String Matching

Finding all occurrences of a pattern in the text.

Applications

- Spell Checkers.
- Search Engines.
- Spam Filters.
- Intrusion Detection System.
- Plagiarism Detection.
- Bioinformatics – DNA Sequencing.
- Digital Forensics.
- Information Retrieval, etc.

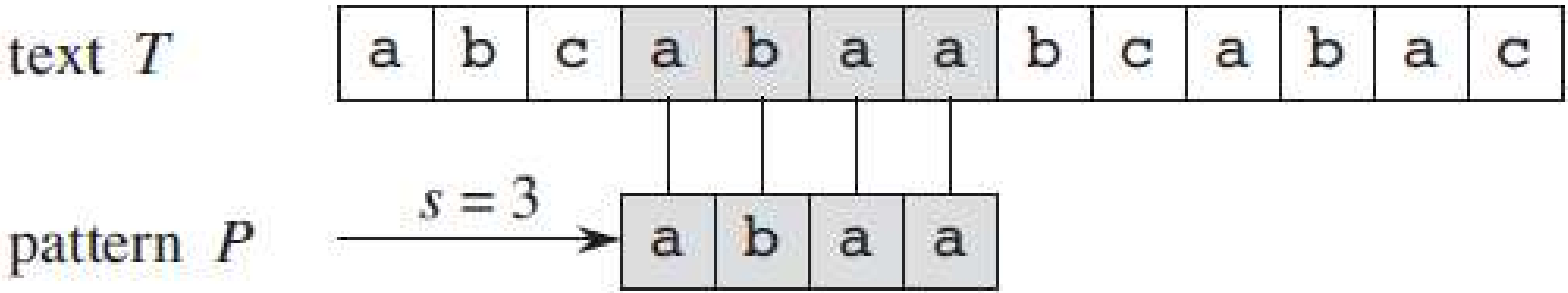
String Matching Problem

- Let,
- Text $T[1..n]$ is an array of length n.
- Pattern $P[1..m]$ is an array of length $m \leq n$.
- P and T are drawn from a finite alphabet Σ.
  - $\Sigma = \{0,1\}$ or $\Sigma = \{a, b, \ldots, z\}$.
- Example:
  - T = a b c a b a a b c a b a c
  - P = a b a a

Contd…

		1	2	3	4	5	6	7	8	9	10	11	12	13
	T	a	b	c	a	b	a	a	b	c	a	b	a	c
shift = 0 ✗	*P*	a	b	a	a									
shift = 1 ✗	*P*		a	b	a	a								
shift = 2 ✗	*P*			a	b	a	a							
shift = 3 ✓	*P*				a	b	a	a						
shift = 4 ✗	*P*					a	b	a	a					
shift = 5 ✗	*P*						a	b	a	a				
shift = 6 ✗	*P*							a	b	a	a			
shift = 7 ✗	*P*								a	b	a	a		
shift = 8 ✗	*P*									a	b	a	a	
shift = 9 ✗	*P*										a	b	a	a

Contd…



- Shift s a valid shift, if P occurs with shift s in T.
  - $0 \le s \le n - m$ and $T[s + 1..s + m] = P[1..m]$.
- Otherwise, shift s is an invalid shift.
- String-matching problem means finding all valid shifts with which a given pattern P occurs in a given text T.

Naive String Matching Algorithm

NAIVE-STRING-MATCHER(T,P)

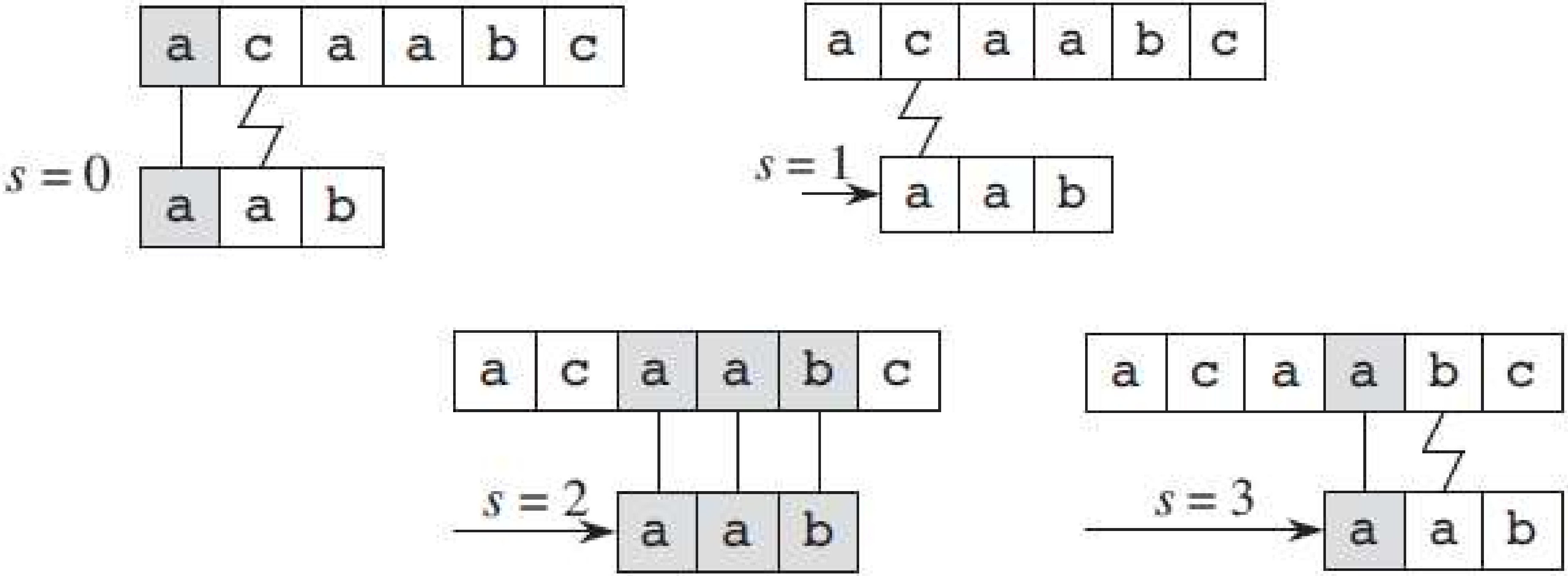
1. $n = T.length$
2. $m = P.length$
3. for $s = 0$ to $n - m$
4. if $P[1..m] == T[s + 1..s + m]$
5. print "Pattern occurs with shift" s

