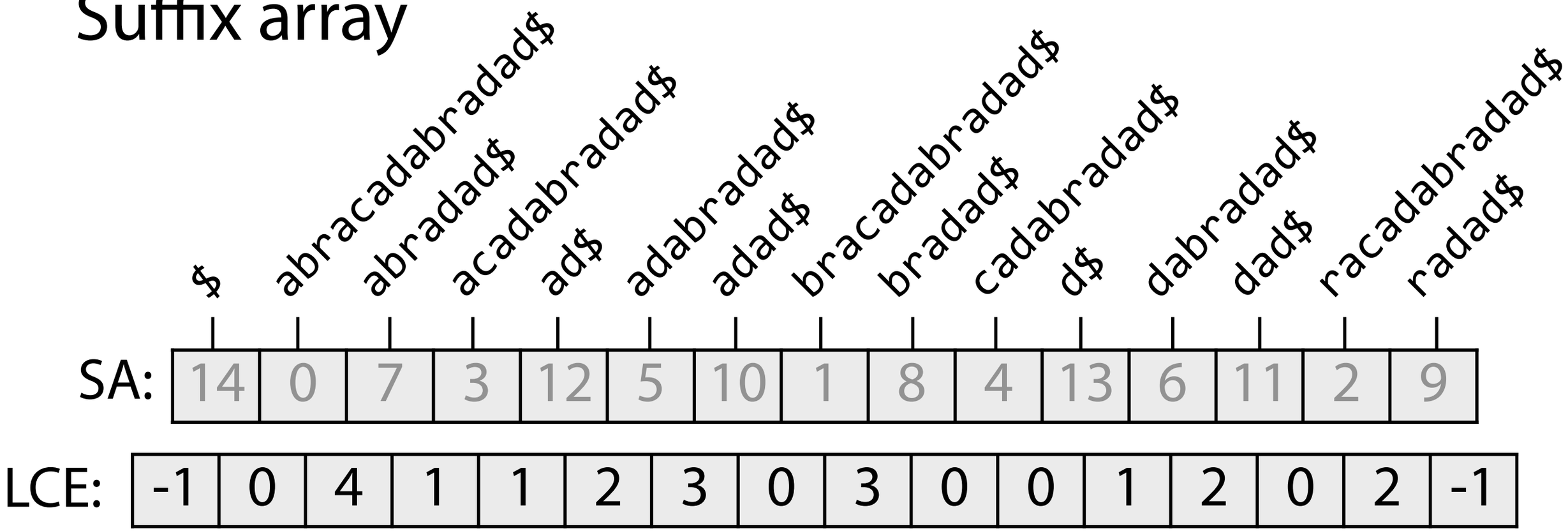


Pre-compute and record LCEs for each adjacent pair of suffixes

Suffix array



Suffix array

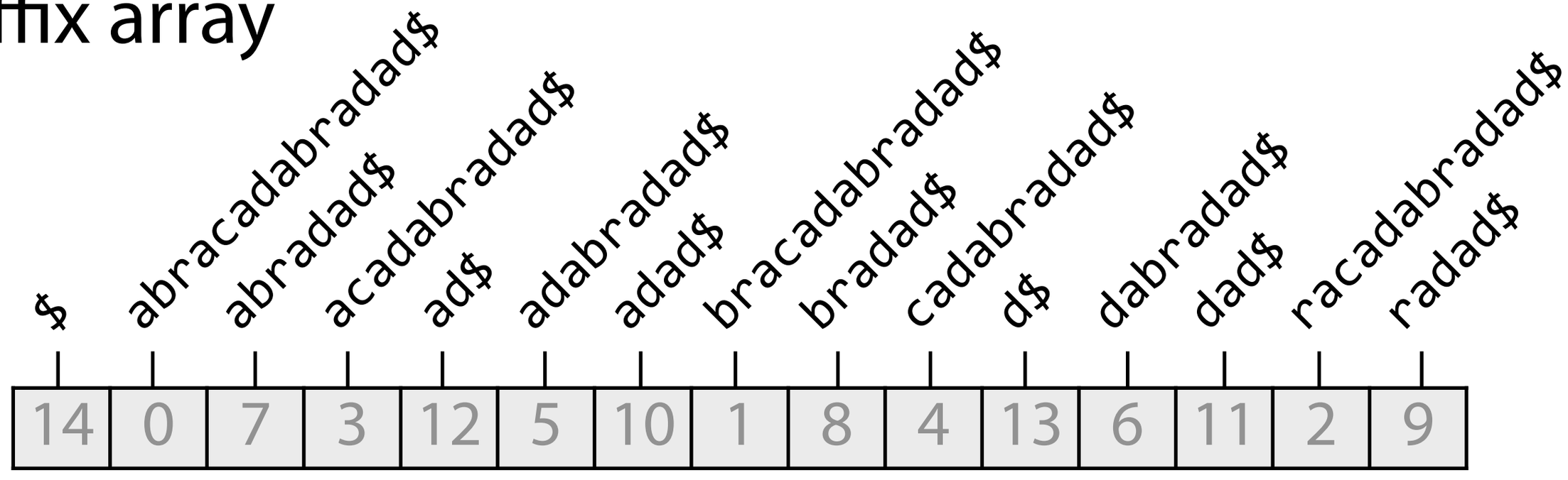


Certain **intervals** of the SA are ℓ-intervals

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

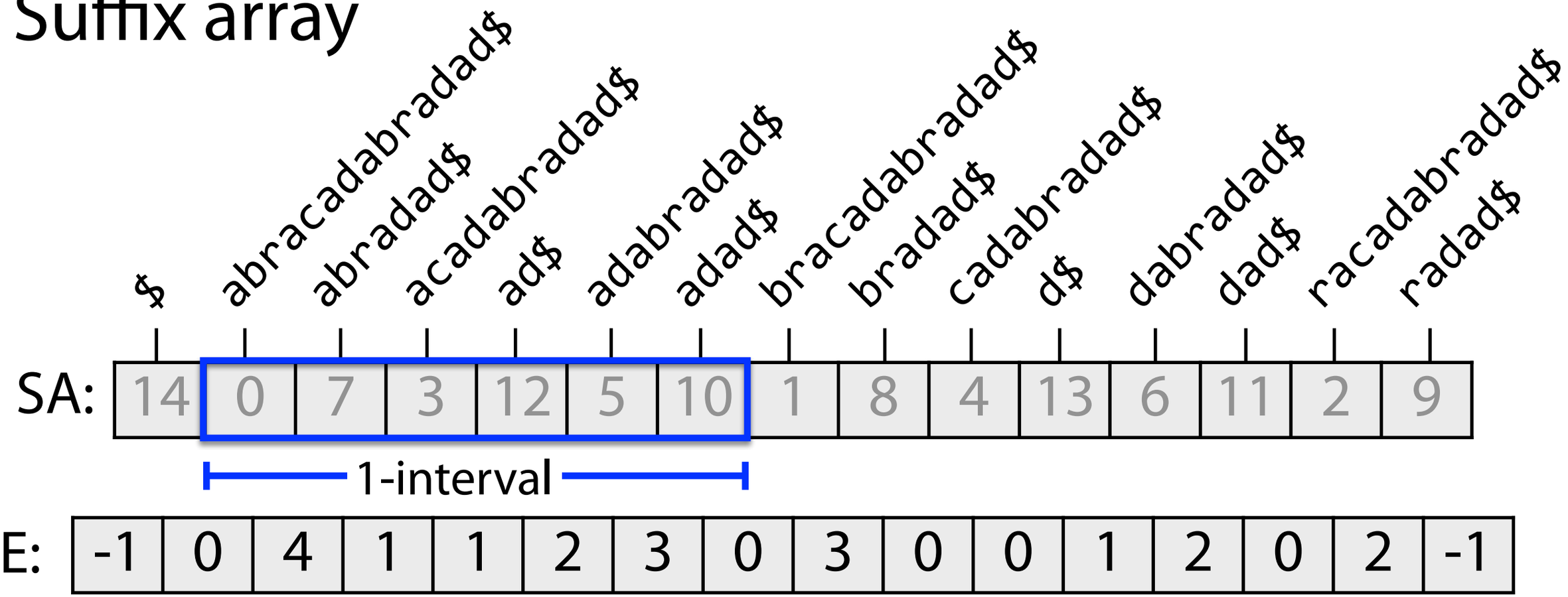


LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array



LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

	$	abracadabradad$	abradad$	acadabradad$	ad$	adabradad$	adad$	bracadabradad$	bradad$	cadabradad$	d$	dabradad$	dad$	racadabradad$	radad$
SA:	14	0	7	3	12	5	10	1	8	4	13	6	11	2	9

1-interval

LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

	$	abracadabradad$	abradad$	acadabradad$	ad$	adabradad$	adad$	bracadabradad$	bradad$	cadabradad$	d$	dabradad$	dad$	racadabradad$	radad$
SA:	14	0	7	3	12	5	10	1	8	4	13	6	11	2	9

1-interval

LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

	$	abracadabradad$	abradad$	acadabradad$	ad$	adabradad$	adad$	bracadabradad$	bradad$	cadabradad$	d$	dabradad$	dad$	racadabradad$	radad$
SA:	14	0	7	3	12	5	10	1	8	4	13	6	11	2	9

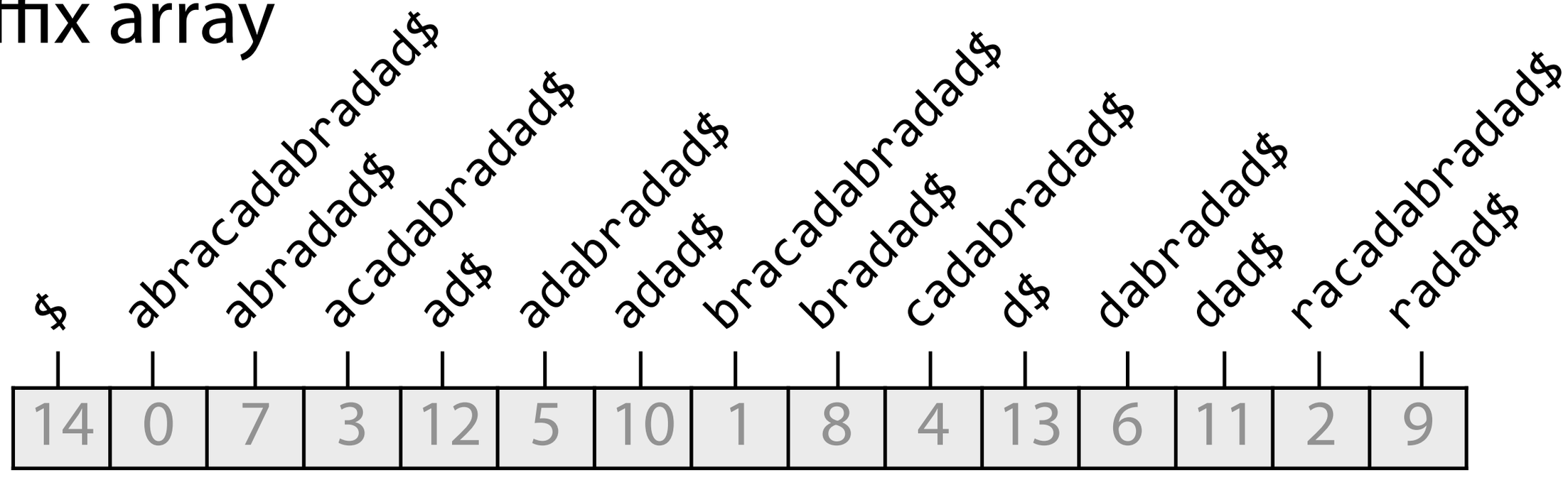
1-interval

LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

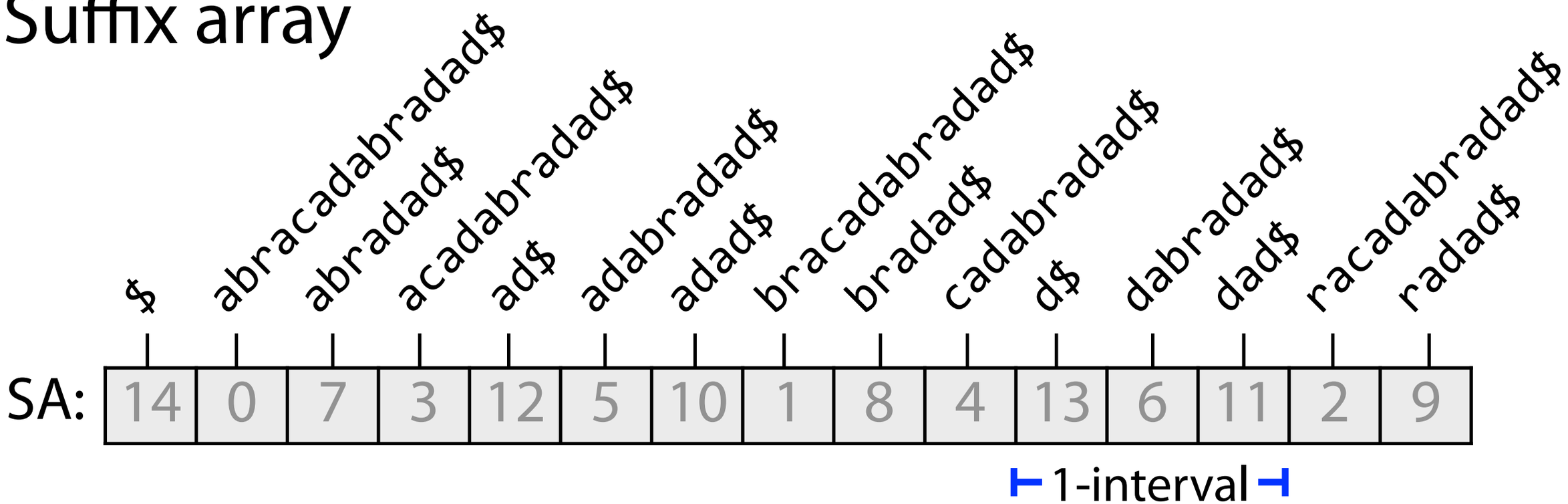


LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array



LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

	$	abracadabradad$	abradad$	acadabradad$	ad$	adabradad$	adad$	bracadabradad$	bradad$	cadabradad$	d$	dabradad$	dad$	racadabradad$	radad$
SA:	14	0	7	3	12	5	10	1	8	4	13	6	11	2	9

1-interval

LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

	$	abracadabradad$	abradad$	acadabradad$	ad$	adabradad$	adad$	bracadabradad$	bradad$	cadabradad$	d$	dabradad$	dad$	racadabradad$	radad$
SA:	14	0	7	3	12	5	10	1	8	4	13	6	11	2	9
LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

See any 2-intervals?

