

Suffix Trees: matching statistics

Ben Langmead



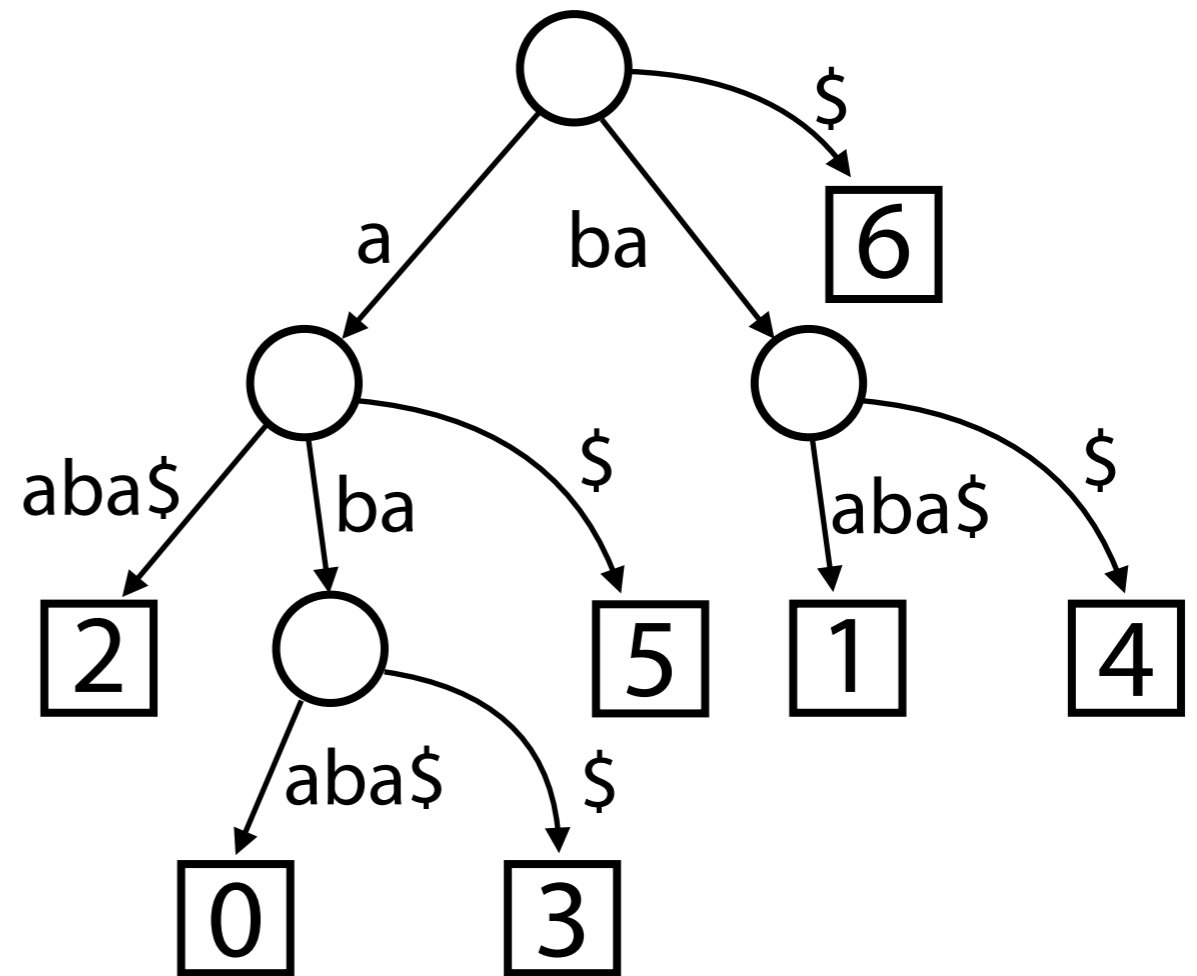
Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).

Matching statistics

We can describe similar substrings between pattern & text with **matching statistics**

At step i compute the length of the longest prefix of the suffix $P[i :]$ that occurs in T

...using suffix links!