- Complexity: $O((n - m + 1)m)$

Example

```
NAIVE-STRING-MATCHER(T, P)
n = T.length
m = P.length
for s = 0 to n − m
    if P[1..m] == T[s + 1..s + m]
        print "Pattern occurs with shift" s
```

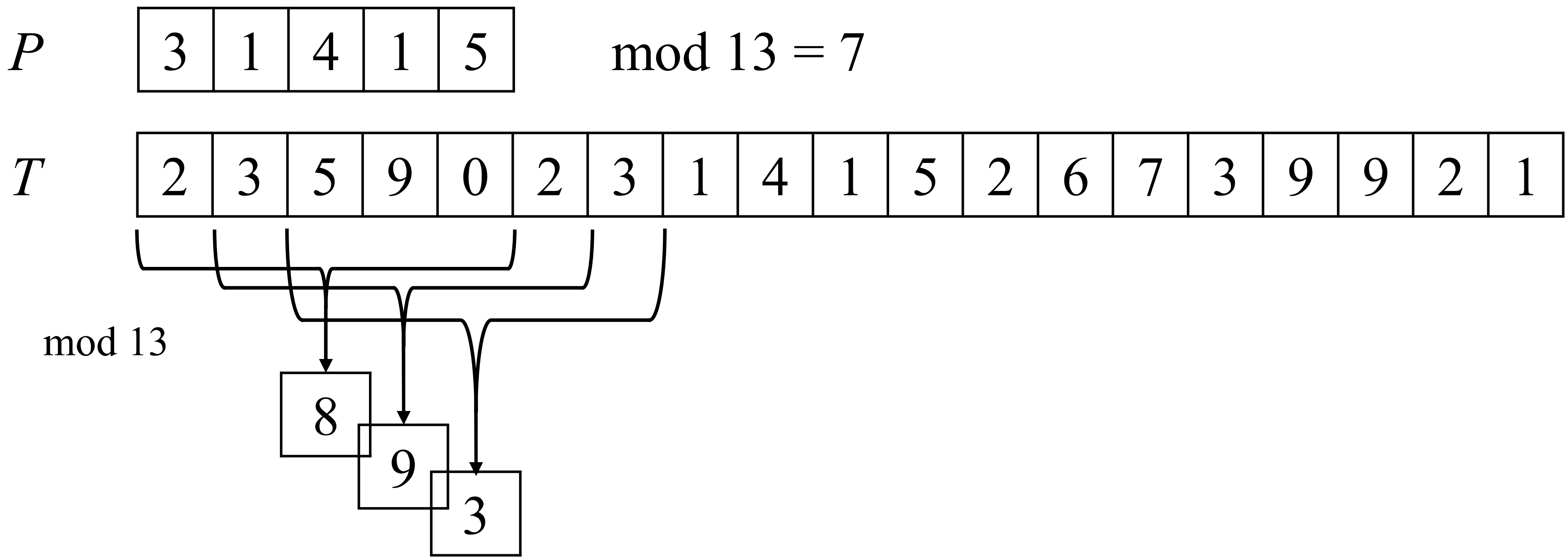
- Text $T = acaabc$, and pattern $P = aab$.

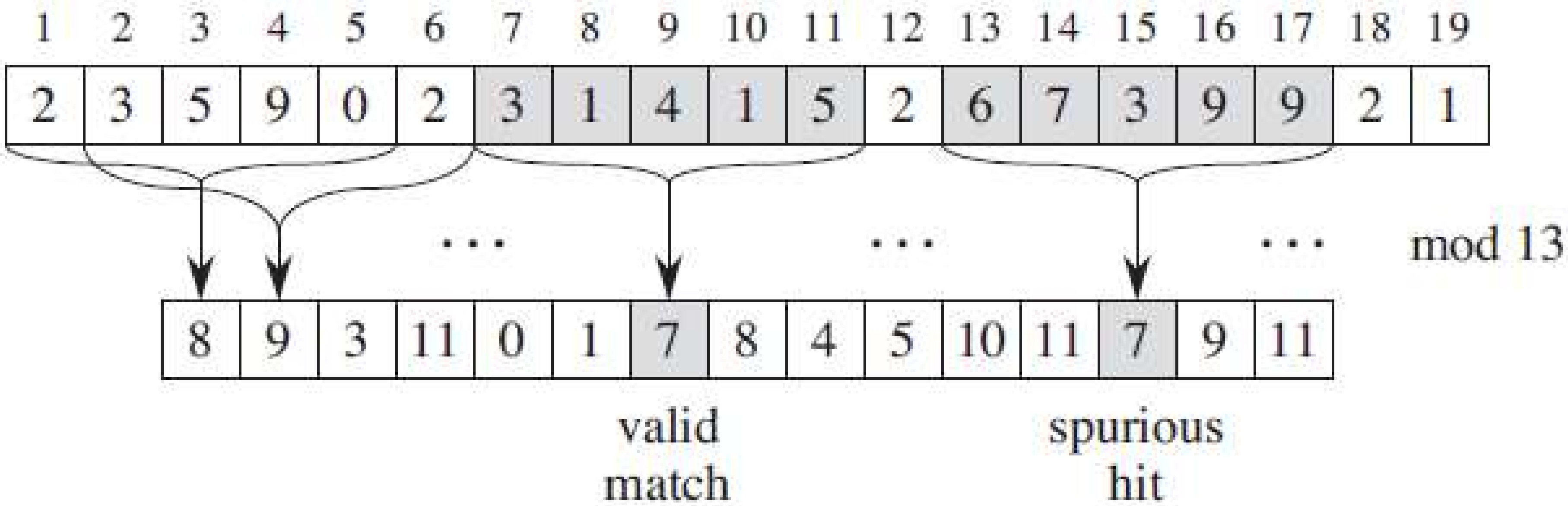


Rabin-Karp Algorithm

- Uses elementary number-theoretic notions.
  - Equivalence of two numbers modulo a third number.
- Let, $\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
- String of k consecutive characters represents a length-k decimal number.
  - Thus, character string 31415 corresponds to the decimal number 31,415.
- Note:
  - In the general case, each character is a digit in radix-d notation, where $d = |\Sigma|$.)