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

	$	abracadabradad$	abradad$	acadabradad$	ad$	adabradad$	adad$	bracadabradad$	bradad$	cadabradad$	d$	dabradad$	dad$	racadabradad$	radad$
SA:	14	0	7	3	12	5	10	1	8	4	13	6	11	2	9

2-ival

LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

	`$`	`abracadabradad$`	`abradad$`	`acadabradad$`	`ad$`	`adabradad$`	`adad$`	`bracadabradad$`	`bradad$`	`cadabradad$`	`d$`	`dabradad$`	`dad$`	`racadabradad$`	`radad$`
SA:	14	0	7	3	12	5	10	1	8	4	13	6	11	2	9

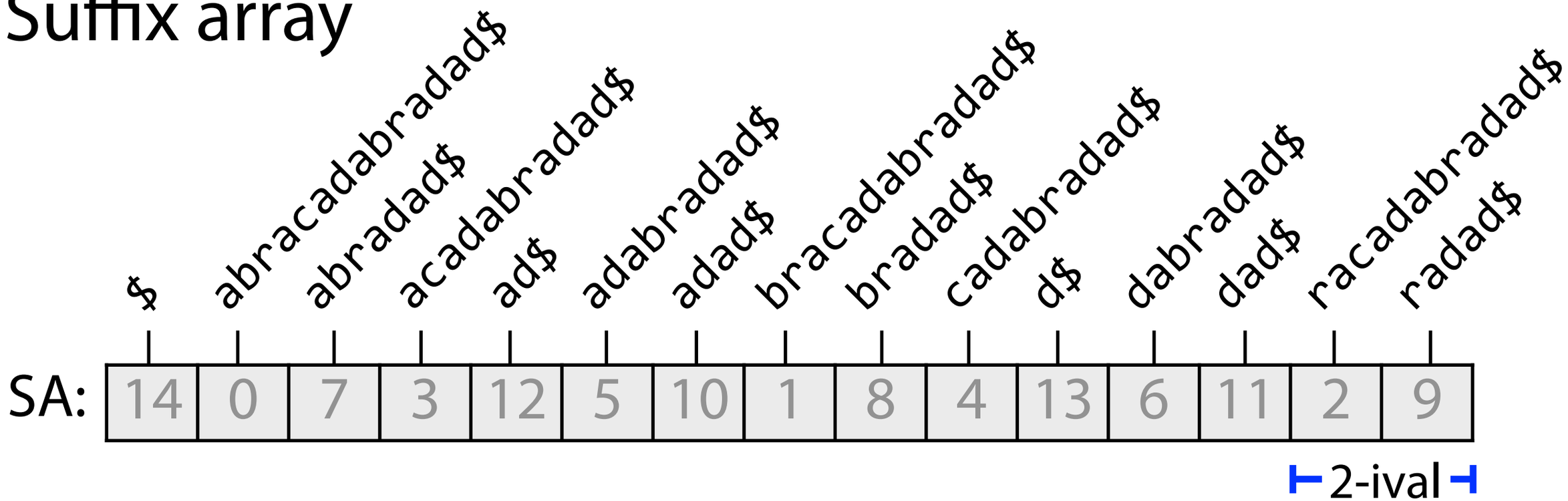
2-ival

LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array

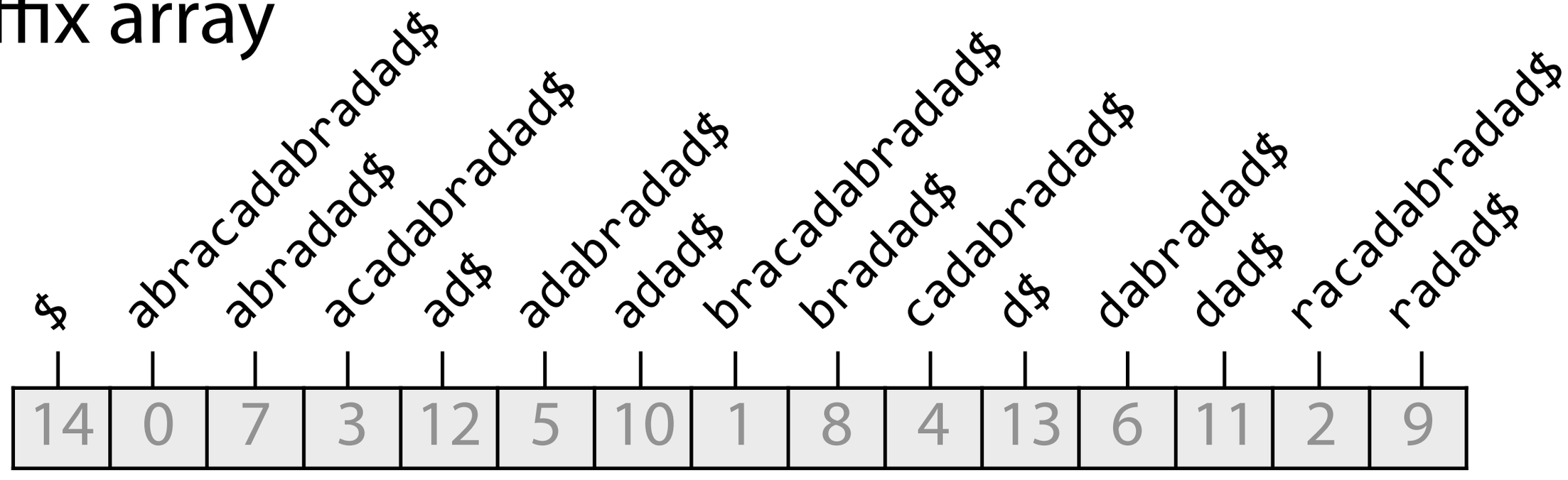


LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

Suffix array



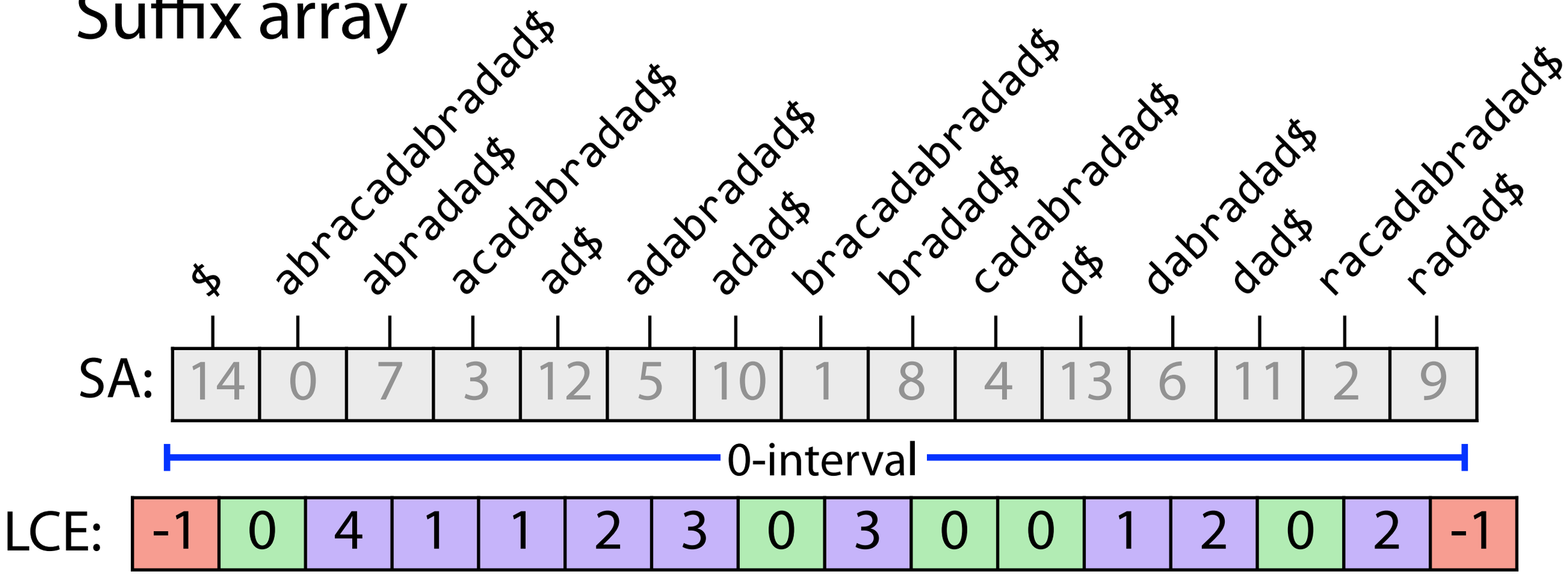
LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

Is there a 0-interval?

An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

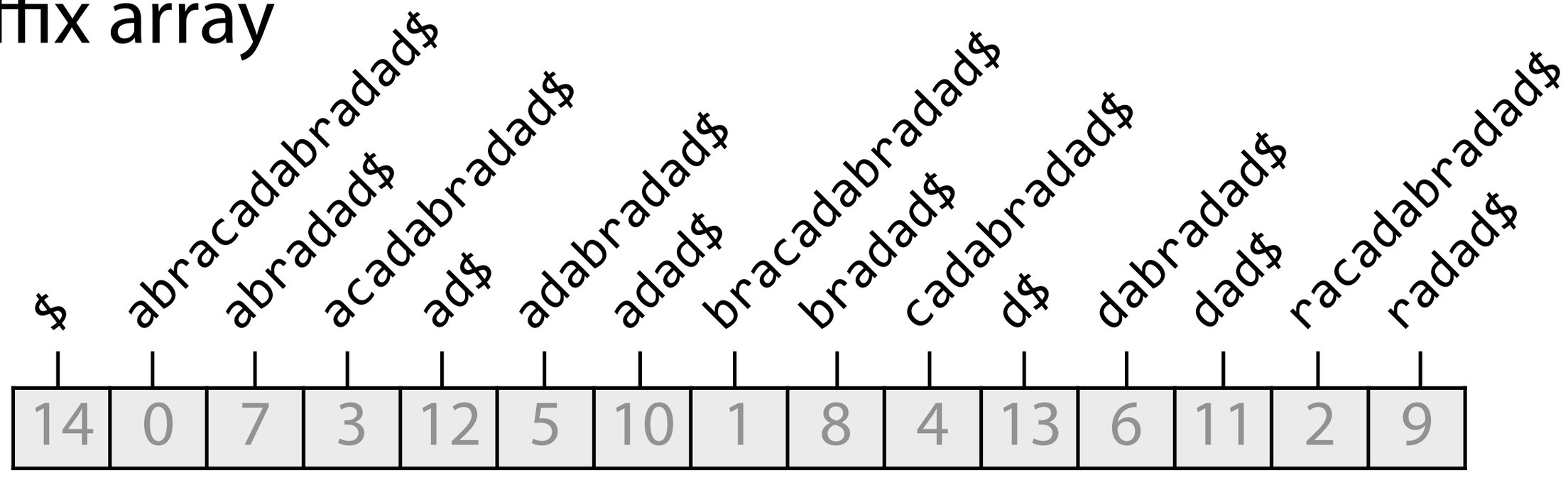
Suffix array



An interval SA$[i, j]$ is an ℓ-interval if:

1. LCEs to either side are $< \ell$
2. At least one LCE in the interval is $= \ell$
3. All other LCEs in the interval are $> \ell$

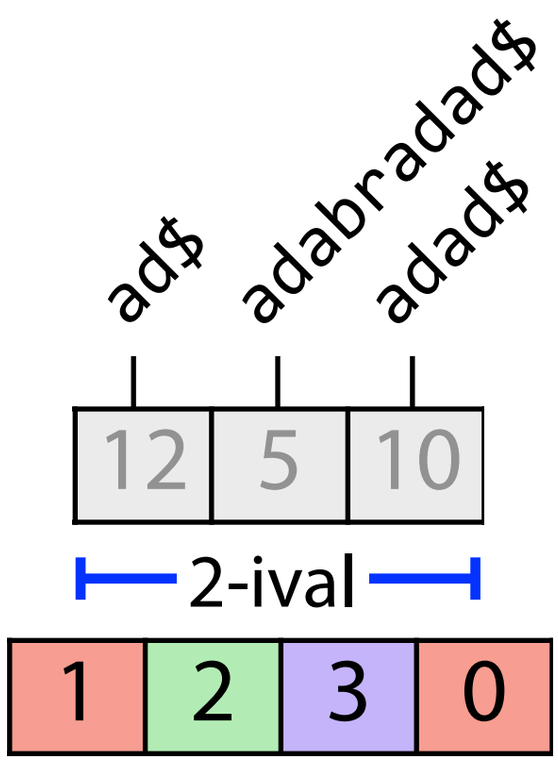
Suffix array



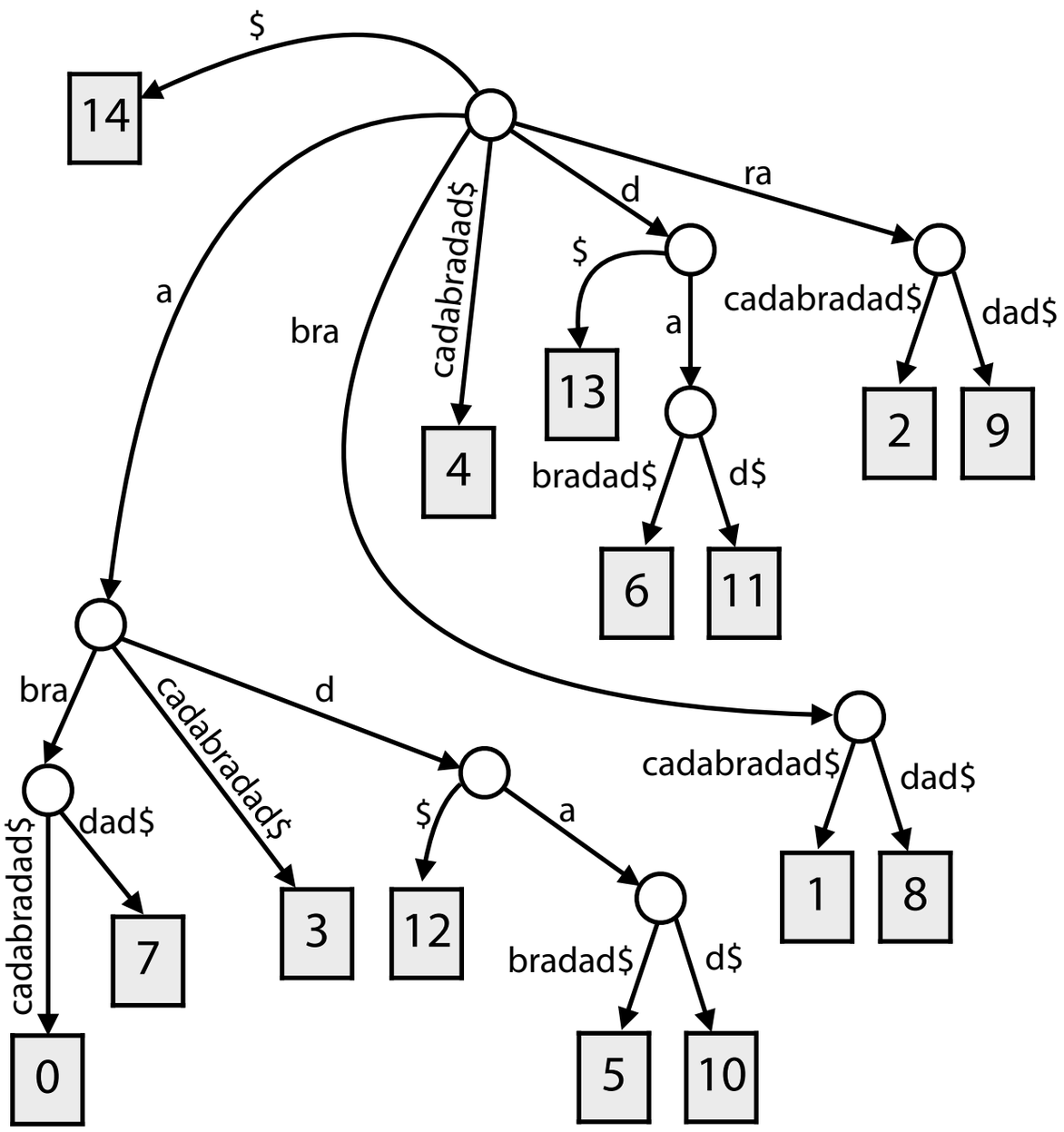
LCE:	-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

ℓ-intervals correspond to **internal nodes of the suffix tree**

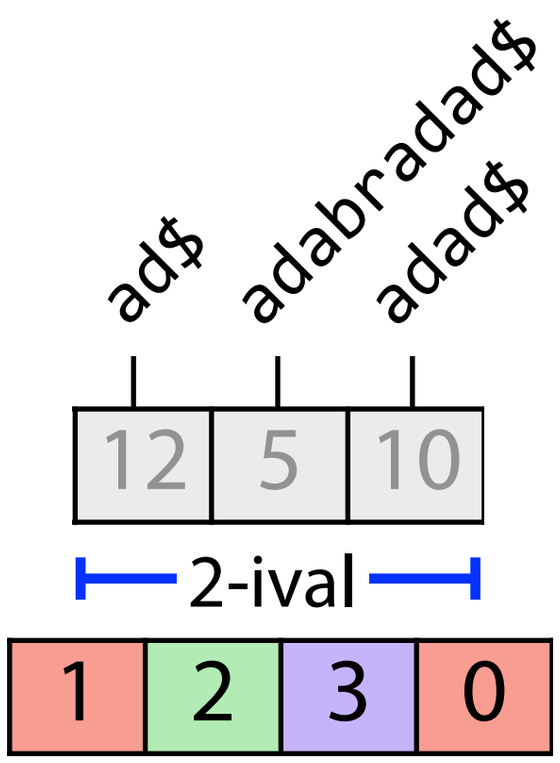
Suffix array

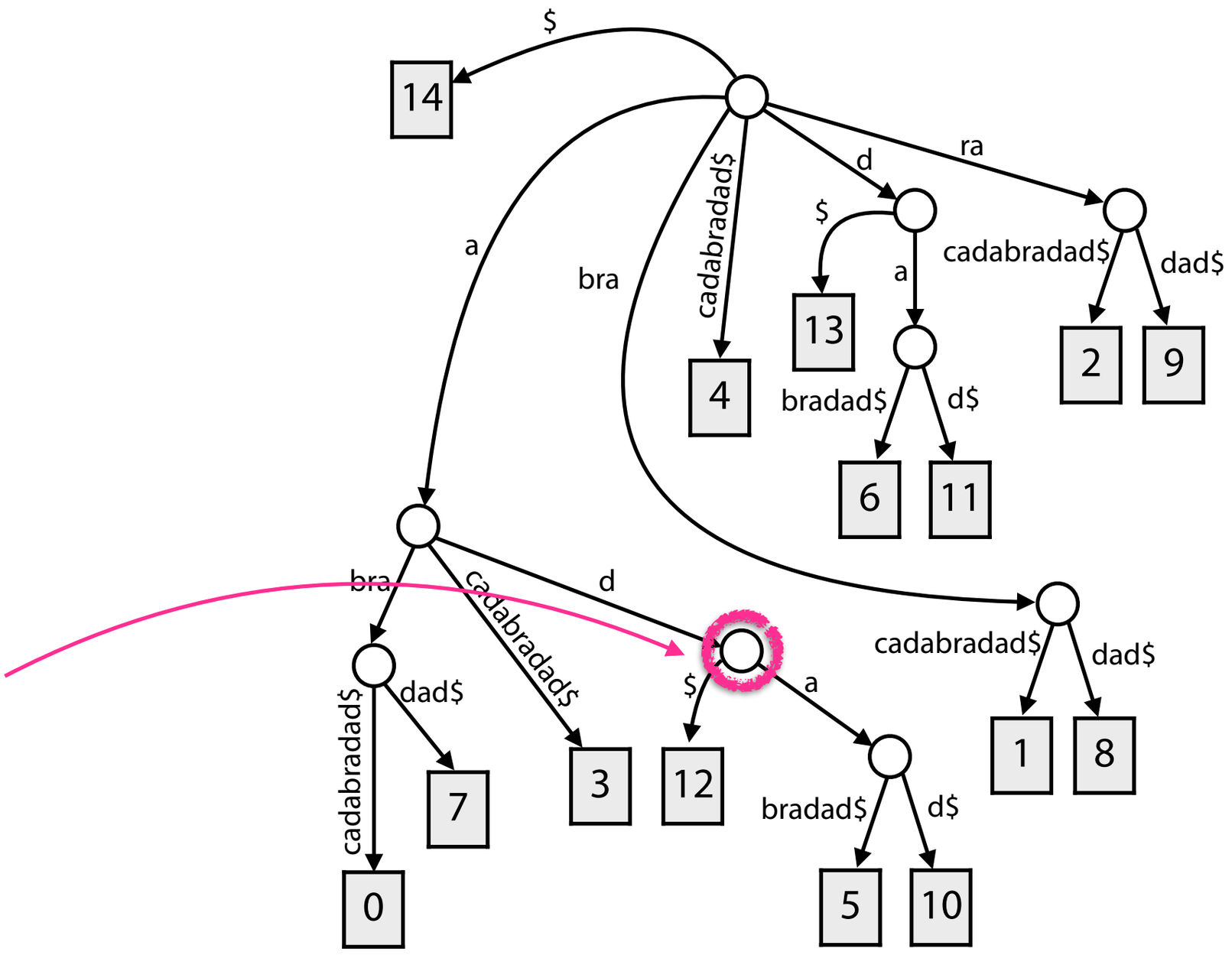


Which node is this?

