기본 알고리즘
제8장

2017-Fall

국민대학교 컴퓨터공학부 최준수

String Matching

- Substring search
  - Find pattern of length M in a text of length N. (typically $N \gg M$)

text → R E C U R S I V E A L G S T A C K Q U

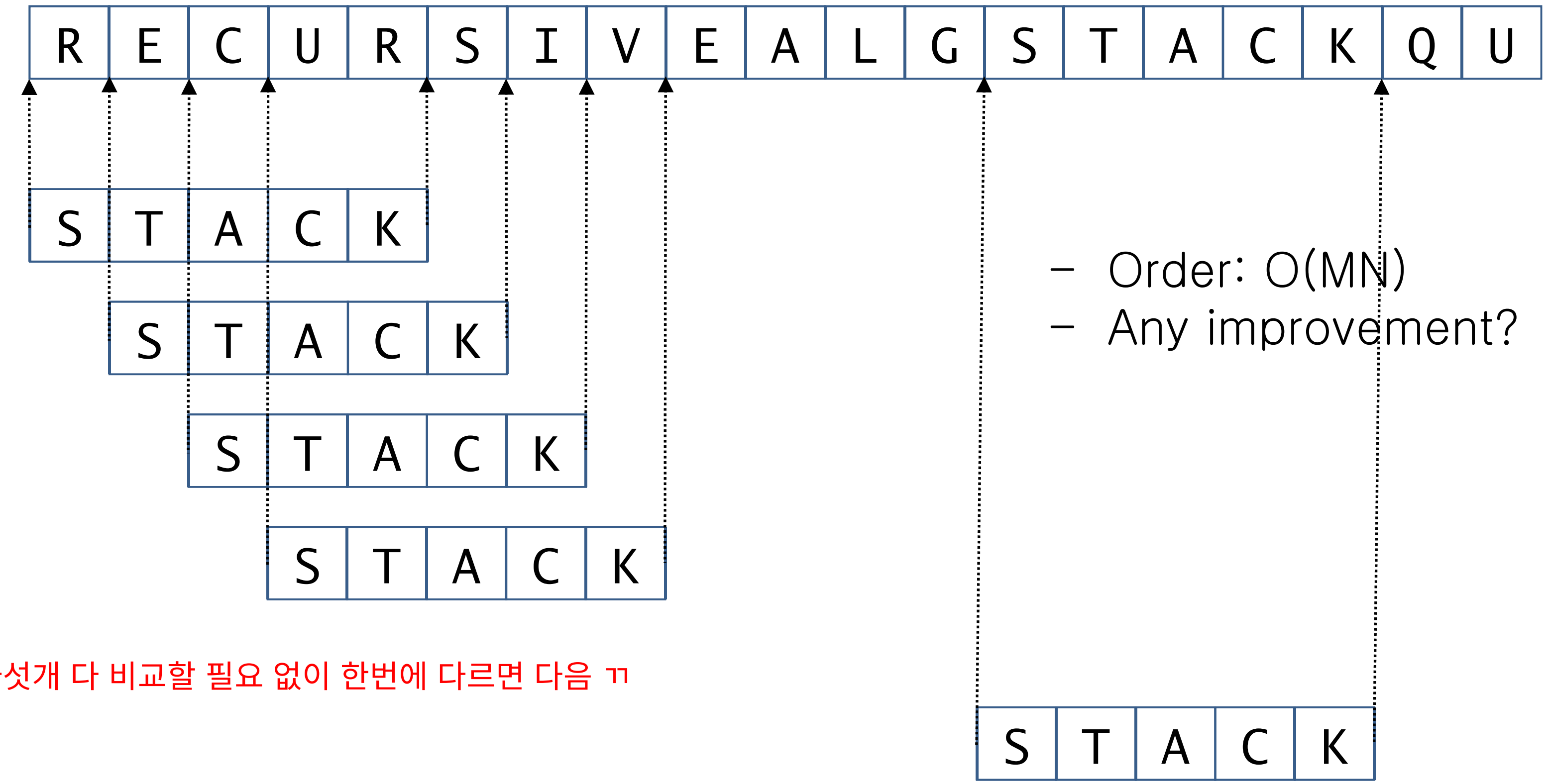
match

pattern → S T A C K

Algorithmic Engineering Laboratory

Brute-Force Substring Search

- Naïve Algorithm 1
  - Check for pattern starting at each text position



- Order: O(MN)
- Any improvement?

다섯개 다 비교할 필요 없이 한번에 다르면 다음 ㄱㄱ

Brute-Force Substring Search

- Naïve Algorithm 2
  - Check for pattern starting at each text position

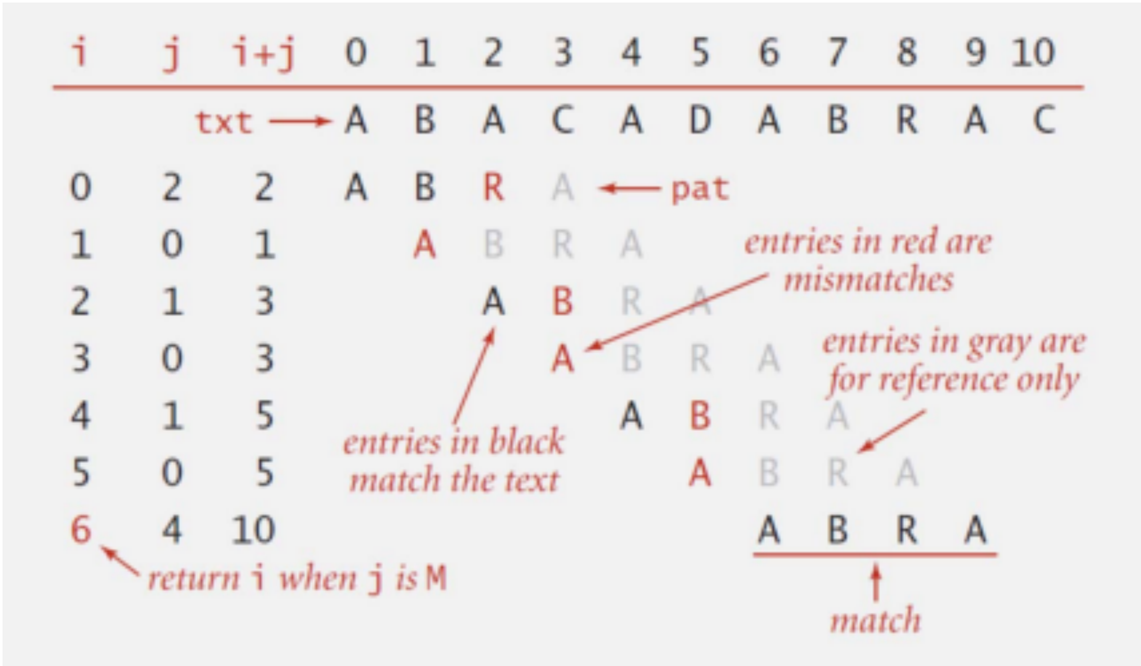
```
int naiveStringMatch(char text[], char pattern[])
{
    int patLength, txtLength;

    patLength = strlen(pattern);
    txtLength = strlen(text);

    for(int i=0; i <= txtLength - patLength; i++)
    {
        for(int j=0; j < patLength; j++)
            if(text[i+j] != pattern[j])
                break;
        if(j == patLength)
            return i;
    }
    return 0;
}
```

다르면 다음 으로

Find just only one sample of pattern.
Modify the code to find all samples of the pattern.