Matching statistics

First an example without the tree.

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :														

$i = 1$ ↑

Length of the longest prefix of suffix $P[i :]$ that occurs in T

Matching statistics

First an example without the tree.

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2													

$i = 1$ ↑

Length of the longest prefix of suffix $P[i :]$ that occurs in T

Matching statistics

First an example without the tree.

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2													

$i = 2$ ↑

Length of the longest prefix of suffix $P[i :]$ that occurs in T

Matching statistics

First an example without the tree.

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2	1												

$i = 2$ ↑

Length of the longest prefix of suffix $P[i :]$ that occurs in T

Matching statistics

First an example without the tree.

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2	1												

$i = 3$ ↑

Length of the longest prefix of suffix $P[i :]$ that occurs in T

Matching statistics

First an example without the tree.

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2	1	5											

$i = 3$ ↑

Length of the longest prefix of suffix $P[i :]$ that occurs in T

Matching statistics

First an example without the tree.

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2	1	5											

$i = 4$ ↑

Length of the longest prefix of suffix $P[i :]$ that occurs in T

Matching statistics

First an example without the tree.

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2	1	5	4										

$i = 4$

Length of the longest prefix of suffix $P[i :]$ that occurs in T

Matching statistics

Let's fill in the rest:

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2	1	5	4	3	5	4	3	2	1	1			

Matching statistics

T :	a	b	r	a	c	a	d	a	b	r	a	d	a	d
P :	c	a	r	a	d	a	d	a	b	r	d			
MS :	2	1	5	4	3	5	4	3	2	1	1			

Matching statistics

T : a|b|r|a|c|a|d|a|b|r|a|d|a|d
P : c|a|r|a|d|a|d|a|b|r|d
MS : 2|1|5|4|3|5|4|3|2|1|1

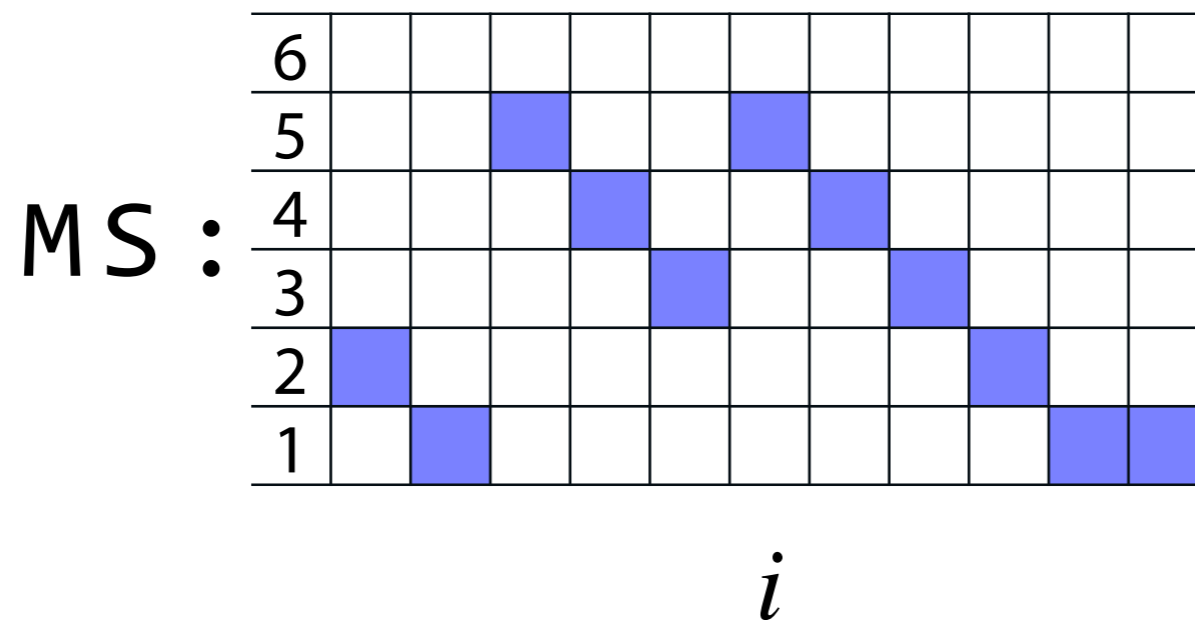
MS :