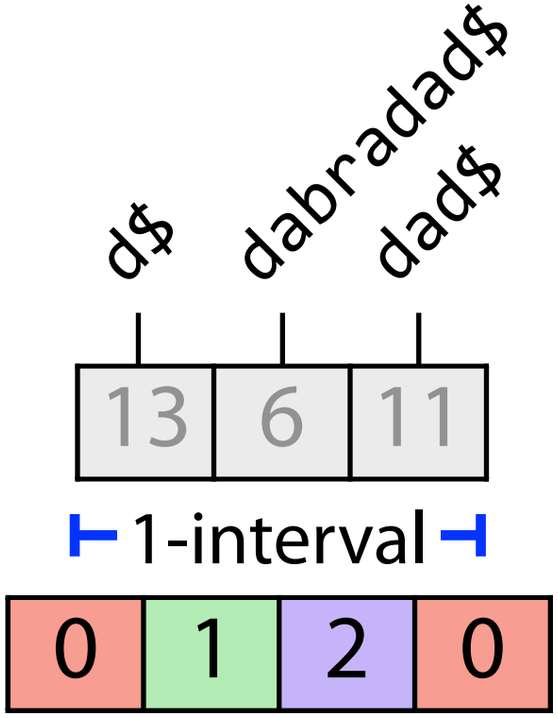


Suffix array

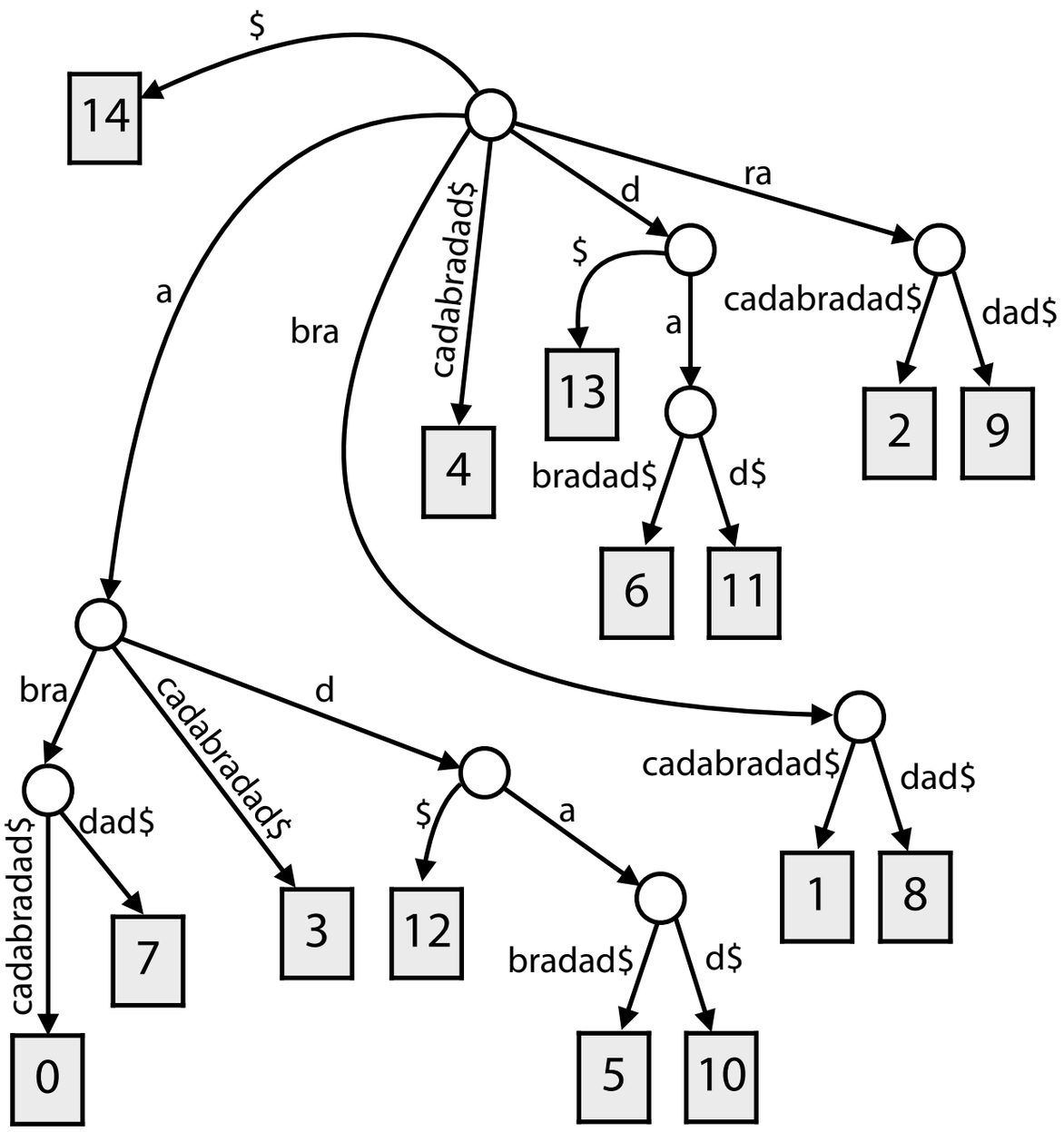




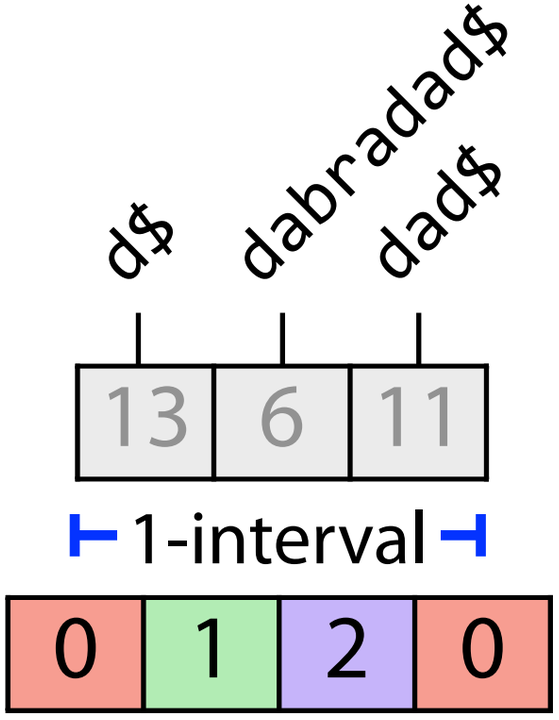
Suffix array

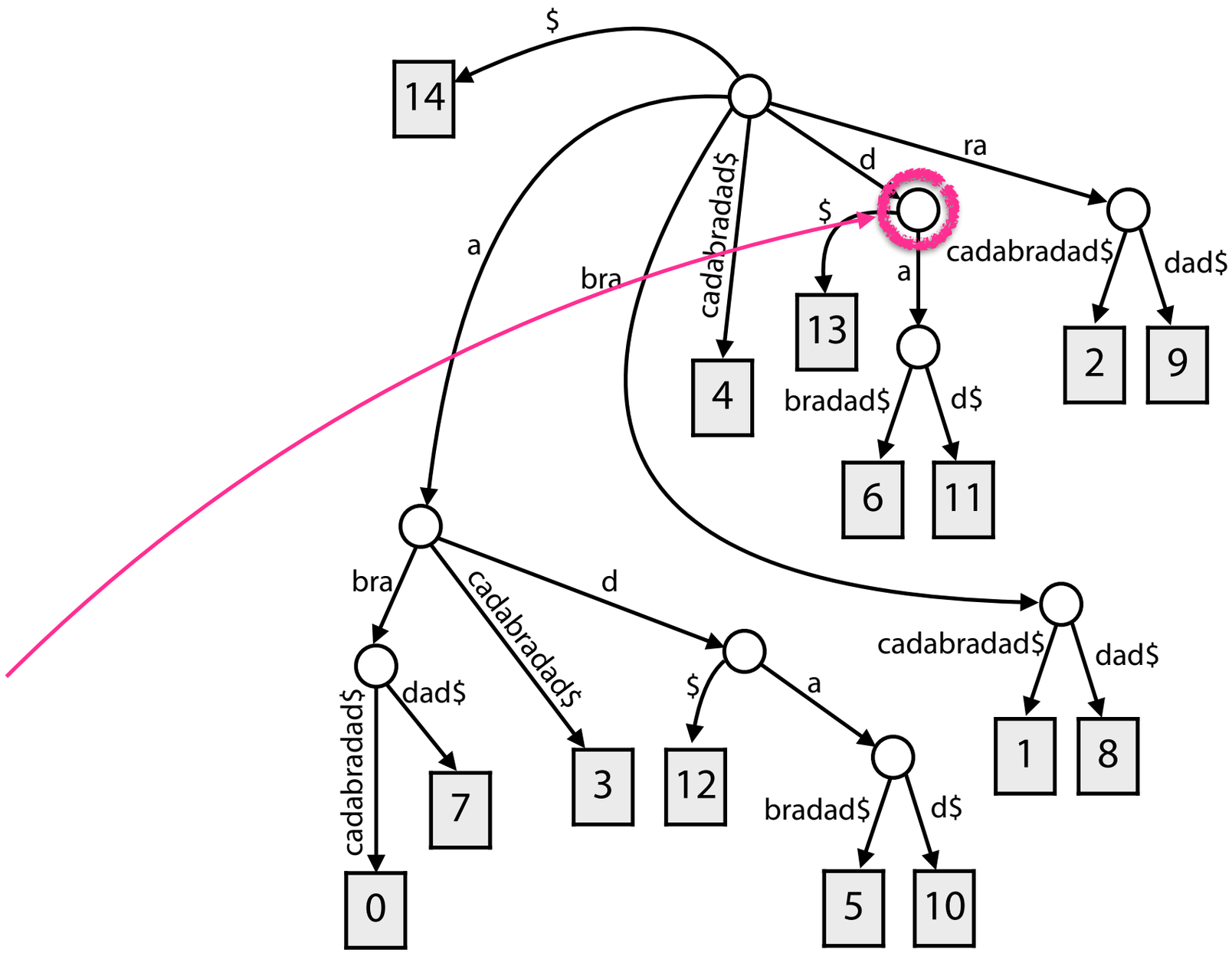


Which node is this?



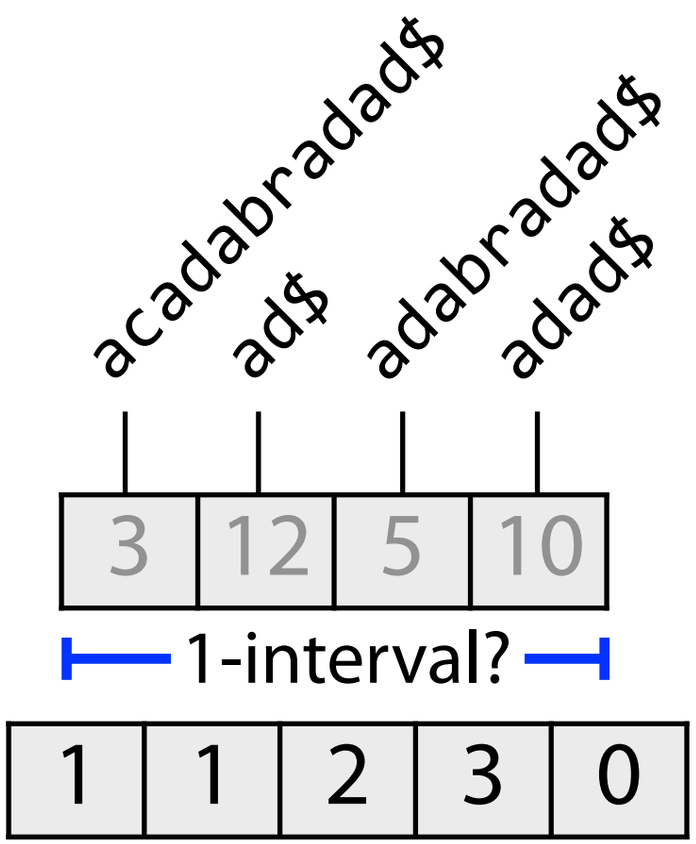
Suffix array

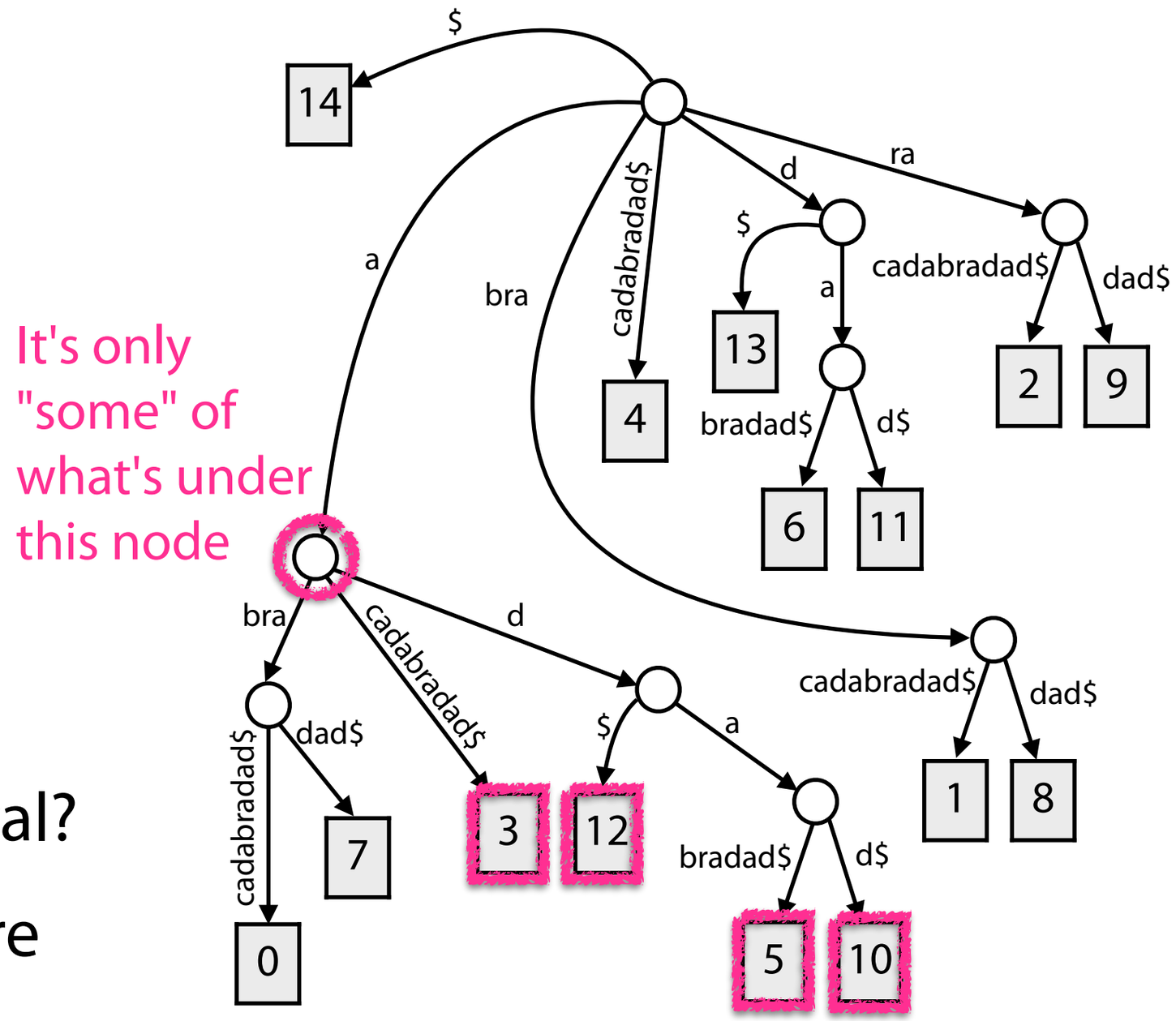




1-intervals are at (label) depth of 1, 2-intervals at depth of 2, etc

Suffix array





Why is this *not* a 1-interval?