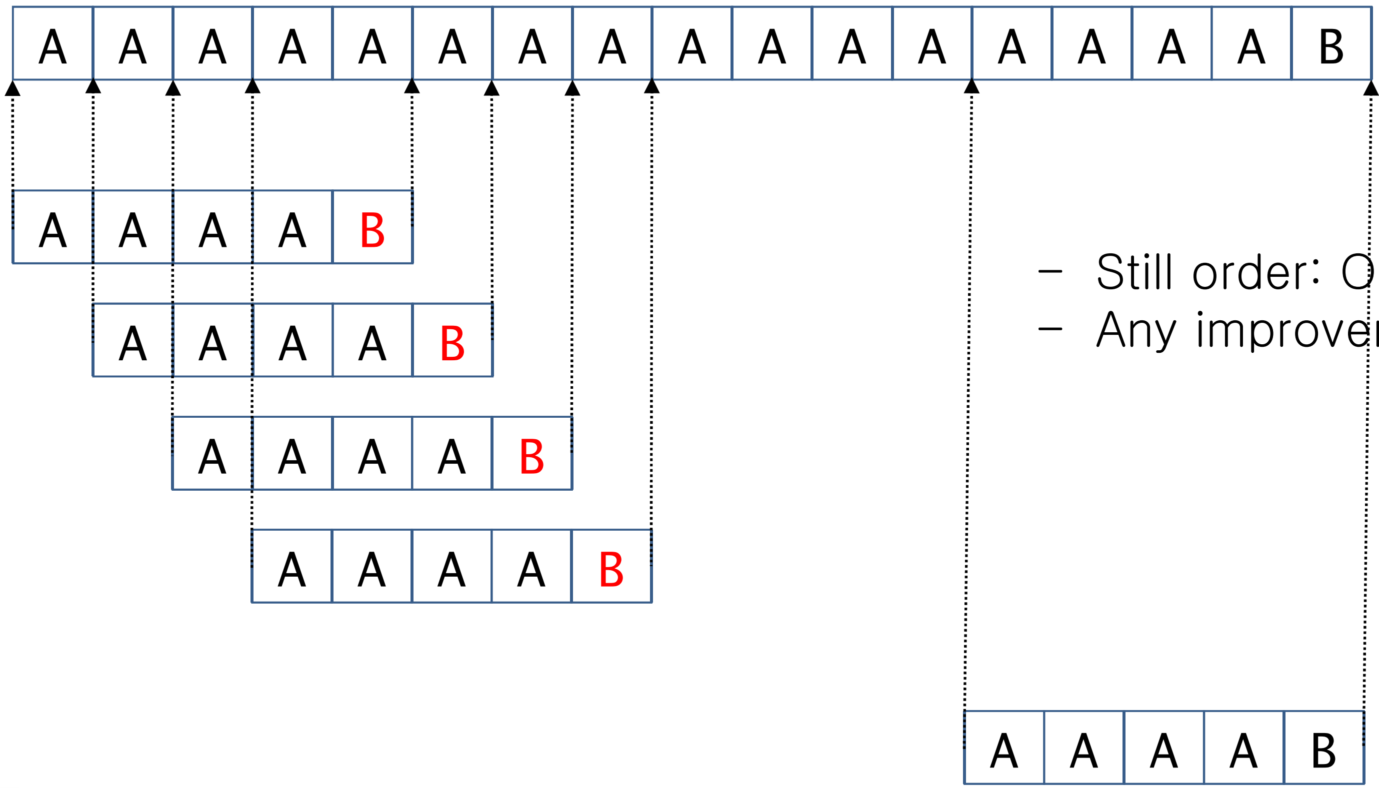
Brute-Force Substring Search

- Naïve Algorithm 2
  - Check for pattern starting at each text position



Brute-Force Substring Search

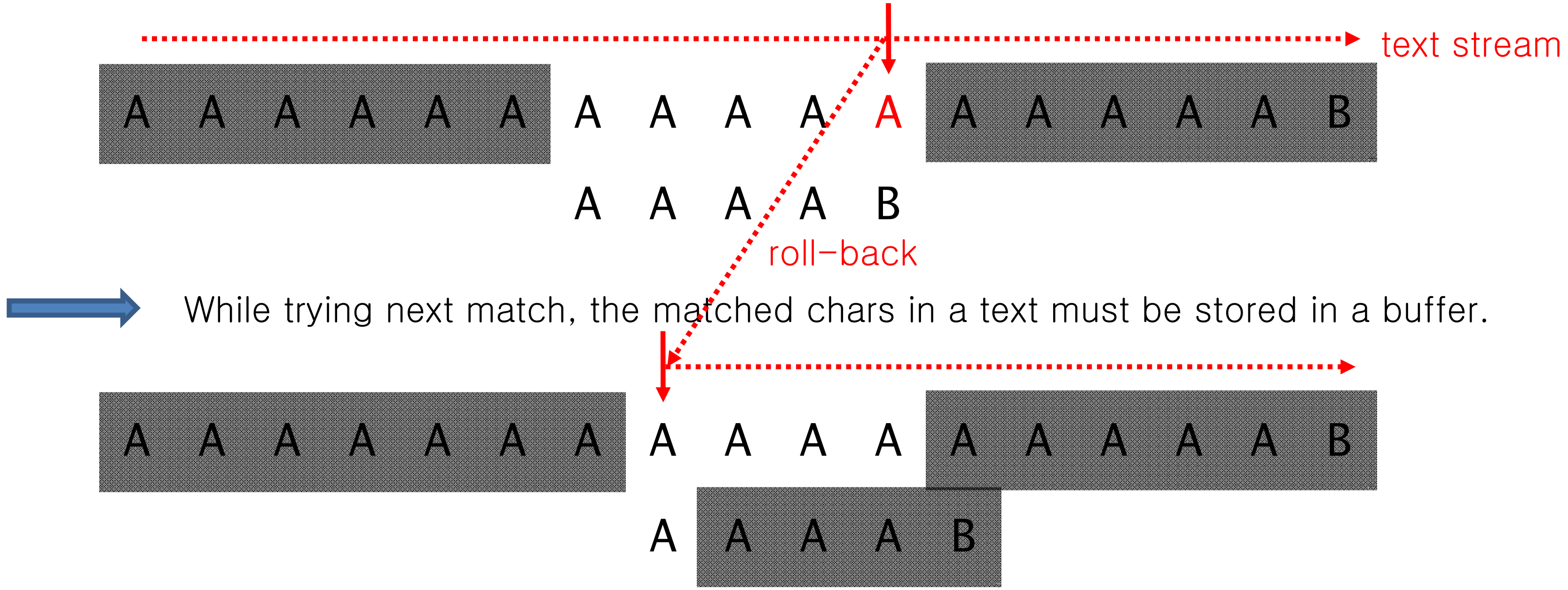
- Naïve Algorithm 2
  - Naïve algorithm can be slow if text and pattern are repetitive



- Still order: O(MN)
- Any improvement?