6											
5											
4											
3											
2											
1											

i

Matching statistics

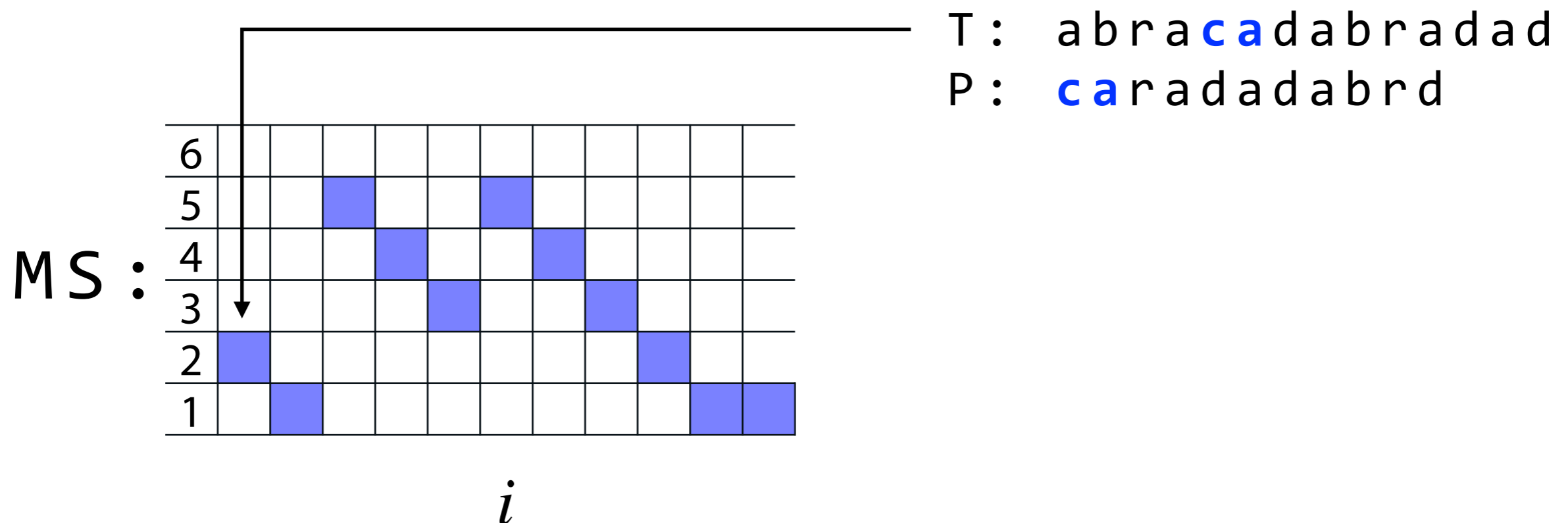
T : a b r a c a d a b r a d a d
P : c a r a d a d a b r d
MS : 2 1 5 4 3 5 4 3 2 1 1



Matching statistics

A "peak" in the matching statistics corresponds to a **Maximal** Exact Match (MEM)