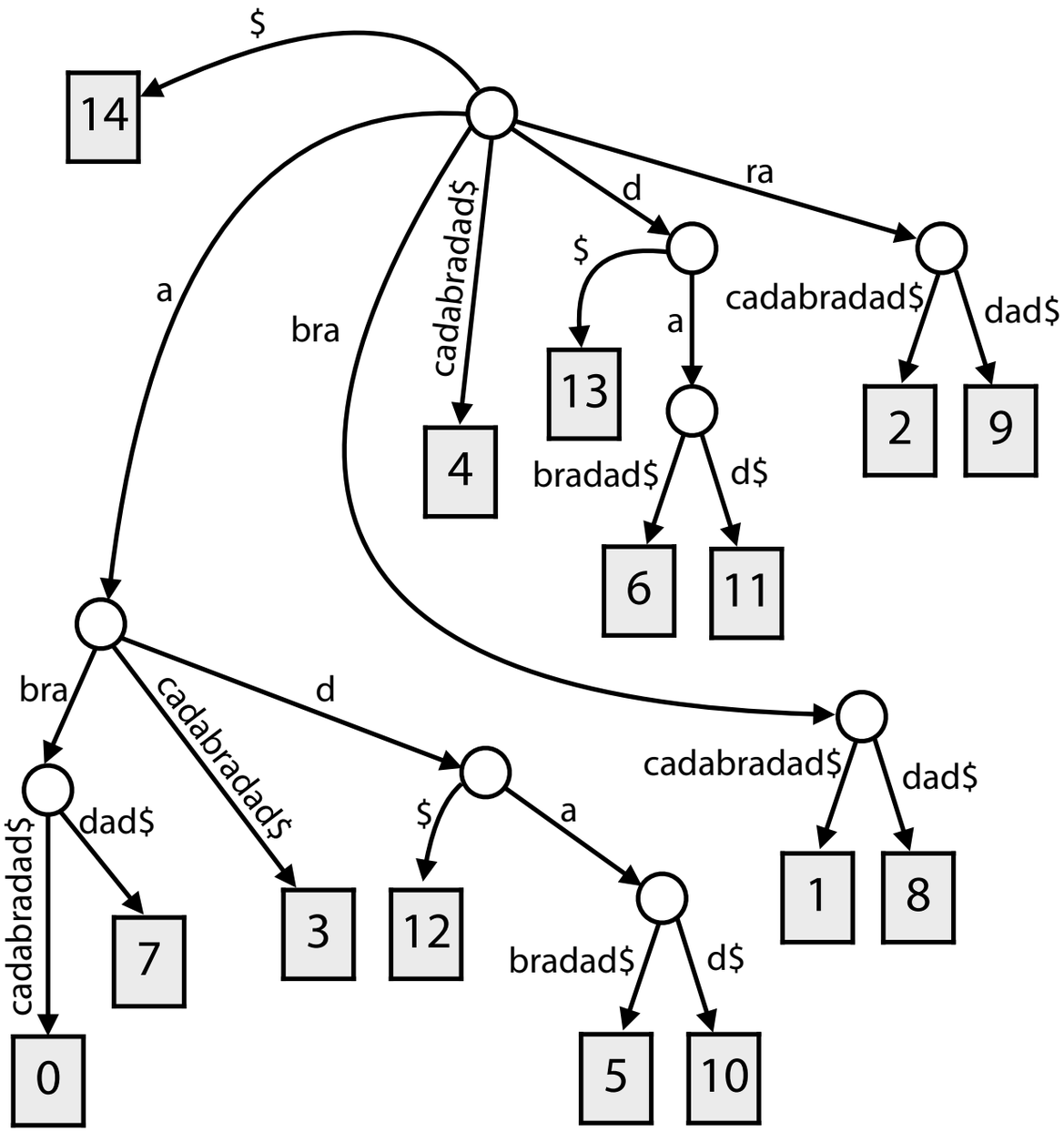
1. LCEs to either side are not both <1

2. It's not an internal node!

Suffix array

What is the "meaning" of the LCEs that are $= \ell$?

-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

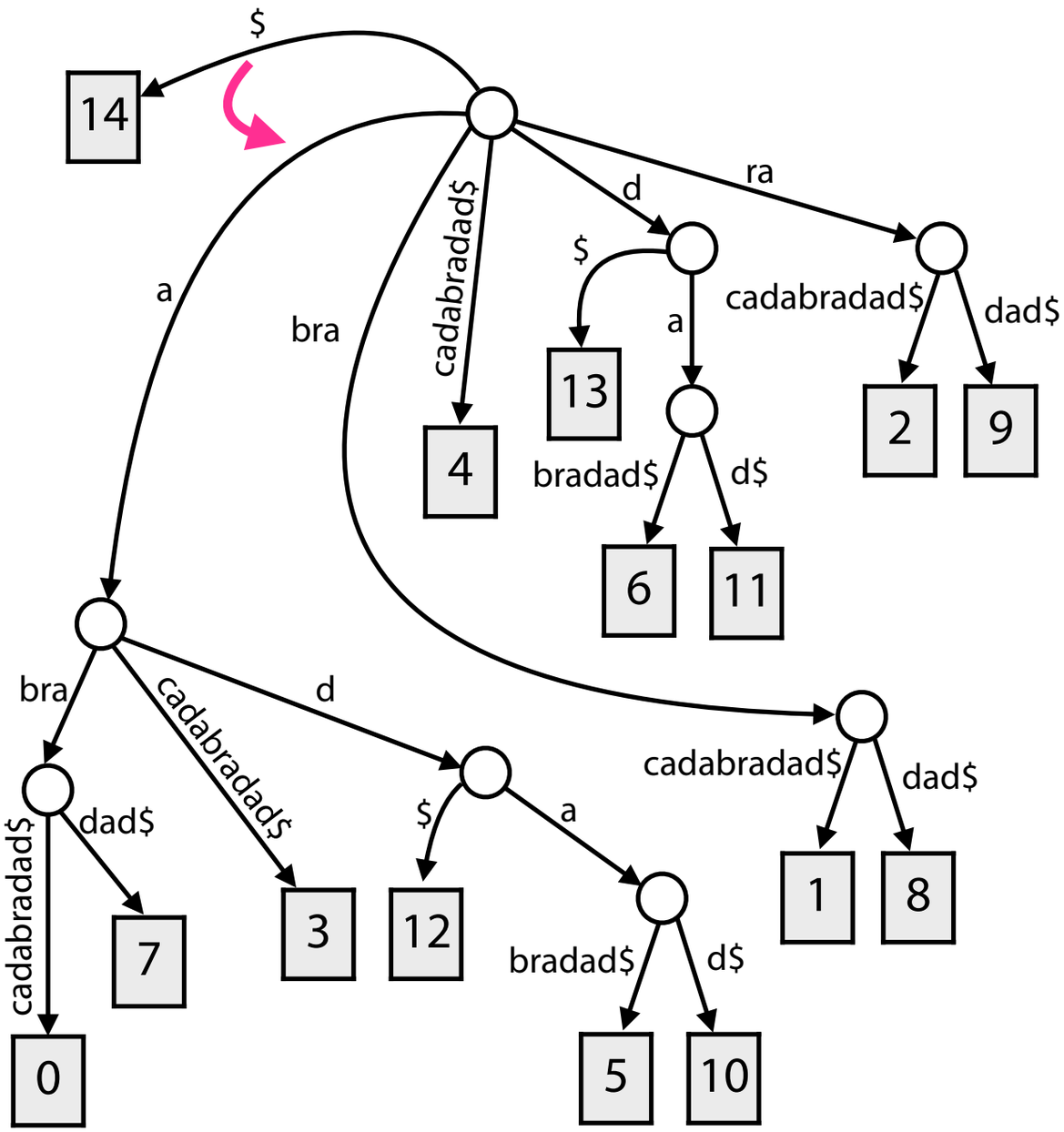


Correspond to "turnovers" from child edge to child edge

Suffix array

What is the "meaning" of the LCEs that are $= \ell$?

-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

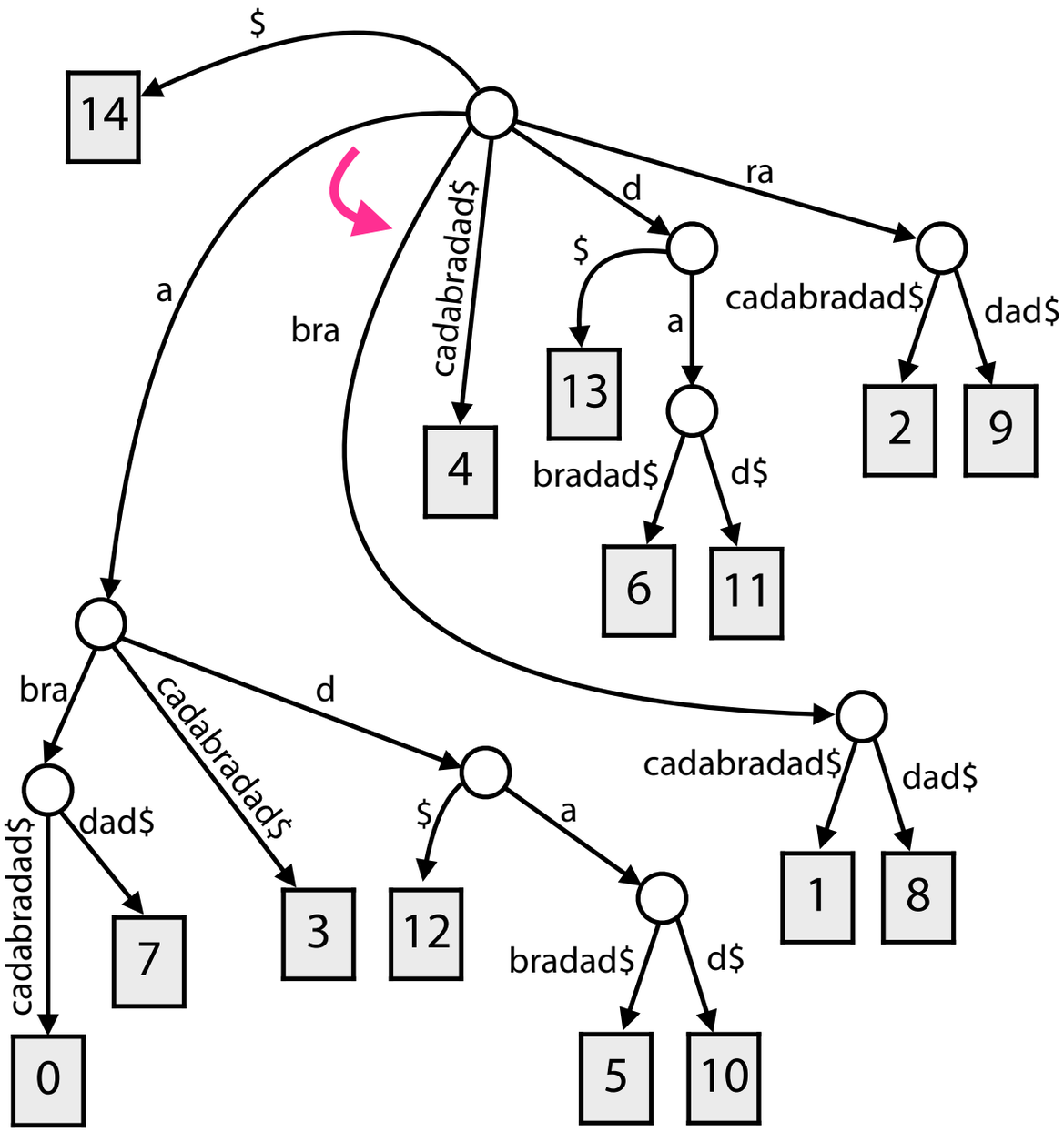


Correspond to "turnovers" from child edge to child edge

Suffix array

What is the "meaning" of the LCEs that are $= \ell$?