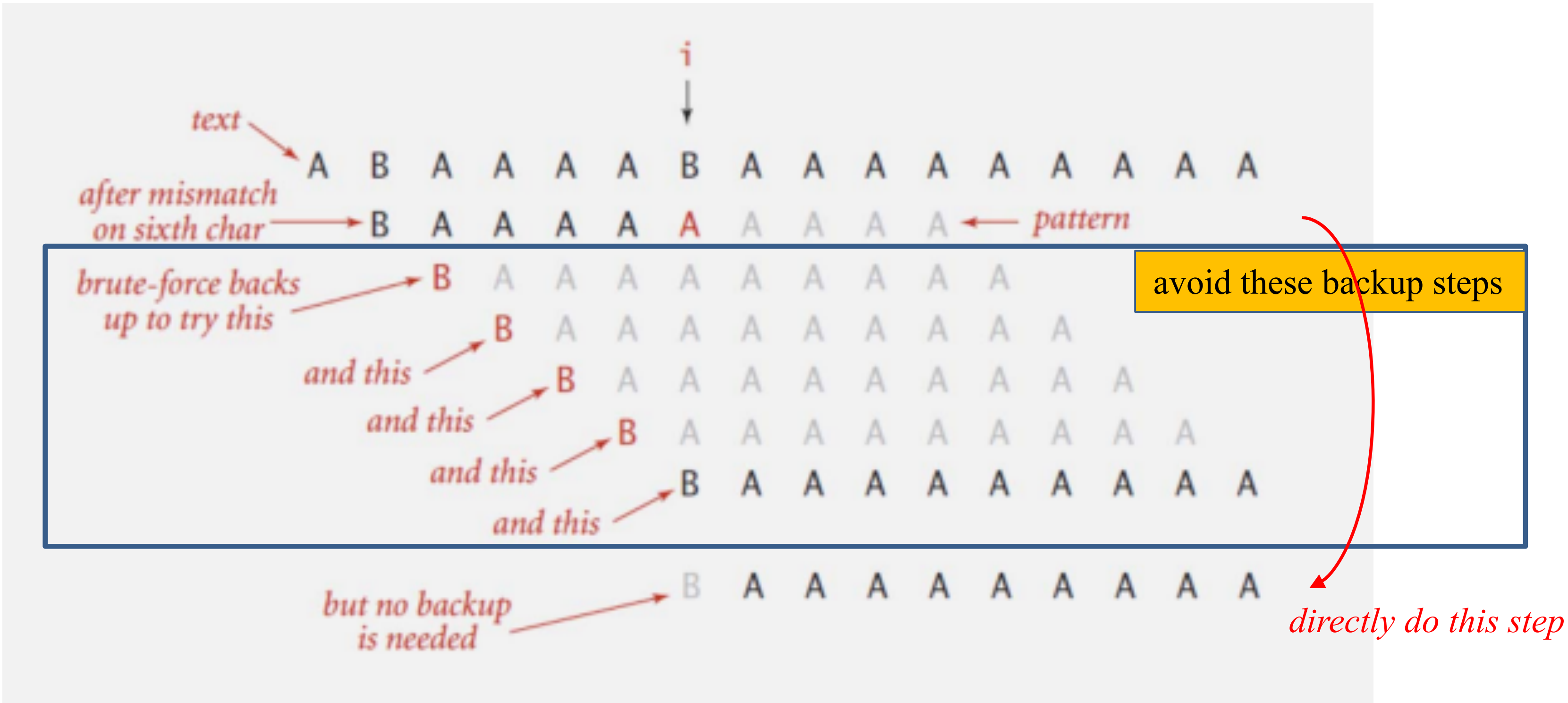
Brute-Force Substring Search

- Improvement
  - Develop a linear time algorithm
  - Avoid backup
    - Naïve algorithm needs backup for every mismatch
    - Thus naïve algorithm cannot be used when input text is a stream.



Knuth-Morris-Pratt(KMP) Algorithm

- KMP algorithm
  - Clever method to always avoid backup problem.

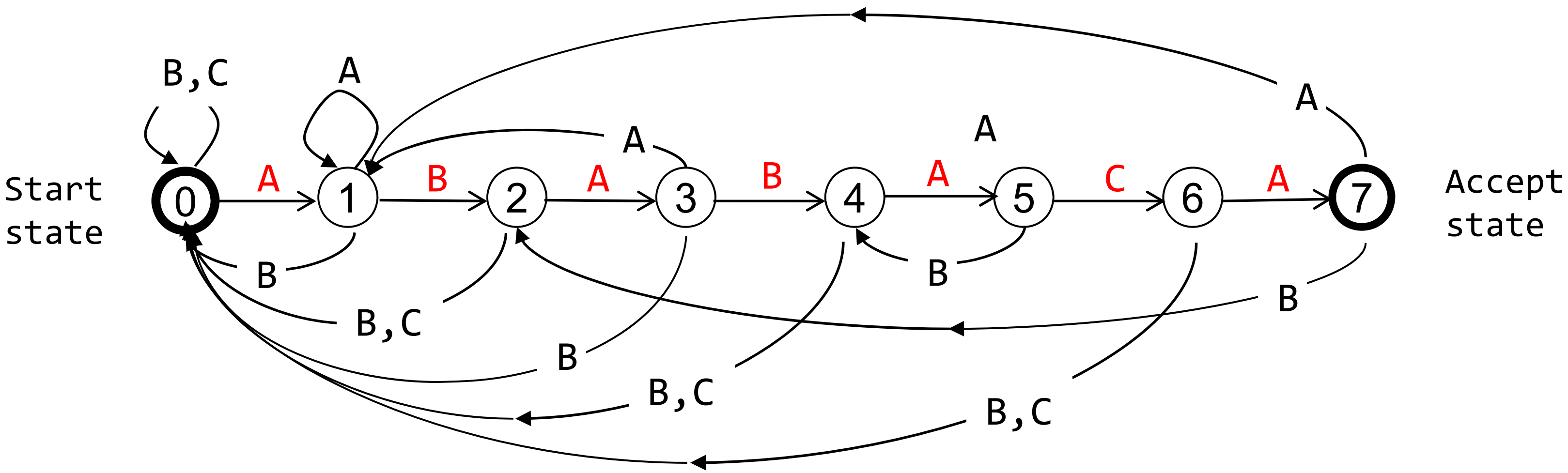


Deterministic Finite Automaton

- DFA(Deterministic Finite State Automaton)
  - Finite number of states (including start and accept states)
  - Exactly one transition for each char
  - Accept if sequence of transitions leads to accept state

롤벡을 안할수 있다.