Example


P
3 1 4 1 5
mod 13 = 7
T
2 3 5 9 0 2 3 1 4 1 5 2 6 7 3 9 9 2 1
mod 13
8
9
3


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
2 3 5 9 0 2 3 1 4 1 5 2 6 7 3 9 9 2 1
… … …
mod 13
8 9 3 11 0 1 7 8 4 5 10 11 7 9 11
valid
match
spurious
hit

A few calculations…

- For a pattern $P\,[1..m]$, let p denote its corresponding value in radix-d notation.

- Using Horner's rule, p can be computed in time $\Theta(m)$.

$$p = P\,[m] + d(P\,[m-1] + d(P\,[m-2] + \ldots + d(P\,[2] + dP[1])\ \ldots)).$$

- Similarly, for a text $T\,[1..n]$, let t_s denotes the radix-d notation value of the length-m substring $T\,[s+1..s+m]$, for $s = 0, 1, \ldots, n-m$.

- Again, t_0 can be computed from $T\,[1..m]$ in $\Theta(m)$.

Contd…

- Each of the remaining values $t_1, t_2, \ldots, t_{n-m}$ can be computed in constant time.
  - Subtracting $d^{m-1}T[s+1]$ removes the high-order digit from t_s, multiplying the result by d shifts the number left by one digit position, and adding $T[s+m+1]$ brings in the appropriate low-order digit.

$$t_{s+1} = d\,(t_s - d^{m-1}T[s+1]) + T[s+m+1].$$

- Let $h = d^{m-1}$, then

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1].$$

- Modulus:
  - p modulo q takes $\Theta(m)$ time.
  - For all t_s, t_s modulo q takes $\Theta(n-m+1)$ time.

Algorithm

RABIN-KARP-MATCHER(T, P, d, q)

```
n = T.length
m = P.length
h = d^(m-1) mod q
p = 0
t_0 = 0
for i = 1 to m                      // preprocessing
    p = (dp + P[i]) mod q
    t_0 = (dt_0 + T[i]) mod q
for s = 0 to n - m                  // matching
    if p == t_s
        if P[1..m] == T[s+1..s+m]
            print "Pattern occurs with shift" s
    if s < n - m
        t_{s+1} = (d(t_s - T[s+1]h) + T[s+m+1]) mod q
```

Example – 1

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- n = 19, m = 5, n – m = 14.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 1: $s = 0, t_0 = 8, p = 7$.

$p == t_0 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_1 = 10\,(8 - 3\,(2)) + 2 \pmod{13}$$

$$t_1 = 10\,(2) + 2 \pmod{13} = 22 \pmod{13} = 9.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 2: $s = 1$, $t_1 = 9$, $p = 7$.

$p == t_1 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_2 = 10\,(9 - 3\,(3)) + 3 \pmod{13}$$

$$t_2 = 10\,(0) + 3 \pmod{13} = 3 \pmod{13} = 3.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \|\Sigma\| = 10$ $q = 13$																			

- $n = 19, m = 5, n - m = 14.$
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3.$
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7.$
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8.$

Step 3: $s = 2, t_2 = 3, p = 7.$

$p == t_2 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_3 = 10\,(3 - 3\,(5)) + 1 \pmod{13} = 10\,(-12) + 1 \pmod{13}$$

Because -12 mod 13 = 1

$$t_3 = 10\,(1) + 1 \pmod{13} = 11 \pmod{13} = 11.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 4: $s = 3$, $t_3 = 11$, $p = 7$.

$p == t_3 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_4 = 10\,(11 - 3\,(9)) + 4 \pmod{13} = 10\,(-16) + 4 \pmod{13}$$

Because -16 mod 13 = 10

$$t_4 = 10\,(10) + 4 \pmod{13} = 104 \pmod{13} = 0.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \|\Sigma\| = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 5: $s = 4$, $t_4 = 0$, $p = 7$.

$p == t_4 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT\,[s+1]) + T\,[s+m+1] \pmod{13}$$

$$t_5 = 10\,(0 - 3\,(0)) + 1 \pmod{13}$$

$$t_5 = 1 \pmod{13} = 1.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 6: $s = 5$, $t_5 = 1$, $p = 7$.

$p == t_5 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_6 = 10\,(1 - 3\,(2)) + 5 \pmod{13} = 10\,(-5) + 5 \pmod{13}$$

$$t_6 = 10\,(8) + 5 \pmod{13} = 85 \pmod{13} = 7.$$

Because -5 mod 13 = 8

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \|\Sigma\| = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 7: $s = 6$, $t_6 = 7$, $p = 7$.

$p == t_6 \rightarrow$ Yes.

Character by character matching $p[1..5] == T[7..11]$.

$\{3\ 1\ 4\ 1\ 5\} == \{3\ 1\ 4\ 1\ 5\}$. Match, hence $s = 6$ is a valid shift.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_7 = 10\,(7 - 3\,(3)) + 2 \pmod{13} = 10\,(-2) + 2 \pmod{13}$$

$$t_7 = 10\,(11) + 2 \pmod{13} = 112 \pmod{13} = 8.$$

Because -2 mod 13 = 11

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 8: $s = 7$, $t_7 = 8$, $p = 7$.

$p == t_7 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT\,[s + 1]) + T\,[s + m + 1]\ (\text{mod } 13)$$

$$t_8 = 10\,(8 - 3\,(1)) + 6\ (\text{mod } 13)$$

$$t_8 = 10\,(5) + 6\ (\text{mod } 13) = 56\ (\text{mod } 13) = 4.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 9: $s = 8$, $t_8 = 4$, $p = 7$.

$p == t_8 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_9 = 10\,(4 - 3\,(4)) + 7 \pmod{13} = 10\,(-8) + 7 \pmod{13}$$

$$t_9 = 10\,(5) + 7 \pmod{13} = 57 \pmod{13} = 5.$$

Because -8 mod 13 = 5

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 10: $s = 9$, $t_9 = 5$, $p = 7$.