-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

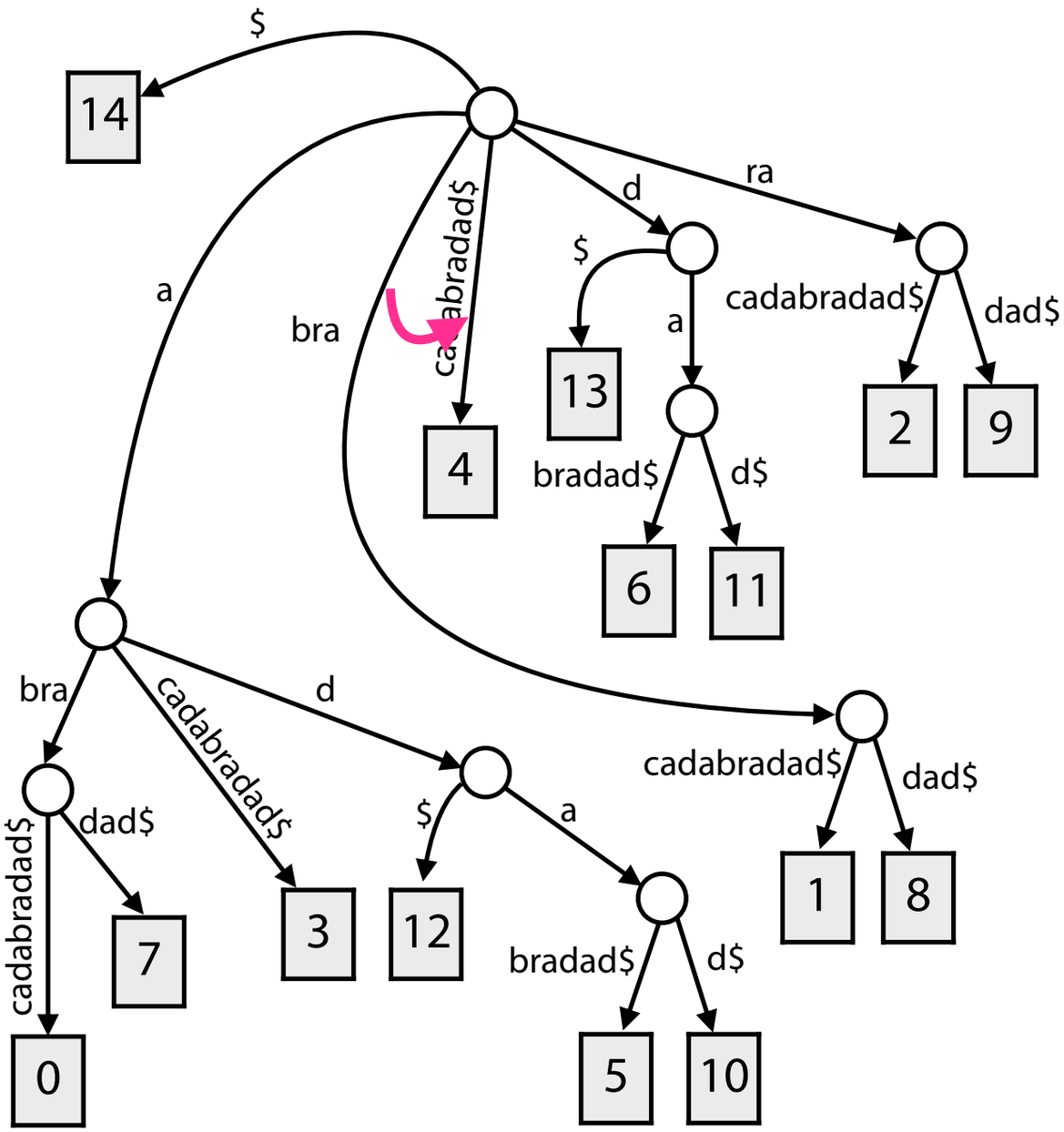


Correspond to "turnovers" from child edge to child edge

Suffix array

What is the "meaning" of the LCEs that are $= \ell$?

-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

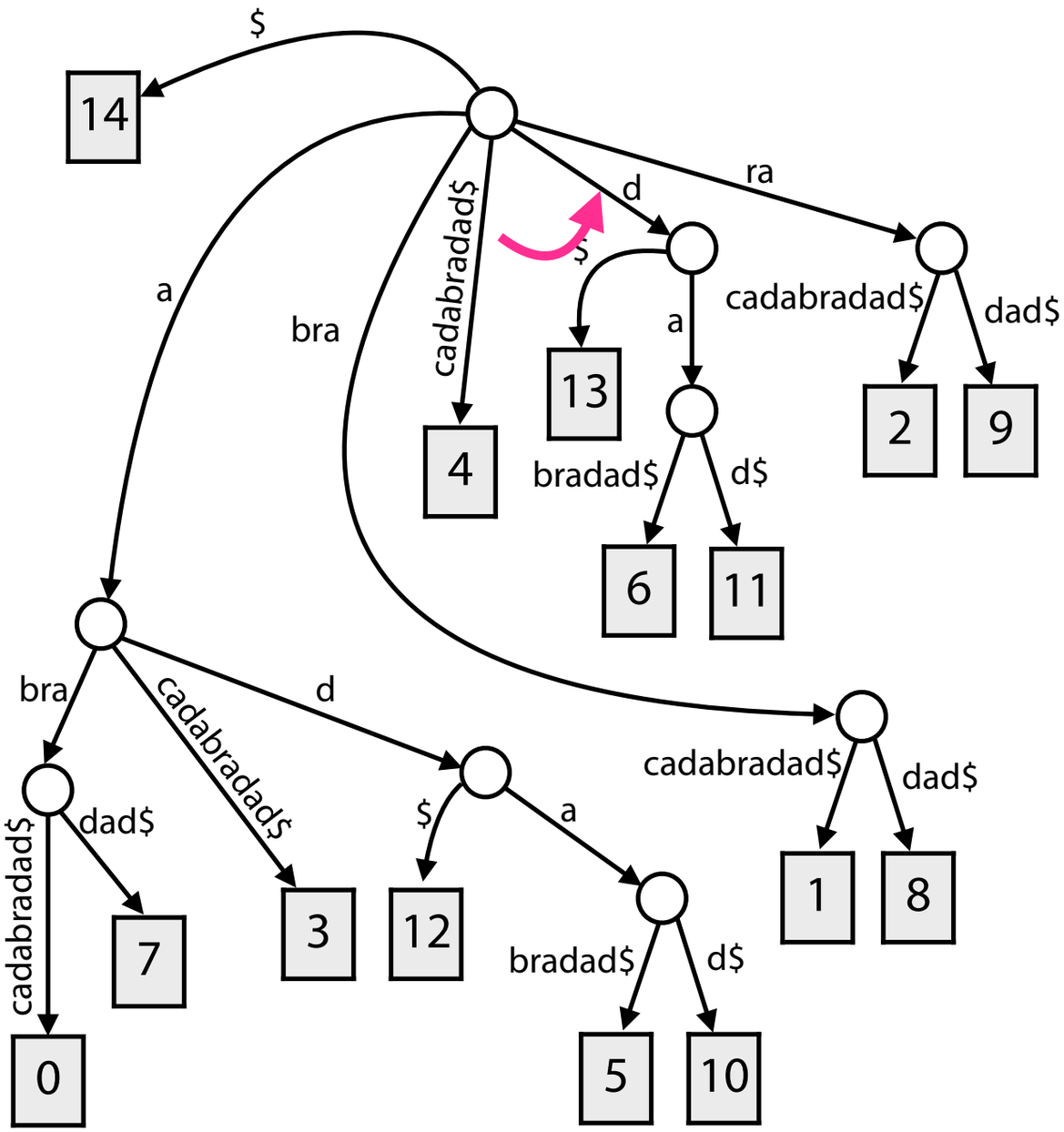


Correspond to "turnovers" from child edge to child edge

Suffix array

What is the "meaning" of the LCEs that are $= \ell$?

-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

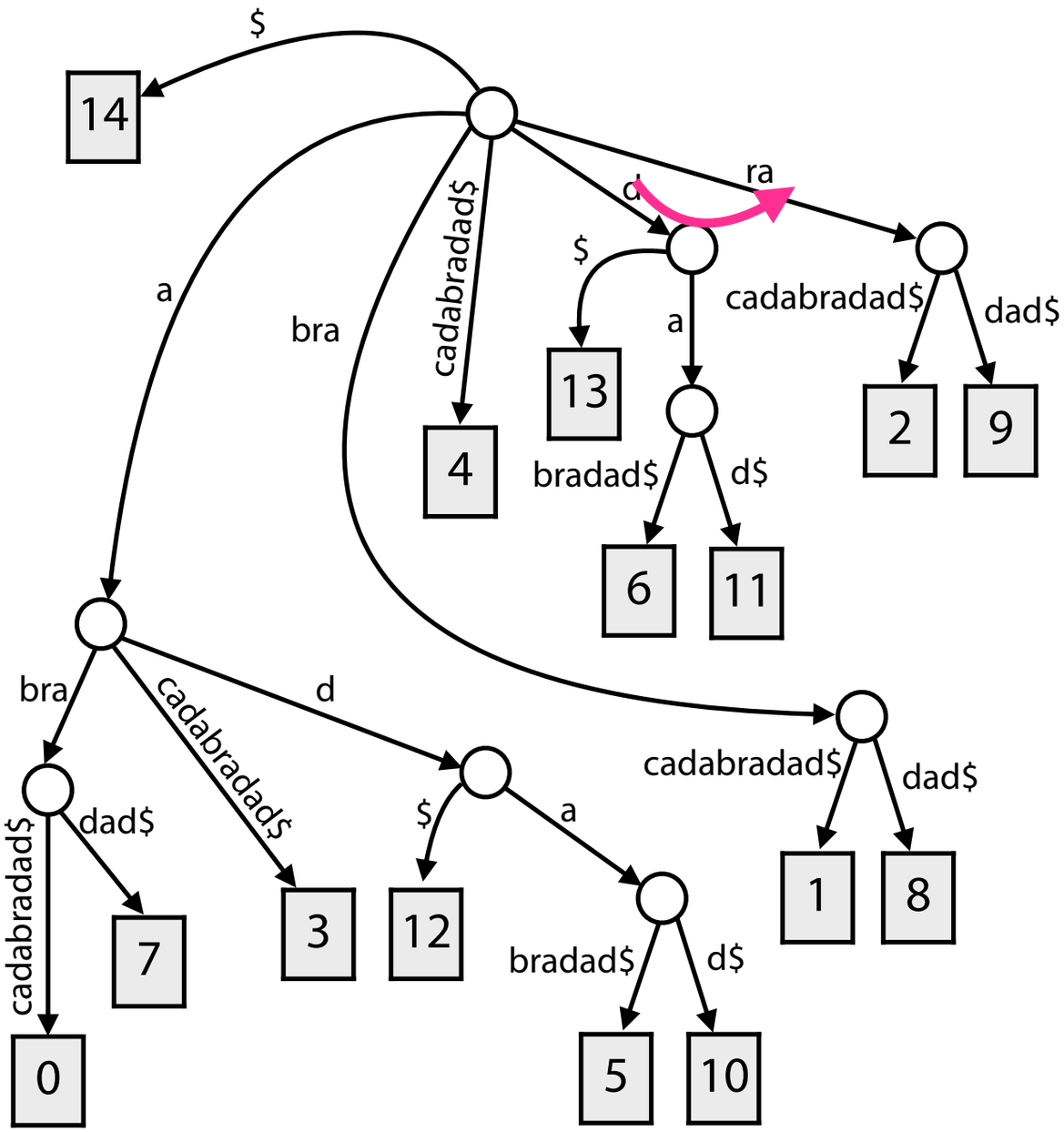


Correspond to "turnovers" from child edge to child edge

Suffix array

What is the "meaning" of the LCEs that are $= \ell$?

-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1



Correspond to "turnovers" from child edge to child edge

Suffix array

ℓ-intervals correspond to internal nodes

LCEs $= \ell$ in an ℓ-interval correspond to child "turnovers"

...so quickly finding $= \ell$ LCEs allows us to quickly find child ℓ-intervals. We can **traverse the tree**!

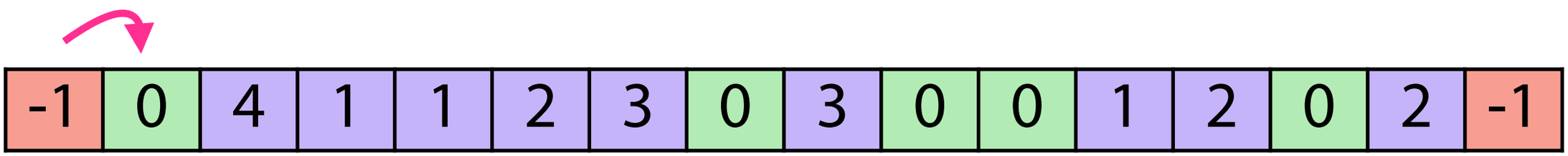
-1	0	4	1	1	2	3	0	3	0	0	1	2	0	2	-1

Suffix array

ℓ-intervals correspond to internal nodes

LCEs $= \ell$ in an ℓ-interval correspond to child "turnovers"

...so quickly finding $= \ell$ LCEs allows us to quickly find child ℓ-intervals. We can **traverse the tree**!