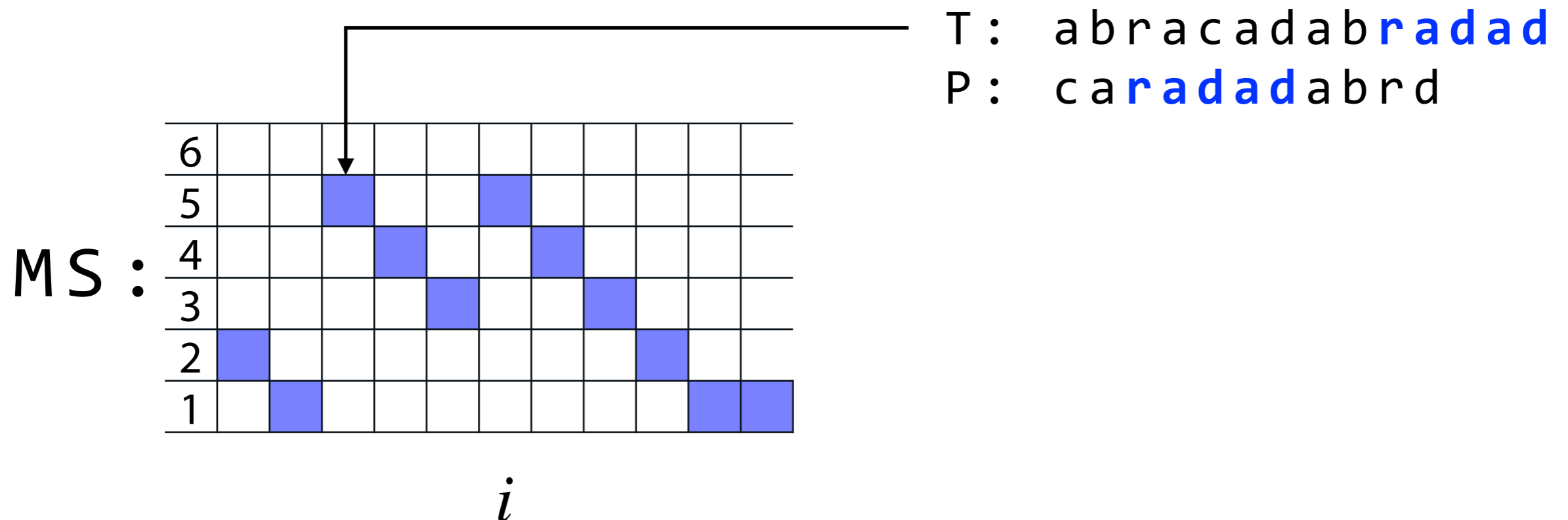
Maximal: can't be extended in either direction without causing a mismatch



Matching statistics

A "peak" in the matching statistics corresponds to a **Maximal** Exact Match (MEM)

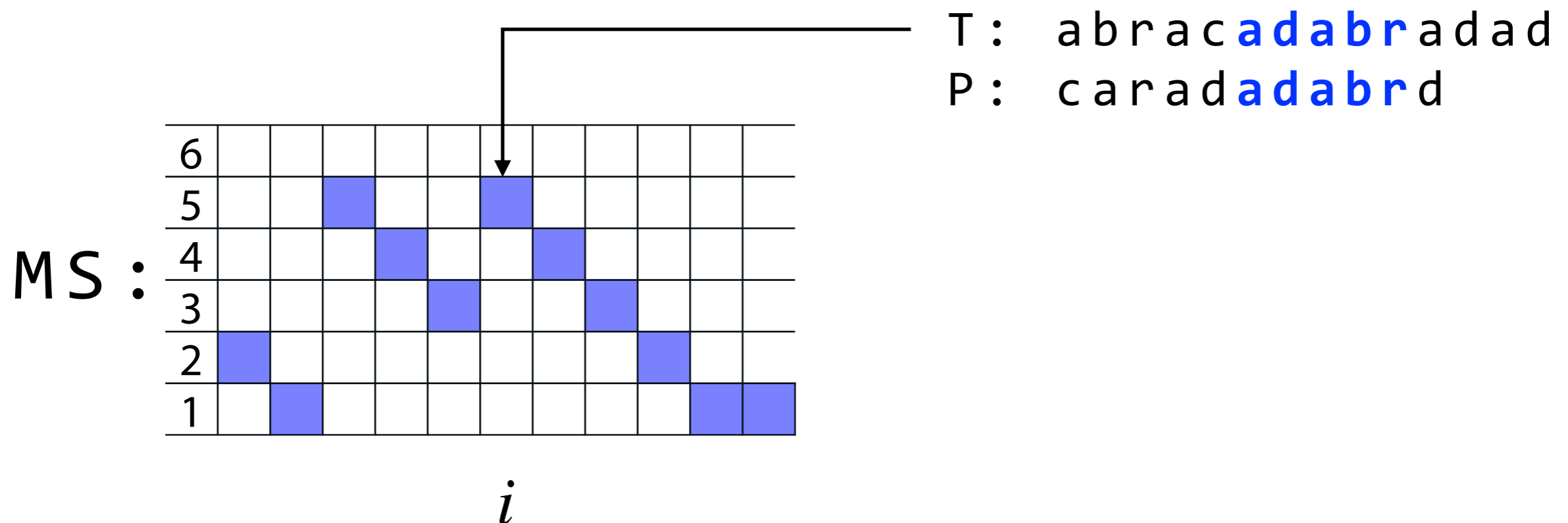
Maximal: can't be extended in either direction without causing a mismatch