DFA for pattern ABABACA

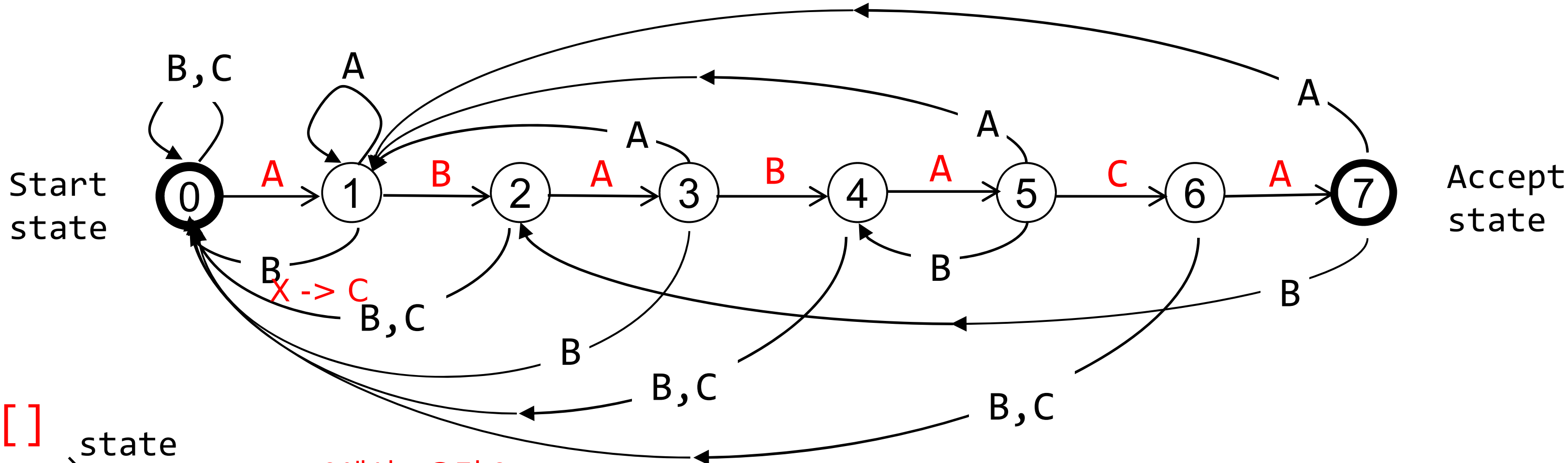


Algorithmic Engineering Laboratory

DFA

DFA for pattern ABABACA

내부적으로는 테이블을 만들어서!!!!!!



DFA[][]

2에서 A오면 3

char \ state	0	1	2	3	4	5	6	7
A	1	1	3	1	5	1	7	1
B	0	2	0	4	0	4	0	2
C	0	0	0	0	0	6	0	0
others	0	0	0	0	0	0	0	0

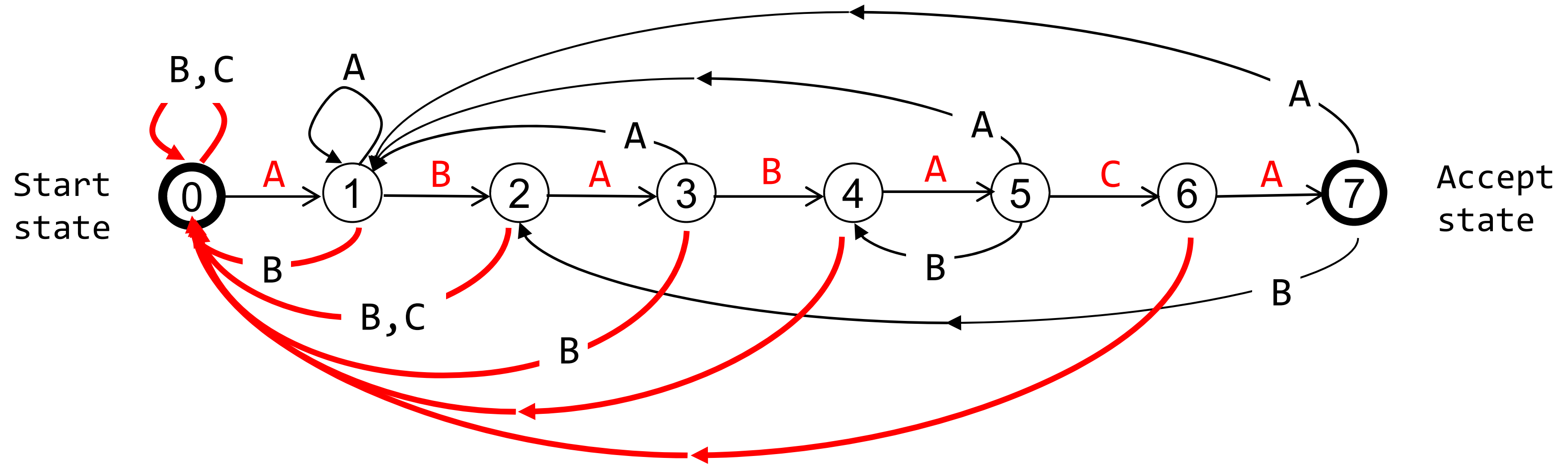
input

```
if in state j reading char c:
    if j is 7, halt and accept
    else move to state DFA[c][j]
```

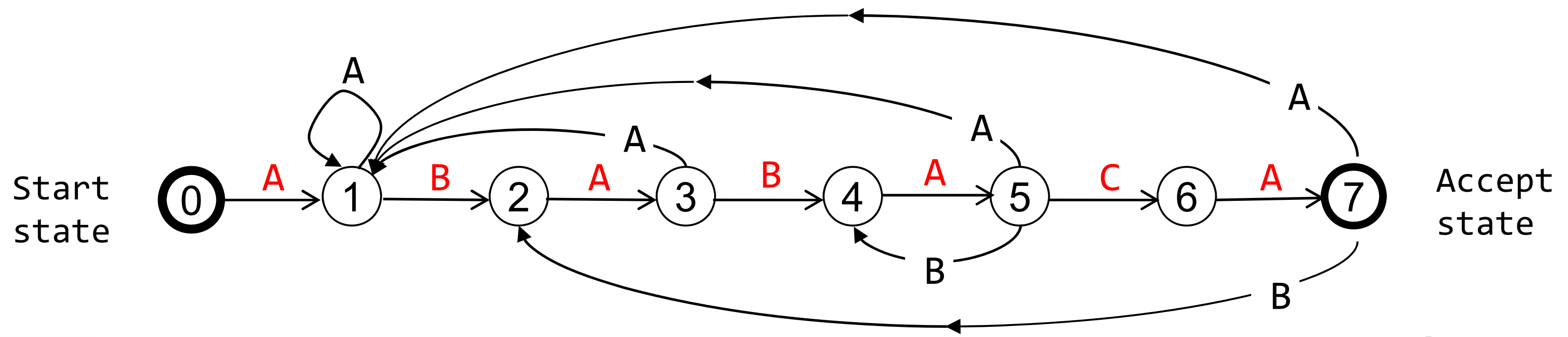
```
Example:
    text:  ABCABAABABABACACA
    state:012012312345456 7
```