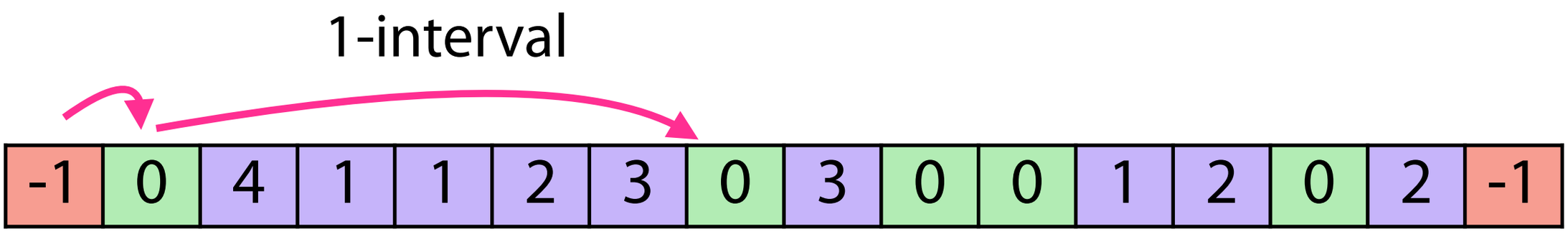


Suffix array

ℓ-intervals correspond to internal nodes

LCEs $= \ell$ in an ℓ-interval correspond to child "turnovers"

...so quickly finding $= \ell$ LCEs allows us to quickly find child ℓ-intervals. We can **traverse the tree**!

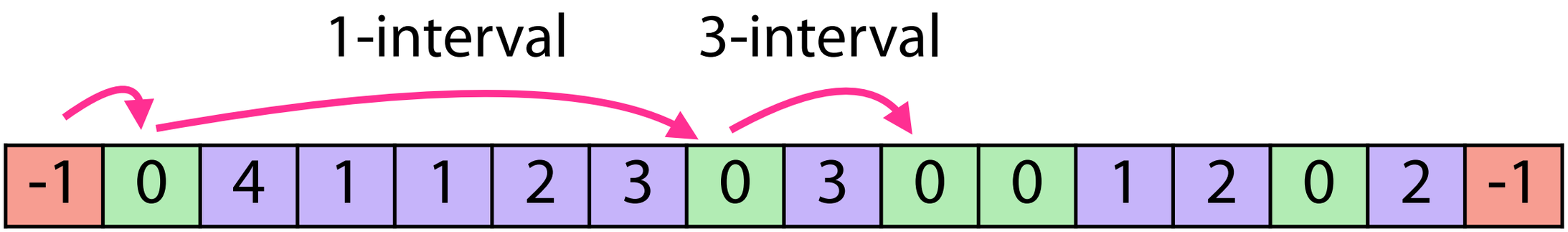


Suffix array

ℓ-intervals correspond to internal nodes

LCEs $= \ell$ in an ℓ-interval correspond to child "turnovers"

...so quickly finding $= \ell$ LCEs allows us to quickly find child ℓ-intervals. We can **traverse the tree**!

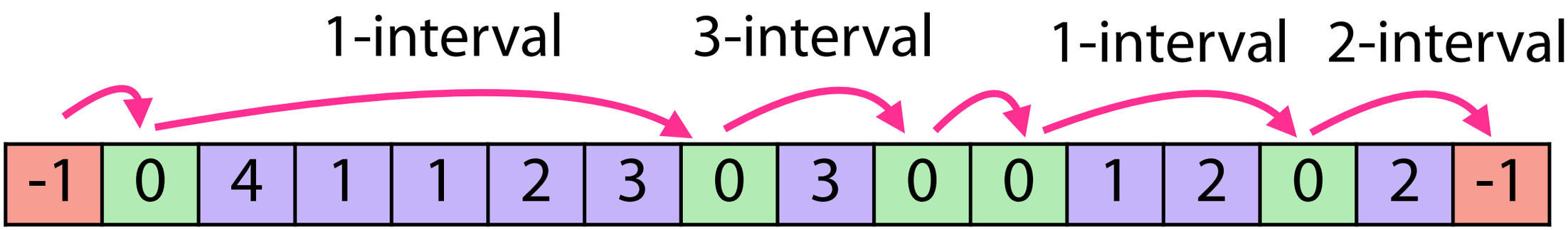


Suffix array

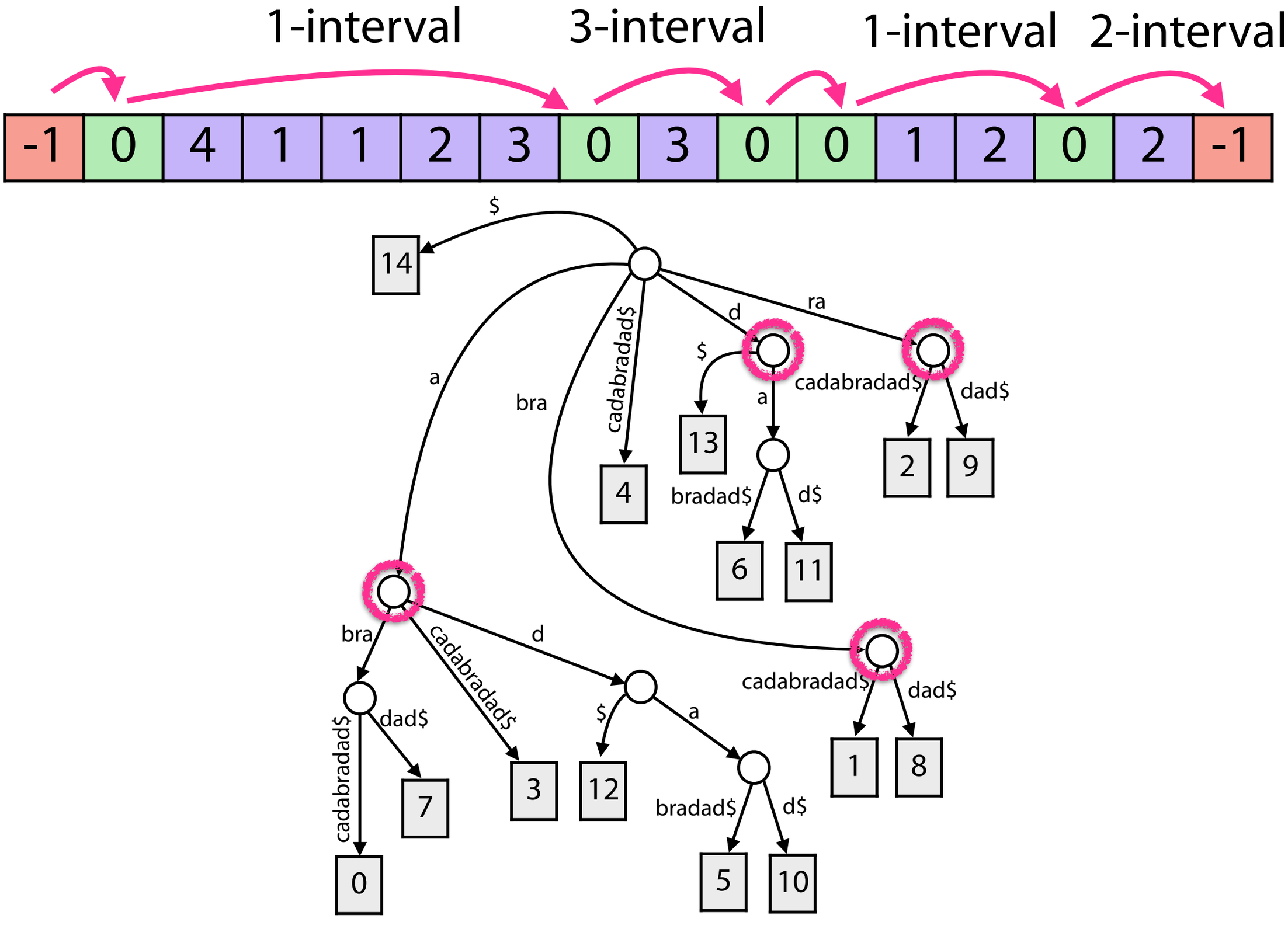
ℓ-intervals correspond to internal nodes

LCEs $= \ell$ in an ℓ-interval correspond to child "turnovers"

...so quickly finding $= \ell$ LCEs allows us to quickly find child ℓ-intervals. We can **traverse the tree**!