$p == t_9 \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_{10} = 10\,(5 - 3\,(1)) + 3 \pmod{13}$$

$$t_{10} = 10\,(2) + 3 \pmod{13} = 23 \pmod{13} = 10.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 11: $s = 10$, $t_{10} = 10$, $p = 7$.

$p == t_{10} \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_{11} = 10\,(10 - 3\,(5)) + 9 \pmod{13} = 10\,(-5) + 9 \pmod{13}$$

$$t_{11} = 10\,(8) + 9 \pmod{13} = 89 \pmod{13} = 11.$$

Because -5 mod 13 = 8

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19, m = 5, n - m = 14.$
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3.$
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7.$
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8.$

Step 12: $s = 11, t_{11} = 11, p = 7.$

$p == t_{11} \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_{12} = 10\,(11 - 3\,(2)) + 9 \pmod{13}$$

$$t_{12} = 10\,(5) + 9 \pmod{13} = 59 \pmod{13} = 7.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 13: $s = 12$, $t_{12} = 7$, $p = 7$.

$p == t_{12} \rightarrow$ Yes.

Character by character matching $p[1..5] == T[13..17]$.

$\{3\ 1\ 4\ 1\ 5\} == \{6\ 7\ 3\ 9\ 9\}$. Mismatch occurs at first character.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_{13} = 10\,(7 - 3\,(6)) + 2 \pmod{13} = 10\,(-11) + 2 \pmod{13}$$

Because -11 mod 13 = 2

$$t_{13} = 10\,(2) + 2 \pmod{13} = 22 \pmod{13} = 9.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \lvert\Sigma\rvert = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 14: $s = 13$, $t_{13} = 9$, $p = 7$.

$p == t_{13} \rightarrow$ No.

$s < 14 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod{13}$$

$$t_{14} = 10\,(9 - 3\,(7)) + 1 \pmod{13} = 10\,(-12) + 1 \pmod{13}$$

Because -12 mod 13 = 1

$$t_{14} = 10\,(1) + 1 \pmod{13} = 11 \pmod{13} = 11.$$

Contd…

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Text (T)	2	3	5	9	0	2	3	1	4	1	5	2	6	7	3	9	9	2	1
Pattern: 3 1 4 1 5																			
$\Sigma = \{0, 1, \ldots, 9\}$ $d = \|\Sigma\| = 10$ $q = 13$																			

- $n = 19$, $m = 5$, $n - m = 14$.
- $h = 10^{5-1} \bmod 13 = 10^4 \bmod 13 = 3$.
- $p = (3 \times 10^4 + 1 \times 10^3 + 4 \times 10^2 + 1 \times 10^1 + 5 \times 10^0) \bmod 13 = 7$.
- $t_0 = (2 \times 10^4 + 3 \times 10^3 + 5 \times 10^2 + 9 \times 10^1 + 0 \times 10^0) \bmod 13 = 8$.

Step 15: $s = 14$, $t_{14} = 11$, $p = 7$.
$p == t_{14} \rightarrow$ No.
$s < 14 \rightarrow$ No.

Step 16: $s = 15$.
Loop terminates.

Example – 2

Index	1	2	3	4	5	6	7	8
Text (T)	a	a	b	b	c	a	b	a
ASCII	97	97	98	98	99	97	98	97
Pattern: c a b								
$\Sigma = \{a, b, \ldots, z\}$	$d = \|\Sigma\| = 26$	$q = 3$						

- $n = 8$, $m = 3$, $n - m = 5$.
- $h = 26^{3-1} \bmod 3$
  $= 26^2 \bmod 3 = 1.$
- $p = (99 \times 26^2 + 97 \times 26^1 + 98 \times 26^0) \bmod 3 = 1.$
- $t_0 = (97 \times 26^2 + 97 \times 26^1 + 98 \times 26^0) \bmod 3 = 2.$

Step 1: $s = 0$, $t_0 = 2$, $p = 1$.

$p == t_0 \rightarrow$ No.

$s < 5 \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT[s+1]) + T[s+m+1] \pmod 3$$

$$t_1 = 26\,(2 - 1\,(97)) + 98 \pmod 3$$

$$t_1 = 26\,(2 - 1\,(1)) + 2 \pmod 3$$

(because 97 mod 3 = 1 and 98 mod 3 = 2)

$$= 26\,(1) + 2 \pmod 3$$

$$= 28 \pmod 3 \qquad = 1.$$

Contd…

Index	1	2	3	4	5	6	7	8
Text (T)	a	a	b	b	c	a	b	a
ASCII	97	97	98	98	99	97	98	97
Pattern: c a b								

Step 2: $s = 1, t_1 = 1, p = 1.$

$p == t_1 \rightarrow$ Yes.

Character by character matching p[1..3] == T[2..4].

{c a b} == {a b b}. Mismatch occurs at first character.

$s < 5 \quad \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT\,[s+1]) + T\,[s+m+1]\ (\text{mod } 3)$$

$$t_2 = 26\,(1 - 1\,(97)) + 99\ (\text{mod } 3)$$

$$t_2 = 26\,(1 - 1\,(1)) + 0\ (\text{mod } 3)$$

(because 97 mod 3 = 1 and 99 mod 3 = 0)

$$= 26\,(0)\ (\text{mod } 3)$$

$$= 0.$$

Contd…

Index	1	2	3	4	5	6	7	8
Text (T)	a	a	b	b	c	a	b	a
ASCII	97	97	98	98	99	97	98	97
Pattern: c a b								

Step 3: $s = 2, t_2 = 0, p = 1.$

$p == t_2 \rightarrow$ No.