DFA

DFA for pattern ABABACA



Simplified Diagram: remove transitions to state 0



Algorithm with DFA

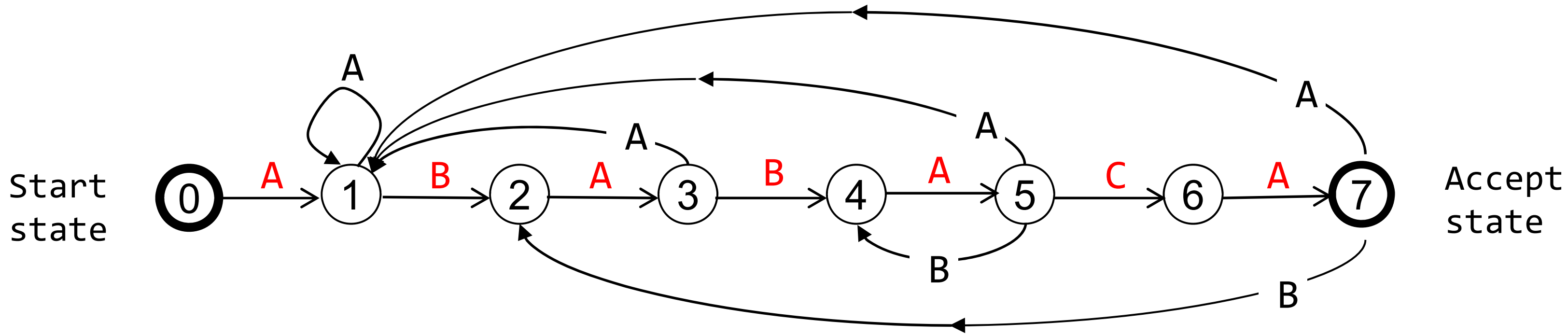
- Difference from naïve algorithm
  - Precomputation of DFA[][] from pattern
  - Text pointer `i` never decrements (no backup)

```
// patLength = strlen(pattern);
int KMP(char text[])
{
    int i, j, txtLength;

    txtLength = strlen(text);

    for(i=0, j=0; i <= txtLength && j < patLength; i++)
        j = DFA[text[i]][j];  // text[i] to be modified

    if(j == patLength)
        return i - patLength;
    else
        return 0;
}
```

- Order: O(N)
  simulation of DFA on text with no backup
- How to build DFA efficiently?

Algorithm with DFA

DFA for pattern ABABACA



Start state

Accept state

state	0	1	2	0	1	2	3	1	2	3	4	5	4	5	6	7	
text		A	B	C	A	B	A	A	B	A	B	A	B	A	C	A	C	A

return 8

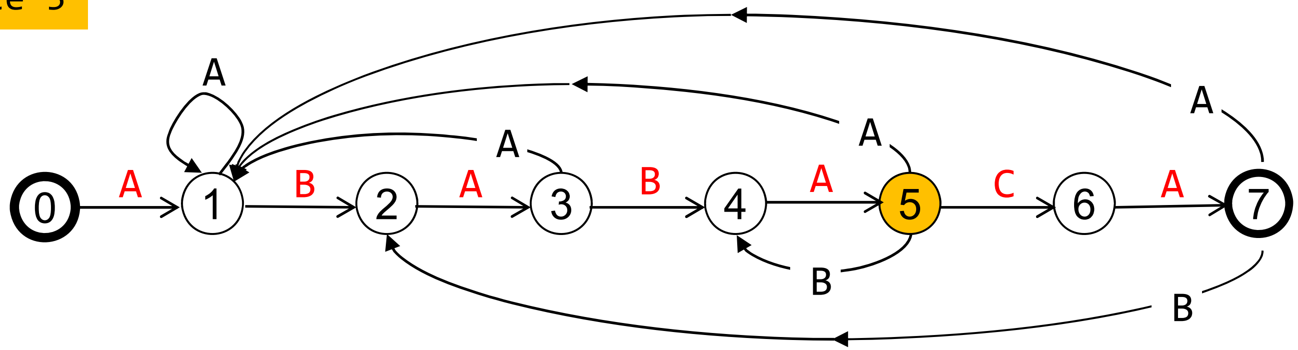
KMU 국민대학교 KOOKMIN UNIVERSITY

Algorithmic Engineering Laboratory

Interpretation of DFA

- The state of DFA represents
  - the number of characters in pattern that have been matched

State 5



text → B C B A A B A B A A B A

Suffix of txt[8]

pattern → A B A B A C A

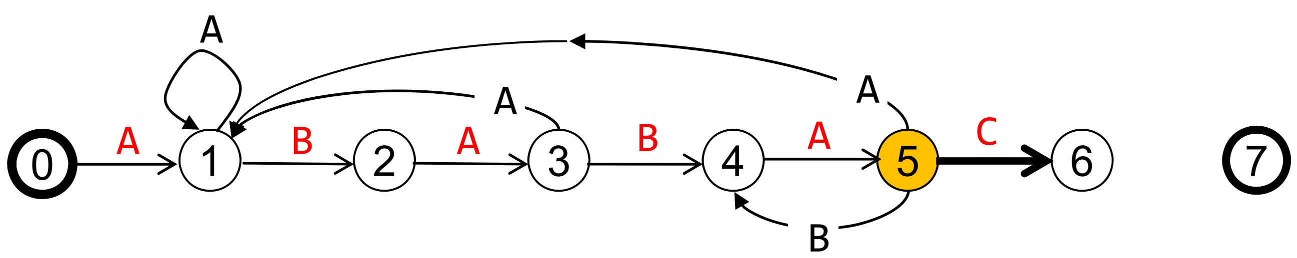
Prefix of pat[]

DFA Construction

- DFA Construction:
  - Suppose that all transitions from state `0` to state `j-1` are already computed
  - Match transition:
    - If in state `j` and next char `c == pat[j]`, then transit to state `j+1`.