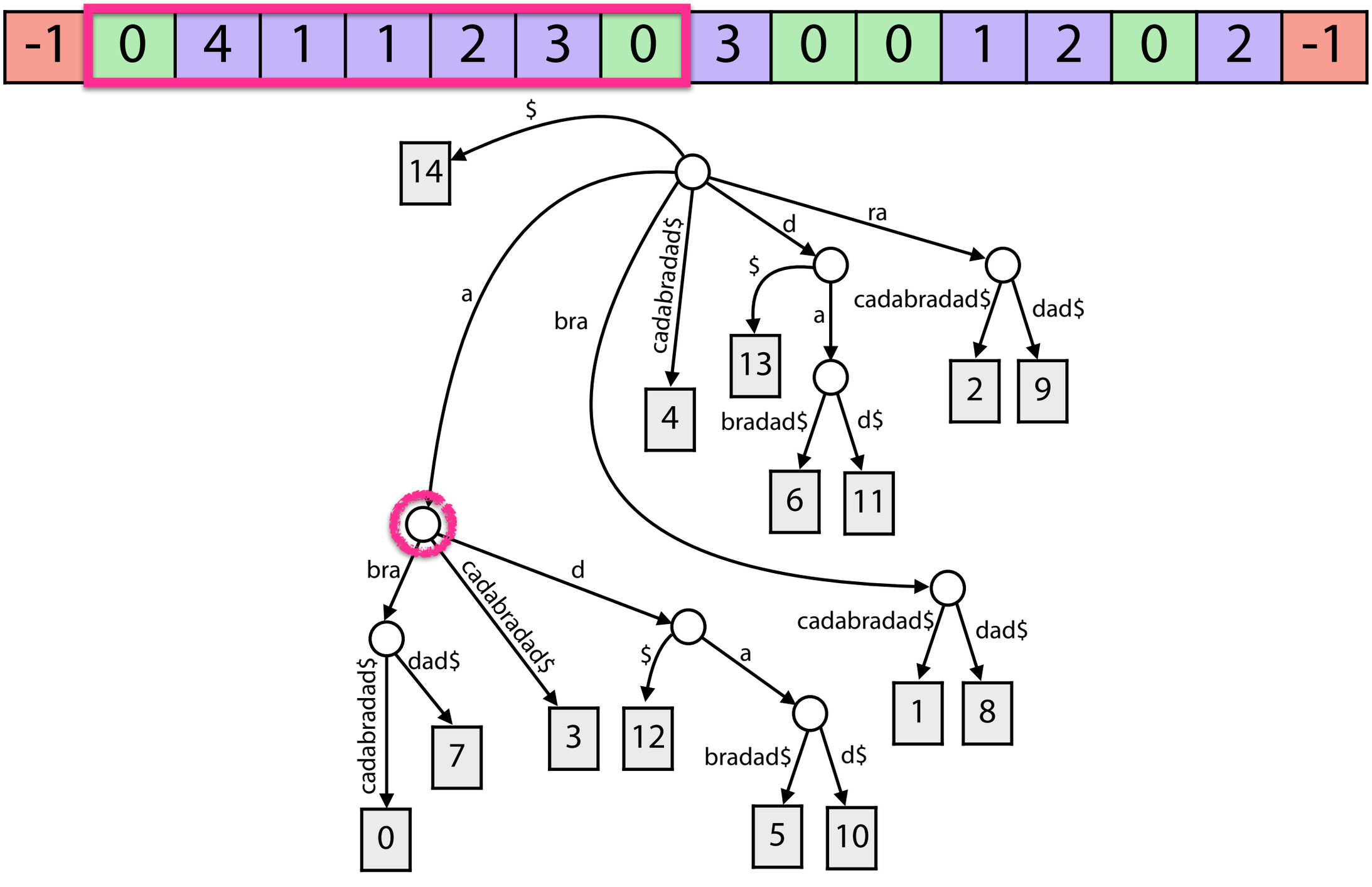


Suffix array

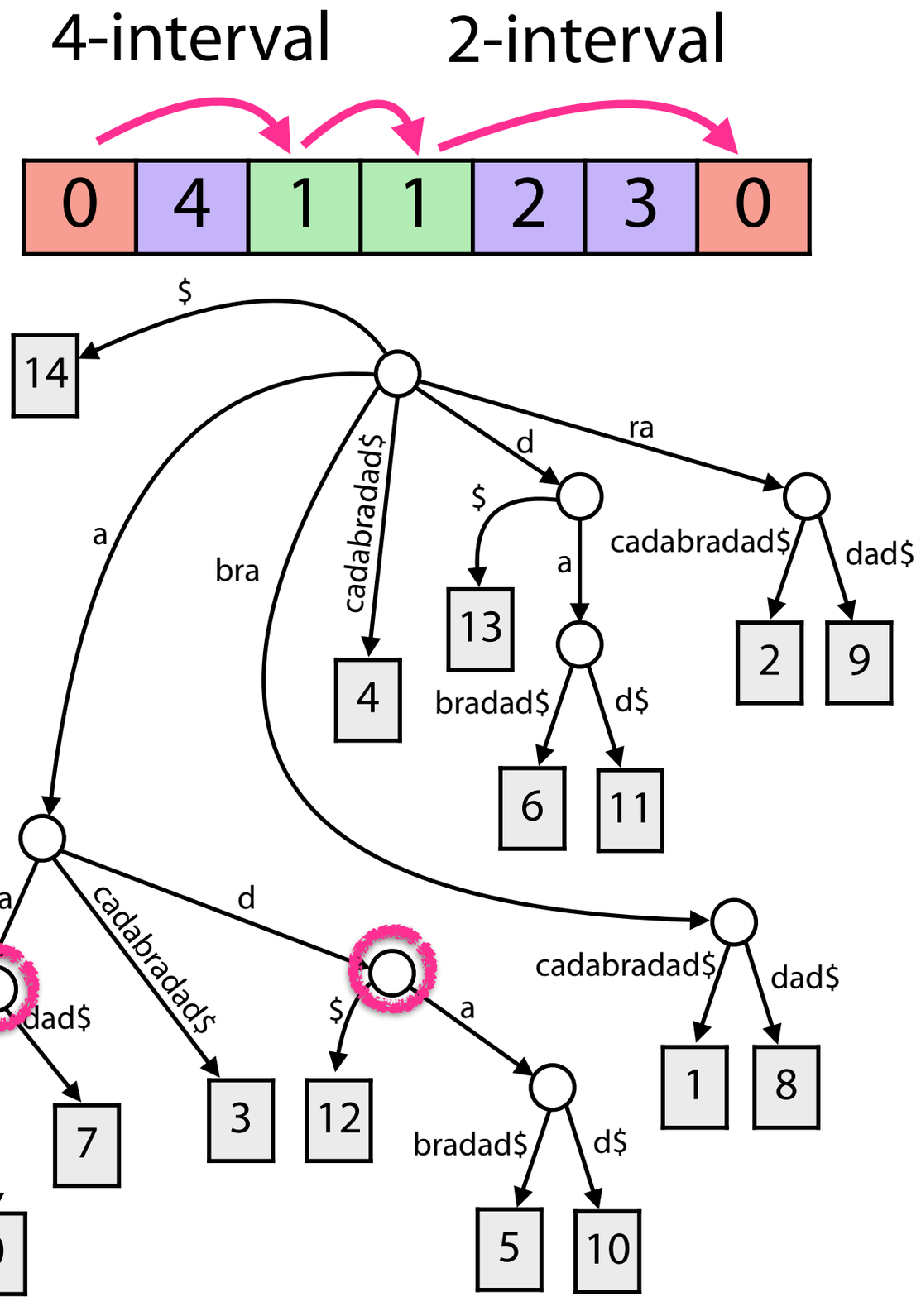


Suffix array

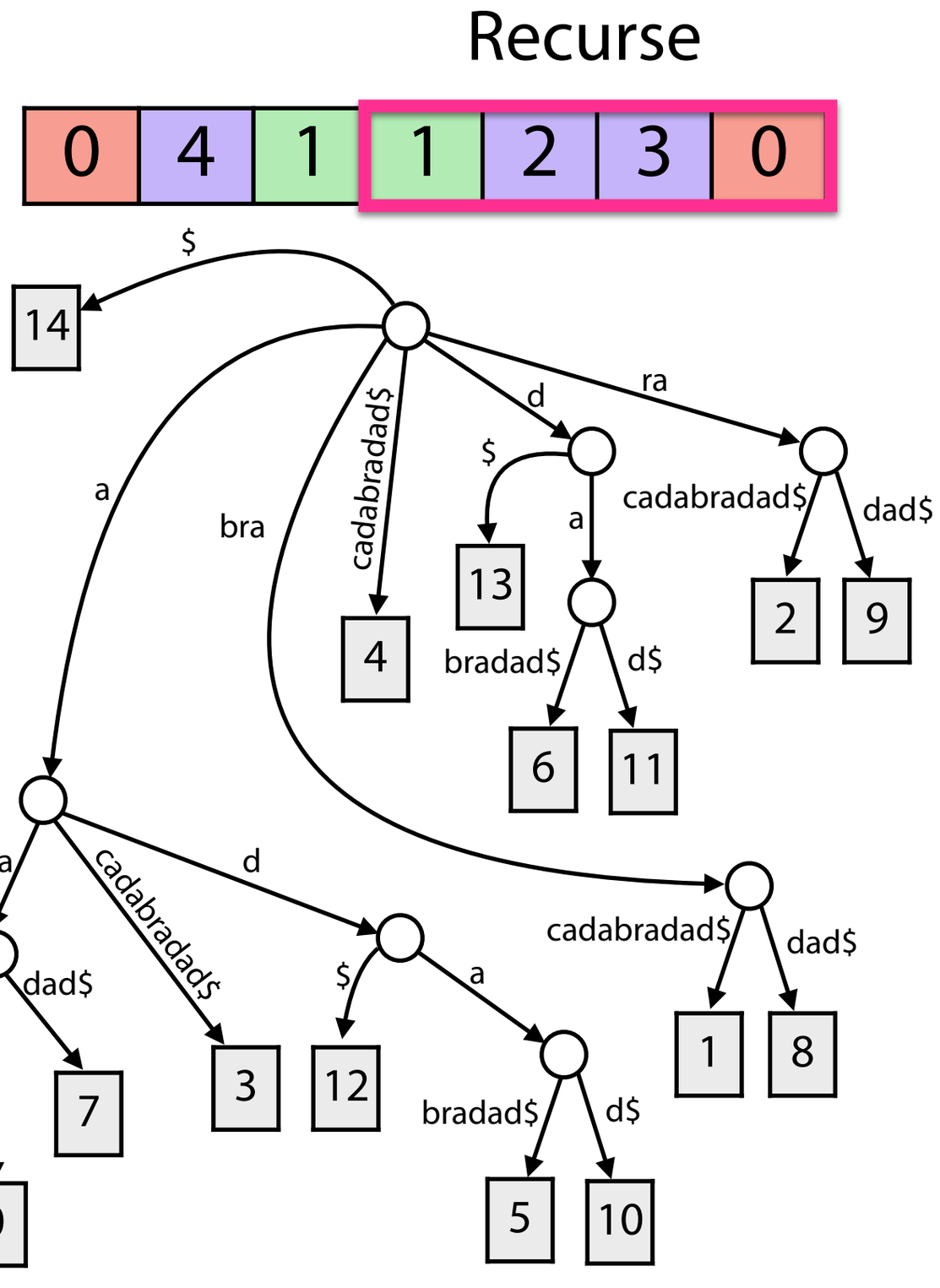
Recurse



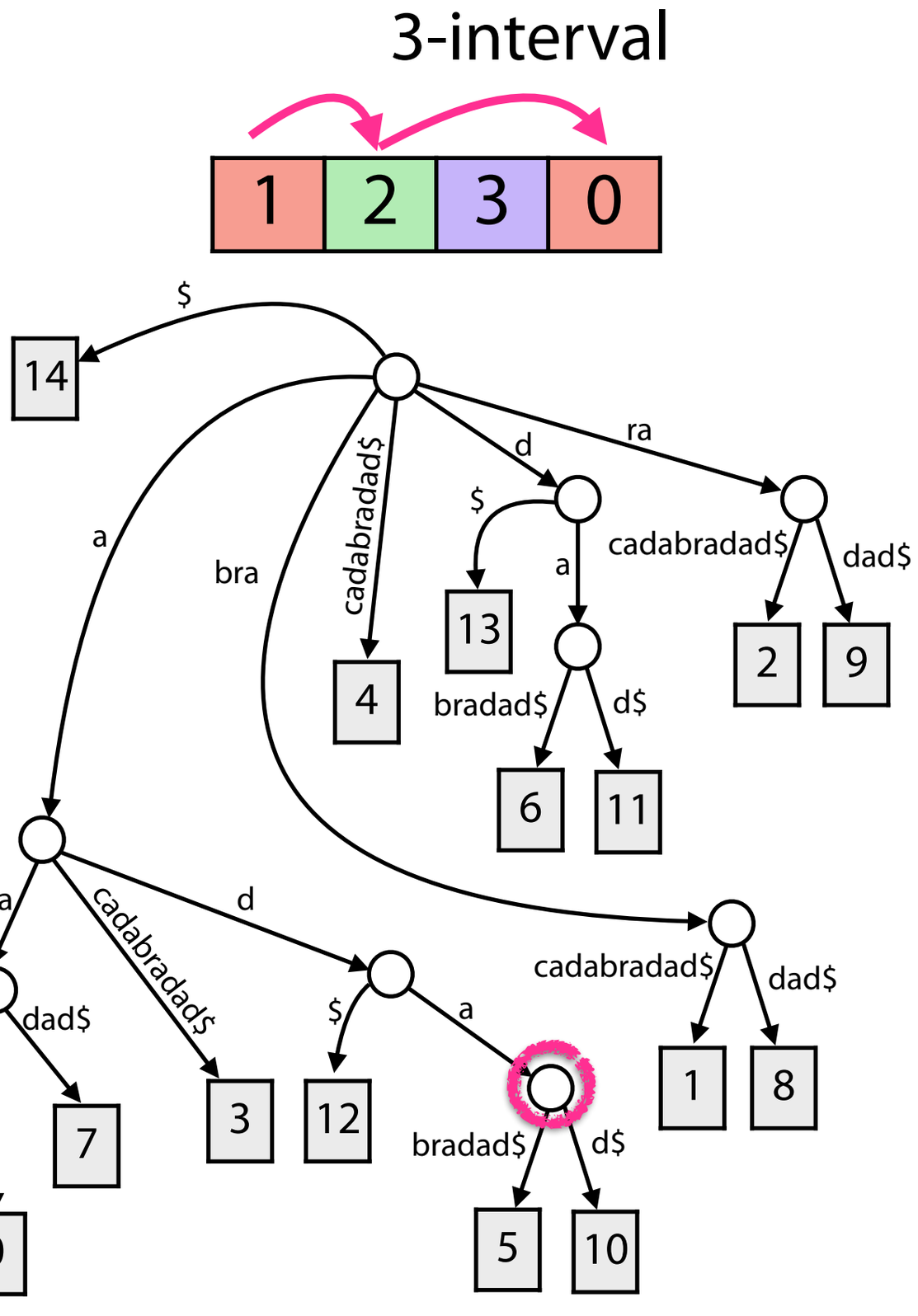
Suffix array



Suffix array

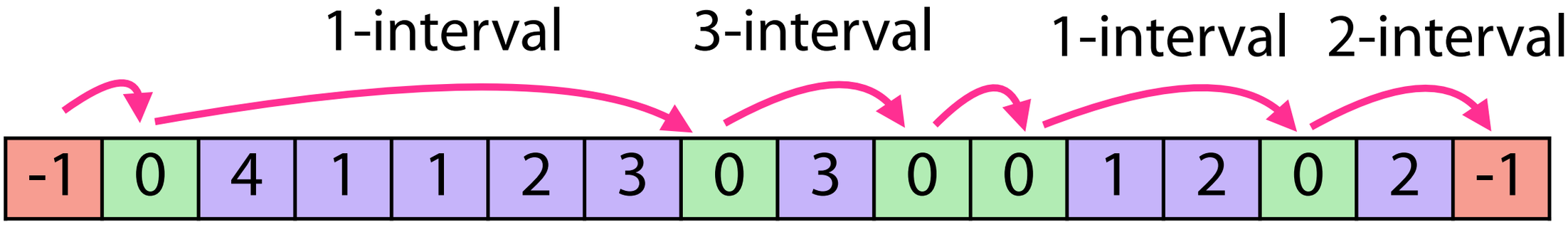


Suffix array



Suffix array

How to accomplish fast jumping between $= \ell$ LCEs?



Pre-compute

Abouelhoda, Mohamed Ibrahim, Stefan Kurtz, and Enno Ohlebusch. "Replacing suffix trees with enhanced suffix arrays." *Journal of discrete algorithms* 2.1 (2004): 53-86.

Rank minimum queries
+
Super cartesian trees

Ohlebusch, Enno, and Simon Gog. "A compressed enhanced suffix array supporting fast string matching." *International Symposium on String Processing and Information Retrieval*. Springer, Berlin, Heidelberg, 2009.