$s < 5 \quad \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT\,[s+1]) + T\,[s+m+1]\ (\text{mod } 3)$$

$$t_3 = 26\,(0 - 1\,(98)) + 97\ (\text{mod } 3)$$

$$t_3 = 26\,(0 - 1\,(2)) + 1\ (\text{mod } 3)$$

(because 97 mod 3 = 1 and 98 mod 3 = 2)

$$= 26\,(0 - 2) + 1\ (\text{mod } 3)$$

$$= 26\,(0 + 1) + 1\ (\text{mod } 3)$$

(because 3's complement of -2 = 1)

$$= 26\,(1) + 1\ (\text{mod } 3) \quad = 27\ (\text{mod } 3) \quad = 0.$$

Contd…

Index	1	2	3	4	5	6	7	8
Text (T)	a	a	b	b	c	a	b	a
ASCII	97	97	98	98	99	97	98	97
Pattern: c a b								

Step 4: $s = 3, t_3 = 0, p = 1.$

$p == t_3 \rightarrow$ No.

$s < 5 \quad \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT\,[s + 1]) + T\,[s + m + 1]\ (\text{mod } 3)$$

$$t_4 = 26\,(0 - 1\,(98)) + 98\ (\text{mod } 3)$$

$$t_4 = 26\,(0 - 1\,(2)) + 2\ (\text{mod } 3)$$

(because 98 mod 3 = 2)

$$= 26\,(0 - 2) + 2\ (\text{mod } 3)$$

$$= 26\,(0 + 1) + 2\ (\text{mod } 3)$$

(because 3's complement of -2 = 1)

$$= 26\,(1) + 2\ (\text{mod } 3) \quad = 28\ (\text{mod } 3) \quad = 1.$$

Contd…

Index	1	2	3	4	5	6	7	8
Text (T)	a	a	b	b	c	a	b	a
ASCII	97	97	98	98	99	97	98	97
Pattern: c a b								

Step 5: $s = 4$, $t_4 = 1$, $p = 1$.

$p == t_4 \rightarrow$ Yes.

Character by character matching $p[1..3] == T[5..7]$.

$\{c\ a\ b\} == \{c\ a\ b\}$. Match, hence $s = 4$ is a valid shift.

$s < 5 \quad \rightarrow$ Yes.

$$t_{s+1} = d\,(t_s - hT\,[s + 1]) + T\,[s + m + 1]\ (\text{mod } 3)$$

$$t_5 = 26\,(1 - 1\,(99)) + 97\ (\text{mod } 3)$$

$$t_5 = 26\,(1 - 1\,(0)) + 1\ (\text{mod } 3)$$

(because 97 mod 3 = 1 and 99 mod 3 = 0)

$$= 26\,(1 - 0) + 1\ (\text{mod } 3)$$

$$= 26\,(1) + 1\ (\text{mod } 3) \quad = 27\ (\text{mod } 3) \quad = 0.$$

Contd…

Index	1	2	3	4	5	6	7	8
Text (T)	a	a	b	b	c	a	b	a
ASCII	97	97	98	98	99	97	98	97
Pattern: c a b								

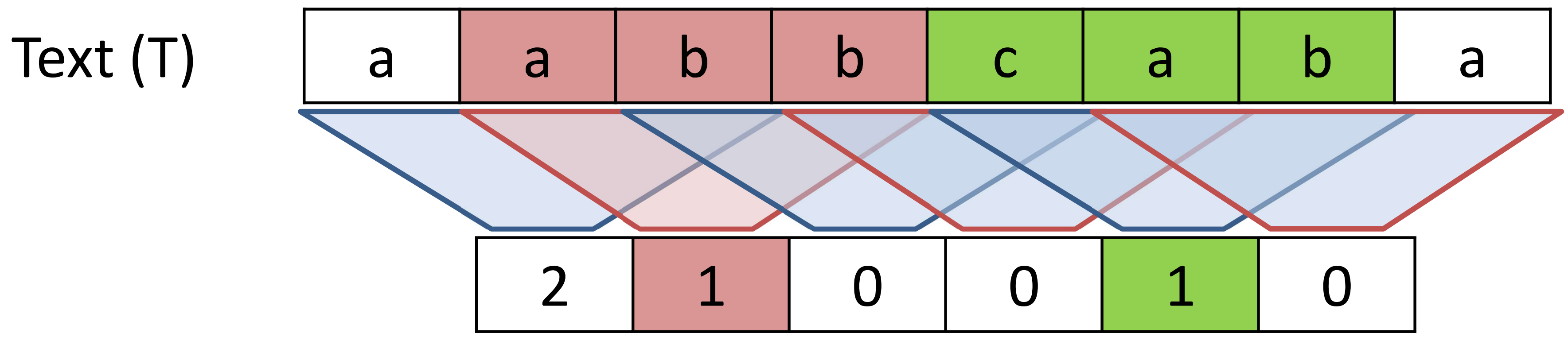
Step 6: $s = 5$, $t_5 = 0$, $p = 1$.

$p == t_5 \rightarrow$ No.

$s < 5 \quad \rightarrow$ No.

Step 7: $s = 6$.

Loop terminates.



Complexity

- Takes $\Theta(m)$ preprocessing time.
- Worst-case running time is $O(m\ (n - m + 1))$.
  - Example: $P = a^m$ and $T = a^n$, each of the $[n - m + 1]$ possible shifts is valid.
- In many applications, there are a few valid shifts (say some constant c). In such applications, the expected matching time is only $O(n - m + 1) + cm)$, plus the time required to process spurious hits.