Pattern: ABABACA

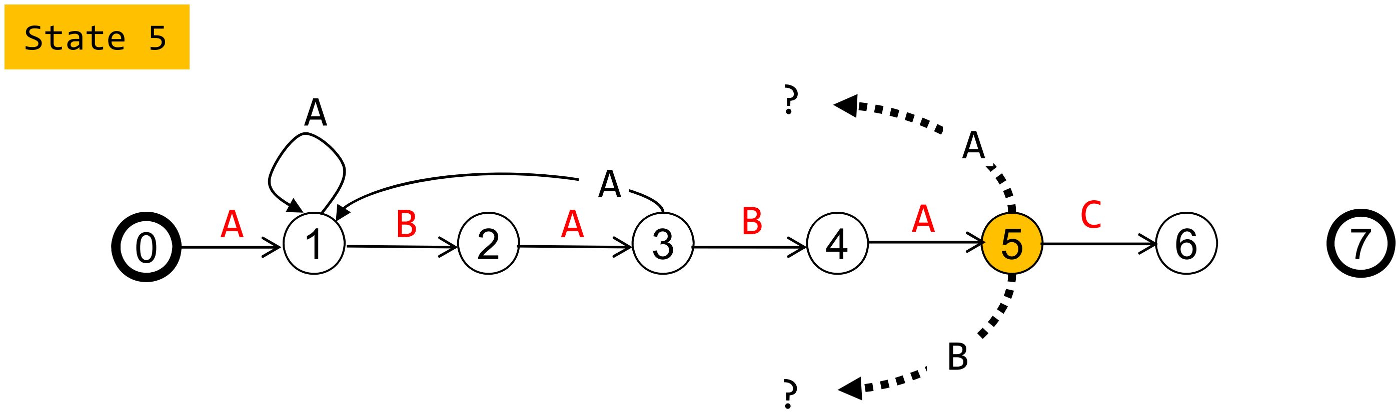
State 5



DFA Construction

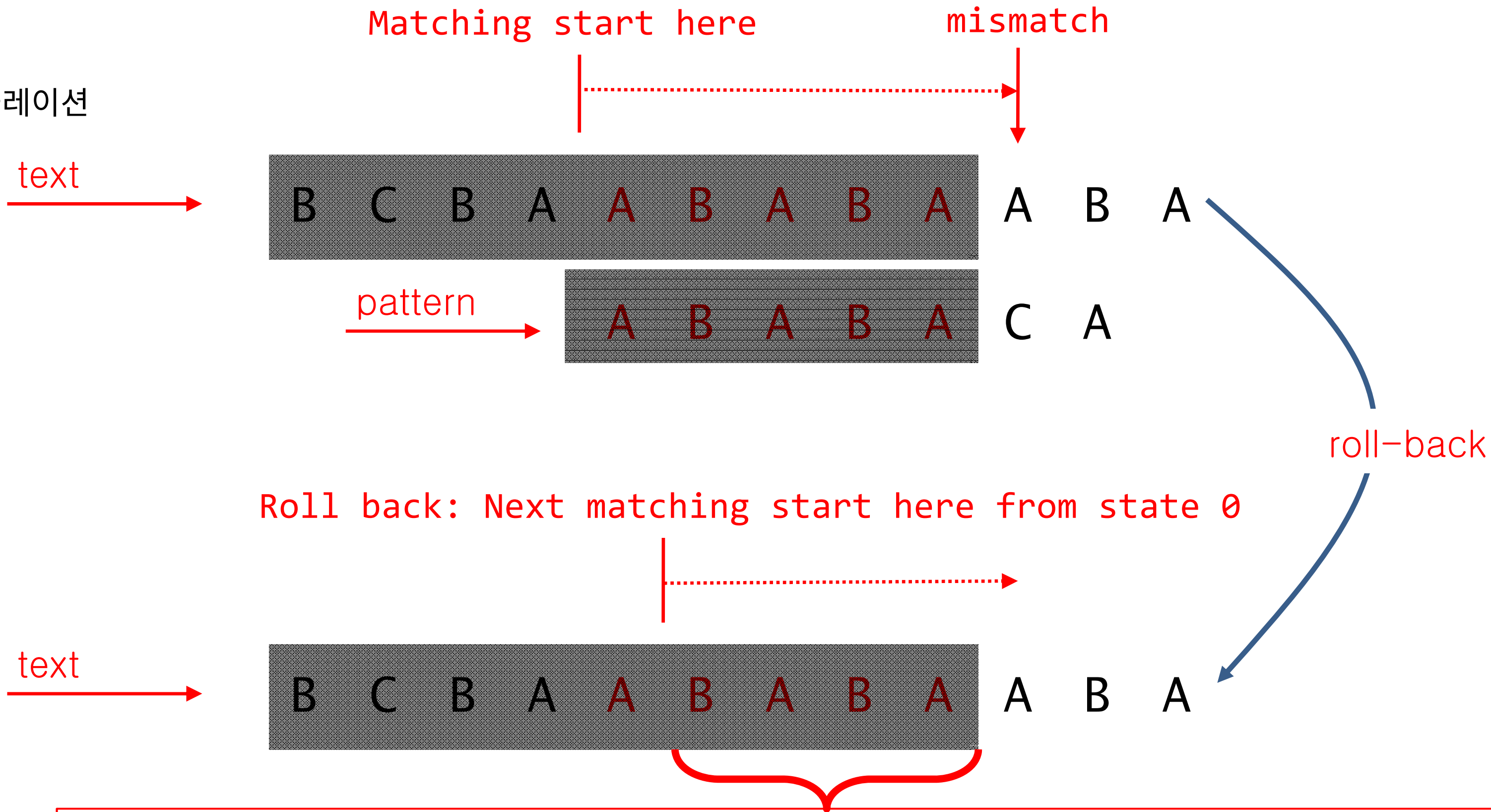
- DFA Construction:
  - Mismatch transition:
    - If in state `j` and next char `c != pat[j]`, then which state to transit?

Pattern: ABABACA



DFA Construction

X를 구하기위한 시뮬레이션



- *The same as* `pat[1] ~ pat[j-1]`
- Roll-back and transit to some state X by matching
  `pat[1] ~ pat[j-1]` from state 0 on DFA.
- Transit to the next state `DFA['A'][X]` for the mismatched char `'A'`.

DFA Construction

- DFA Construction:
  - Mismatch transition:
    - If in state `j` and next char `c != pat[j]`, then the last `j-1` characters of input text are `pat[1]` ~ `pat[j-1]`, followed by `c`.
  - Compute `DFA[c][j]`:
    - Simulate `pat[1]` ~ `pat[j-1]` on DFA from state 0 and let `X` be the current state
    - Then `DFA[c][j] = DFA[c][X]`