Matching statistics

A "peak" in the matching statistics corresponds to a **Maximal** Exact Match (MEM)

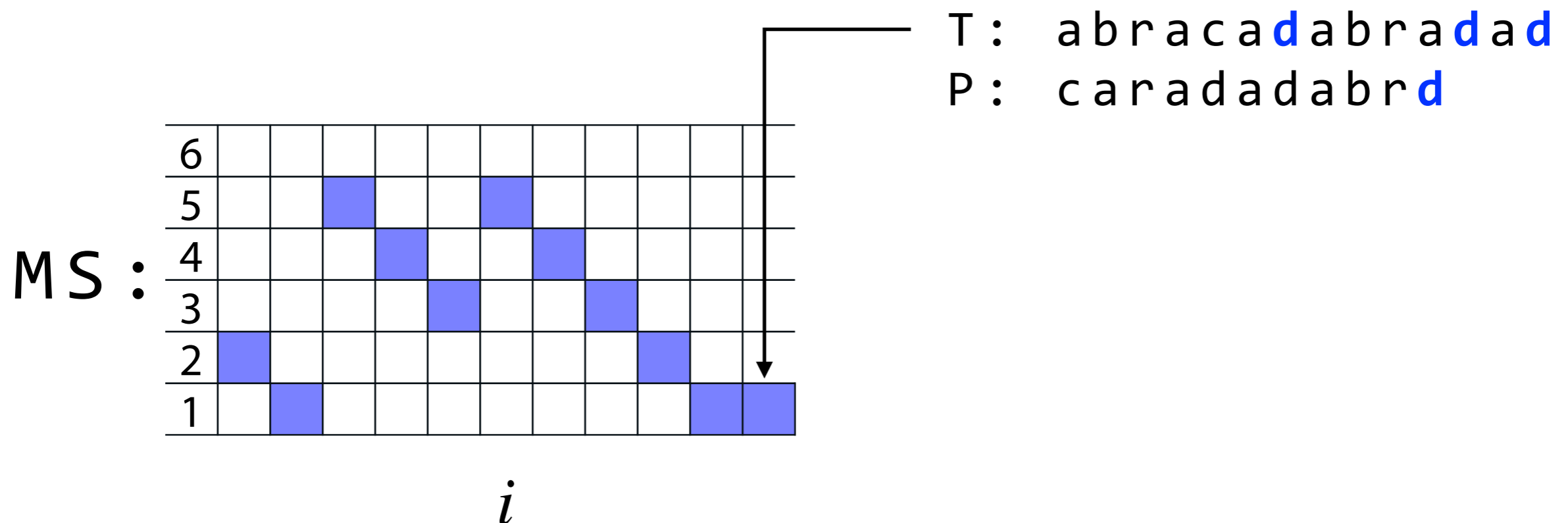
Maximal: can't be extended in either direction without causing a mismatch



Matching statistics

A "peak" in the matching statistics corresponds to a **Maximal** Exact Match (MEM)

Maximal: can't be extended in either direction without causing a mismatch



Matching statistics: summary

A way to describe how well substrings of the pattern match substrings of the text

Don't need to pick a substring length ahead of time;
MSs are "maximal" in the direction of matching

MS "peaks" are Maximal Exact Matches (MEMs)

Basic tool for whole-genome alignment, read alignment (in genomics), approximate matching in general

Next: what's the algorithm?