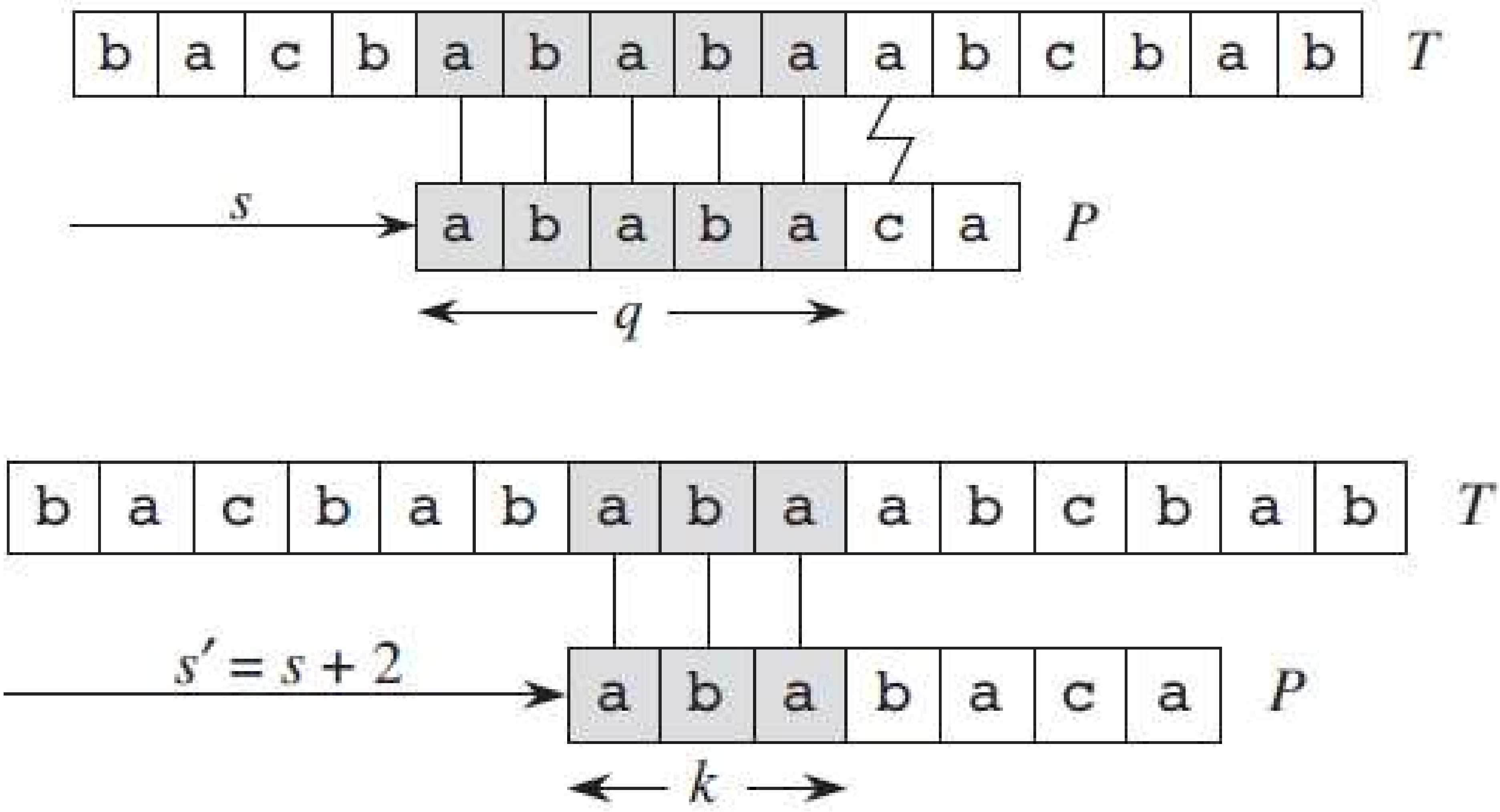
Contd…

- Probabilistic analysis
  - The probability of a false positive hit for a random input is $1/q$.
  - The expected number of false positive hits is $O(n/q)$.
  - The expected run time is $O(n) + O(m(v + n/q)))$, if v is the number of valid shifts.
- Choosing $q \geq m$ and having only a constant number of hits, then the expected matching time is $O(n + m)$.
- Since $m \leq n$, this expected matching time is $O(n)$.

Knuth-Morris-Pratt Algorithm

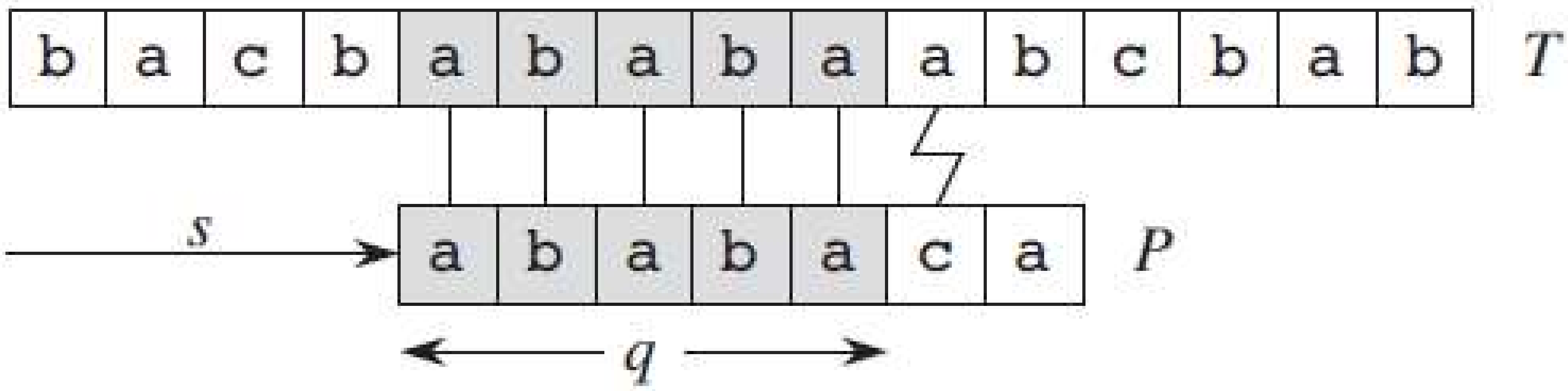
- Based on the concept of prefix function for a pattern.
  - Encapsulates knowledge about how the pattern matches against shifts of itself.
  - This information can be used to avoid testing of invalid shifts.
- T: b a c b a b a b a a b c b a b
- P: a b a b a c a

Contd…



b a c b a b a b a a b c b a b
T
s
a b a b a c a
P
q
b a c b a b a b a a b c b a b
T
s′ = s + 2
a b a b a c a
P
k