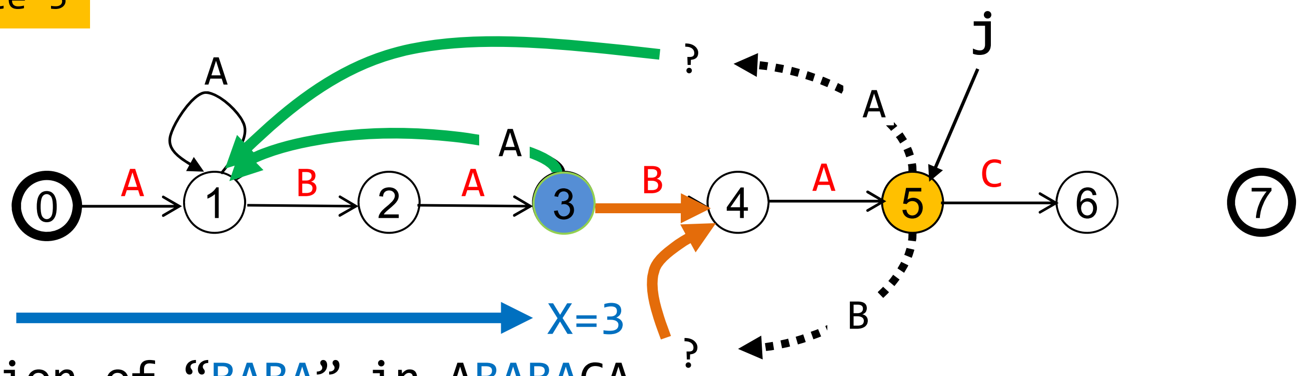
DFA Construction

- DFA Construction:
  - Mismatch transition:
    - `DFA[c][j] = DFA[c][X]`

```
DFA['A'][5] = DFA['A'][3] = 1
DFA['B'][5] = DFA['B'][3] = 4
```

Pattern: ABABACA

State 5



Simulation of "BABA" in ABABACA

DFA Construction

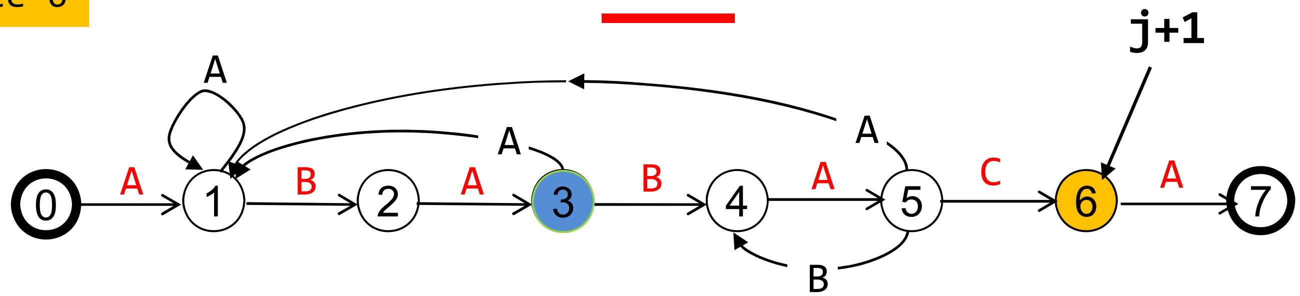
- DFA Construction:
  - Mismatch transition:
    - If in state `j` and next char `c != pat[j]`, then the last `j-1` characters of input text are `pat[1]` ~ `pat[j-1]`, followed by `c`.
  - To compute `DFA[c][j]`:
    - Simulate `pat[1]` ~ `pat[j-1]` on DFA (*still under construction*) and let the current state `X`.
    - take a transition `c` from state `X`.
    - Running time : require `j` steps.
    - But, if we maintain state `X`, it takes only constant time!

DFA Construction

- DFA Construction:
  - Maintaining state X:
    - Finished computing transitions from state `j`.
    - Now, now move to next state `j+1`.
    - Then what the new state(X′) of X be?

State 6

Pattern: ABABACA



X=3

Simulation of “BABA” in ABABACA

X′ = ?

Simulation of “BABAC” in ABABACA

- Simulation requires j+1 steps
- But, X′ = DFA[′C′][X]

X′ = DFA[′C′][3] = 0

DFA Construction

- DFA Construction: A Linear Time Algorithm
  - For each state `j`:
    - Match case: set `DFA[pat[j]][j]=j+1`
    - Mismatch case: Copy `DFA[][X]` to `DFA[][j]`
    - Update `X`

```
int DFA[MAX_SIZE][MAX_SIZE];  /* initially all elements are 0 */
// int R;            /* text character set size */

void constructDFA(char pat[])
{

    int patLength = strlen(pat);

    DFA[pat[0]][0] = 1;

    for(int X=0, j=0; j<patLength; j++)
    {
        for(int c=0; c<R; c++)              // copy mismatch cases
            DFA[c][j] = dfa[c][X];

        DFA[pat[j]][j] = j+1;               // copy match case
        X = DFA[pat[j]][X];                 // update X
    }
}
```

DFA Construction

- DFA Construction: Example

DFA[][]

j start from 1

0123456
Pattern: ABABACA

char \ state	0	1	2	3	4	5	6	7
A	**1**	1	**3**	1	**5**	1	**7**	1
B	0	**2**	0	**4**	0	4	0	2
C	0	0	0	0	0	**6**	0	0
others	0	0	0	0	0	0	0	0

X	0	0	1	2	3	0	1	1

DFA[B][0] DFA[A][0] DFA[B][1] DFA[A][2] DFA[C][3] DFA[A][0] DFA[A][1]

X는 근접한 곳으로 가기위한..

Algorithm with DFA

- Question:
  - Text에 나타나는 모든 pattern 을 찾을 수 있는가?
    - Text: AAAAAAAAAA
    - Pattern: AAAAA
    - 해답: 0, 1, 2, 3, 4, 5