Contd…

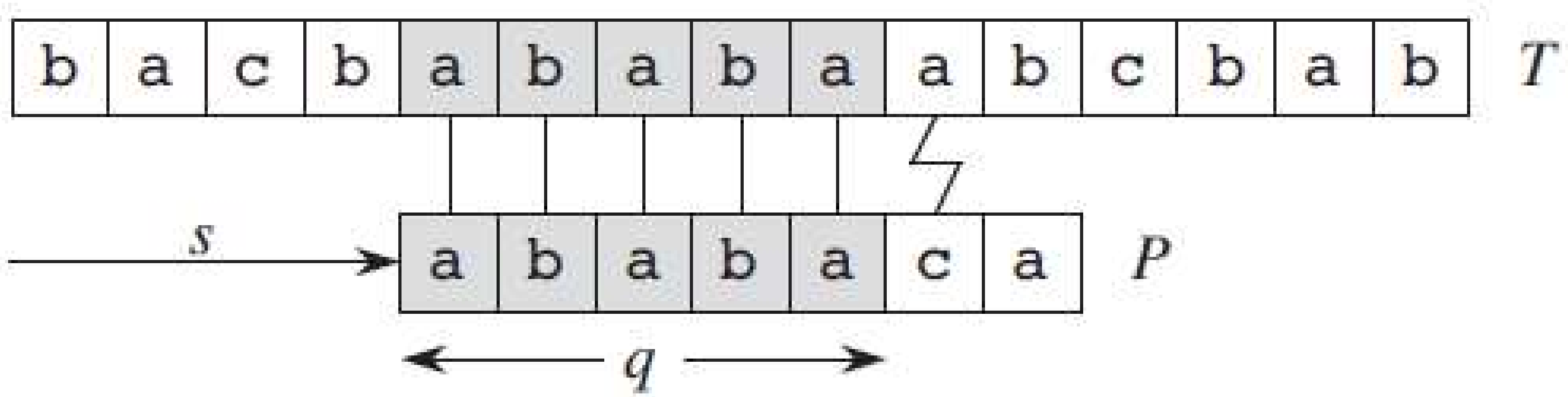


- Given that pattern characters P [1..q] match text characters T [s + 1..s + q], what is the least shift
  s' > s such that for some k < q,
  P [1..k] = T[s' + 1..s' + k], where s' + k = s + q?
  i.e. s' = s + (q – k)
- Best case k = 0. Skips (q – 1) shifts.

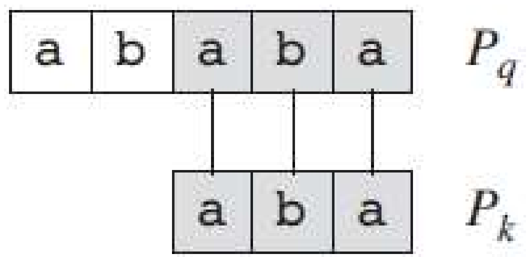
Contd…

Note: $P_k = P[1..k]$. Similarly, $P_q = P[1..q]$.

- In other words, knowing that P_q is a suffix of T_{s+q}, find the longest proper prefix P_k of P_q that is also a suffix of T_{s+q}.



- Since suffix P_k of T_{s+q} should also be suffix of P_q. Thus, an equivalent statement is "determine the greatest $k < q$, such that P_k is a suffix of P_q".



Prefix Function

q	1	2	3	4	5	6	7
k	0	0	1	2			

- P: a b a b a c a
- $P_q = P[1..q]$
- $P_1 = \{a\}$. $P_k = \{\}$. No matching prefix and suffix of P_1.
- $P_2 = \{a\ b\}$. $P_k = \{\}$. No matching prefix and suffix of P_2.
- $P_3 = \{a\ b\ a\}$. $P_k = \{a\}$.

Prefix	Suffix	k
a	a	1
a b	b a	2

$P_4 = \{a\ b\ a\ b\}$. $P_k = \{a\ b\}$.

Prefix	Suffix	k
a	b	1
a b	a b	2
a b a	b a b	3

Contd...

q	1	2	3	4	5	6	7
k	0	0	1	2	3	0	1

- P: a b a b a c a
- P_q = P[1..q]
- P_5 = {a b a b a}. P_k = {a b a}.
  Keep the longest.

Prefix	**Suffix**	**k**
a	a	1
a b	b a	2
a b a	a b a	3
a b a b	b a b a	4

- P_7 = {a b a b a c a}. P_k = {a}.

Prefix	**Suffix**	**k**
a	a	1
a b	c a	2
a b a	a c a	3
a b a b	b a c a	4
a b a b a	a b a c a	5
a b a b a c	b a b a c a	6

- P_6 = {a b a b a c}. P_k = {}. No matching prefix and suffix of P_6 as 'c' does not appear in any of the proper prefix.

Prefix Function

$q \rightarrow i$	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$k \rightarrow \pi[i]$	0	0	1	2	3	0	1

- Given a pattern P [1..m], the prefix function for the pattern P is the function $\Pi : \{1, 2, \ldots, m\} \rightarrow \{0, 1, \ldots, m - 1\}$ such that

$\Pi [q] = \max \{k : k < q$ and P_k is a proper suffix of P_q.

- It contains the length of the longest prefix of P that is a proper suffix of P_q.

Contd…

COMPUTE-PREFIX-FUNCTION(P)

```
m = P.length
let π[1..m] be a new array
π[1] = 0
k = 0
for q = 2 to m
    while k > 0 and P[k + 1] ≠ P[q]
        k = π[k]
    if P[k + 1] == P[q]
        k = k + 1
    π[q] = k
return π
```

KMP Algorithm

```
KMP-MATCHER(T, P)
n = T.length
m = P.length
π = COMPUTE-PREFIX-FUNCTION(P)
q = 0                                    // number of characters matched
for i = 1 to n                           // scan the text from left to right
    while q > 0 and P[q + 1] ≠ T[i]
        q = π[q]                         // next character does not match
    if P[q + 1] == T[i]
        q = q + 1                        // next character matches
    if q == m                            // is all of P matched?
        print "Pattern occurs with shift" i − m
        q = π[q]                         // look for the next match
```

Complexity

- The COMPUTE-PREFIXFUNCTION runs in $\Theta(m)$ time as the while loop in lines 6–7 executes at most $m - 1$ times altogether.
    1. Line 4 starts k at 0, and the only way to increase k is the increment operation in line 9, which executes at most once per iteration of the for loop of lines 5–10. Thus, the total increase in k is at most $m - 1$.
    2. Second, since $k < q$ upon entering the for loop and each iteration of the loop increments q and $k < q$ always. Assignments in lines 3 and 10 ensure that $\Pi[q] < q$ for all $q = 1, 2, \ldots, m$, which means that each iteration of the while loop decreases k.
    3. Third, k never becomes negative.
    - Altogether, the total decrease in k from the while loop is bounded from above by the total increase in k over all iterations of the for loop, which is $m - 1$.
- Using similar analysis, the matching time of KMP-MATCHER is $\Theta(n)$.

Example (Preprocessing)

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

- m = 7, Π[1] = 0, k = 0.

Step 1: q = 2, k = 0 P[k + 1] == P[q]

(if) P[1] == P[2]. Mismatch.

Π[2] = 0.

Step 2: q = 3, k = 0. P[k + 1] == P[q]

(if) P[1] == P[3]. Match. k++ = 1

Π[3] = 1.

Step 3: q = 4, k = 1. P[k + 1] == P[q]

(if) P[2] == P[4]. Match. k++ = 2

Π[4] = 2.

Contd…

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 4: q = 5, k = 2 P[k + 1] == P[q]

(if) P[3] == P[5]. Match. k++ = 3

Π[5] = 3.

Step 5: q = 6, k = 3. P[k + 1] == P[q]

(while) P[4] == P[6]. Mismatch. k = Π[k] = 1

(while) P[2] == P[6]. Mismatch. k = Π[k] = 0

(if) P[1] == P[6]. Mismatch.

Π[6] = 0.

Step 6: q = 7, k = 0. P[k + 1] == P[q]

(if) P[1] == P[7]. Match. k++ = 1

Π[7] = 1.

Step 7: q = 8 **Loop terminates.**

Contd…
(Matching)

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

- T:

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

- P:

a	b	a	b	a	c	a

Step 1: i = 1, q = 0 P[q + 1] == T[i]

(if) P[1] == T[1]. Mismatch.

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

✗

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 2: i = 2, q = 0 P[q + 1] == T[i]

(if) P[1] == T[2]. Match. q++ = 1

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 3: i = 3, q = 1 P[q + 1] == T[i]

(while) P[2] == T[3]. Mismatch.

q = Π[q] = 0

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

✗

a	b	a	b	a	c	a

(if) P[1] == T[3]. Mismatch.

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 4: i = 4, q = 0 P[q + 1] == T[i]

(if) P[1] == T[4]. Mismatch.

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 5: i = 5, q = 0 P[q + 1] == T[i]

(if) P[1] == T[5]. Match. q++ = 1

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 6: i = 6, q = 1 P[q + 1] == T[i]

(if) P[2] == T[6]. Match. q++ = 2

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 7: i = 7, q = 2 P[q + 1] == T[i]

(if) P[3] == T[7]. Match. q++ = 3

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 8: i = 8, q = 3 P[q + 1] == T[i]

(if) P[4] == T[8]. Match. q++ = 4

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 9: i = 9, q = 4 P[q + 1] == T[i]

(if) P[5] == T[9]. Match. q++ = 5

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 10: i = 10, q = 5 P[q + 1] == T[i]

(while) P[6] == T[10]. Mismatch. q = Π[q] = 3

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

(if) P[4] == T[10]. Match. q++ = 4

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Shifts skipped = 1 and first three characters are not compared. Comparison starts from P[4].

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 11: i = 11, q = 4 P[q + 1] == T[i]

(if) P[5] == T[11]. Match. q++ = 5

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

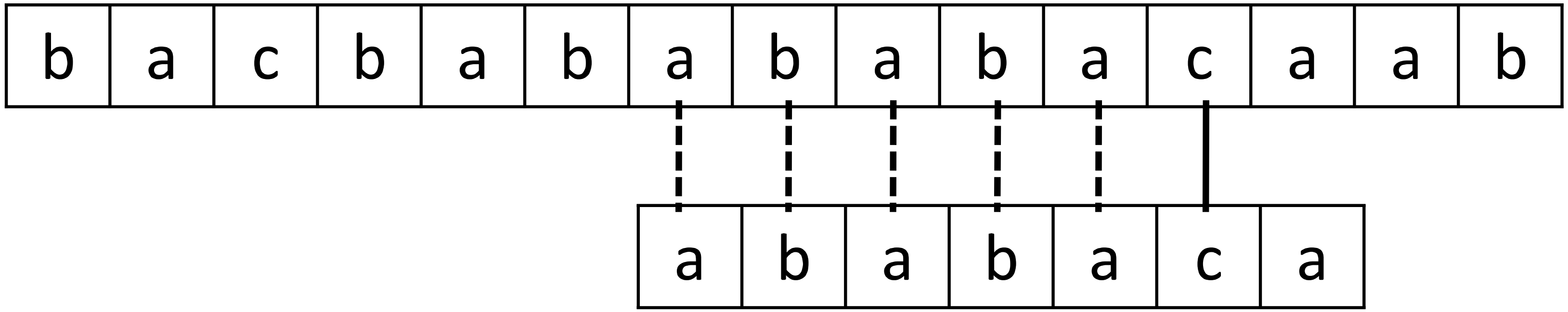
Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 12: i = 12, q = 5 P[q + 1] == T[i]

(if) P[6] == T[12]. Match. q++ = 6



Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 13: i = 13, q = 6 P[q + 1] == T[i]

(if) P[7] == T[13]. Match. q++ = 7

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

- q == m. Yes.
  - Pattern occurs with shift i – m =13 – 7 = 6.
  - q = Π[q] = 1.

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 14: i = 14, q = 1 P[q + 1] == T[i]

(while) P[2] == T[14]. Mismatch. q = Π[q] = 0

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Shifts skipped = 5 and first character is not compared. Comparison starts from P[2].

(if) P[1] == T[14]. Match. q++ = 1

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Contd…

n = 15
m = 7

i	1	2	3	4	5	6	7
$P[i]$	a	b	a	b	a	c	a
$\pi[i]$	0	0	1	2	3	0	1

Step 15: i = 15, q = 1 P[q + 1] == T[i]

(if) P[2] == T[15]. Match. q++ = 2

b	a	c	b	a	b	a	b	a	b	a	c	a	a	b

a	b	a	b	a	c	a

Step 16: i = 16

Loop terminates.