Prefix/Suffix

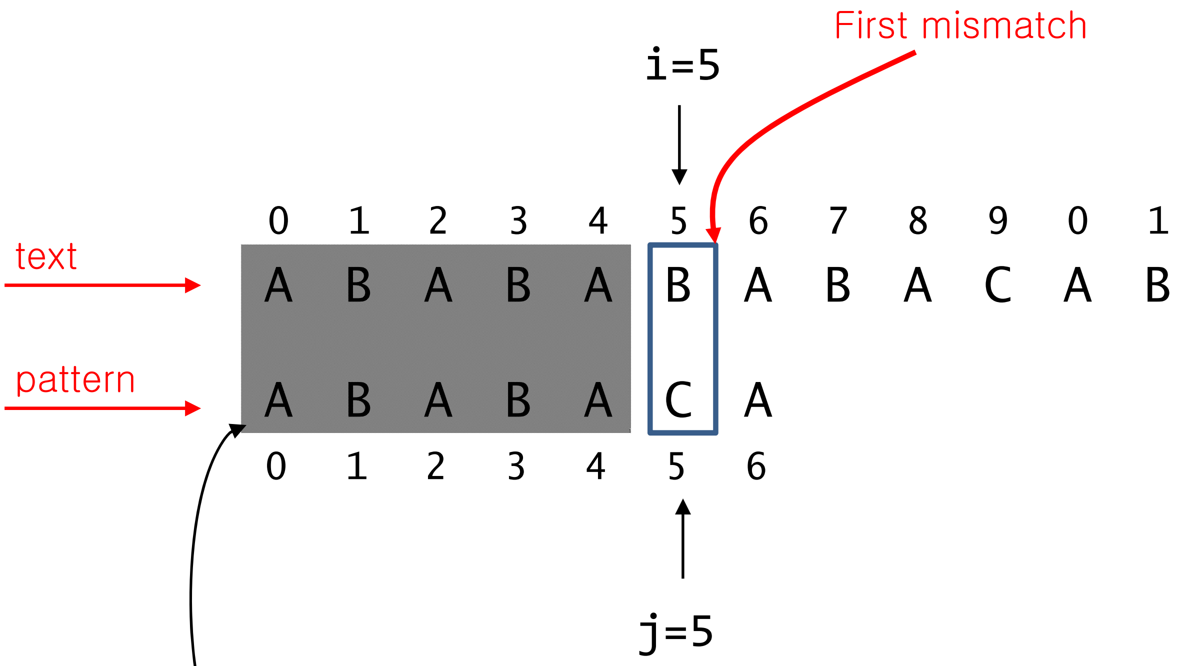
- Prefix / Suffix of a Text

bananada

	Prefix	Suffix
NULL string →	bananada	bananada
	bananada	bananada
	bananada	bananada
	bananada	bananada
	bananada	bananada
	bananada	bananada
	bananada	bananada
	bananada	bananada
	bananada	bananada

KMP Algorithm

- Naïve algorithm again:



These matched characters have very important information.
Do not waste this information!!!

KMP Algorithm

- Avoid roll-backs in naïve algorithm:
  - Example 1:

mismatch

text → ? ? A A A B ? ? ? ? ? ?

pattern → A A A B X Y Z

한칸씩옮기지말고 점프!!

avoid these roll-back steps

A A A B ? ? ?

A A A B ? ? ?

A A A B ? ? ?

First possible next position of pattern without roll-backs → A A A B ? ? ?

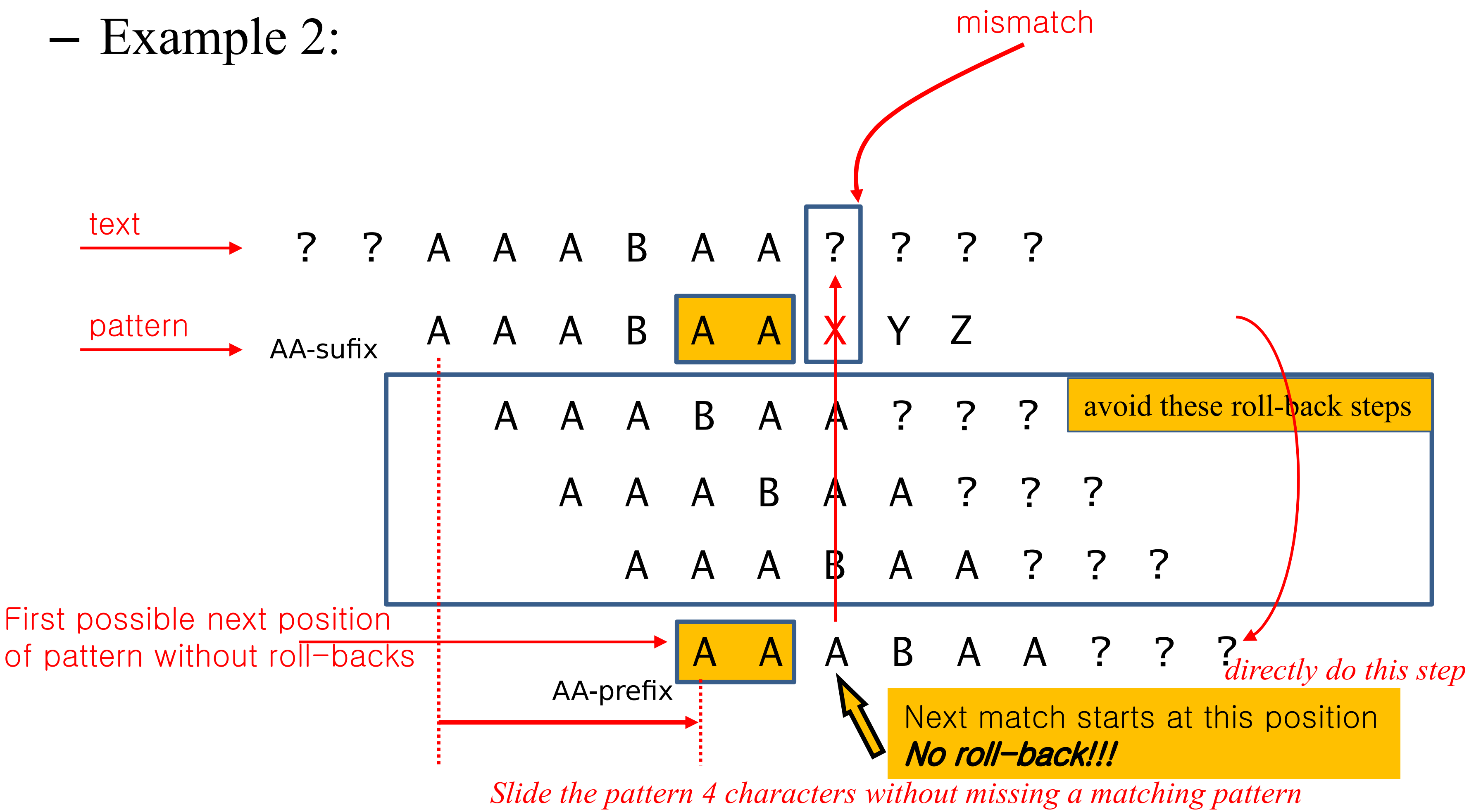
directly do this step

Next match starts at this position

Slide the pattern 4 characters without missing a matching pattern

KMP Algorithm

- Avoid roll-backs in naïve algorithm:
  - Example 2:



KMP Algorithm

- Prefix and Proper Suffix of the Prefix

pattern → A A A B A A X Y Z

$\underbrace{\text{A A A B A A}}_{\text{Prefix}}$ X Y Z

Proper Suffix of the Prefix "**AAAA**X**BAA**"

A A A B A A X Y Z

A A A B A A X Y Z

A A A B A A X Y Z

A A A B A A X Y Z

A A A B A A X Y Z

A A A B A A X Y Z

Not a ***proper*** suffix of the prefix (the same as the prefix)

KMP Algorithm

- Maximum Overlap of a Prefix
  - the longest proper suffix that is equal to prefix of the prefix

pattern

A A A B A A X Y Z

Prefix

Maximum Overlap of a Prefix "**AAAABAA**"

A A A B A A

A A A B A A

Proper Suffix of the Prefix "**AAAABAA**"

A A A B A A X Y Z

A A A B A A X Y Z

A A A B A A X Y Z

A A A B A A X Y Z

A A A B A A X Y Z

KMP Algorithm

- Maximum Overlap of a Prefix
  - the longest proper suffix that is equal to prefix of the prefix

- Example:

proper suffix = 가장긴 prefix에 prefix

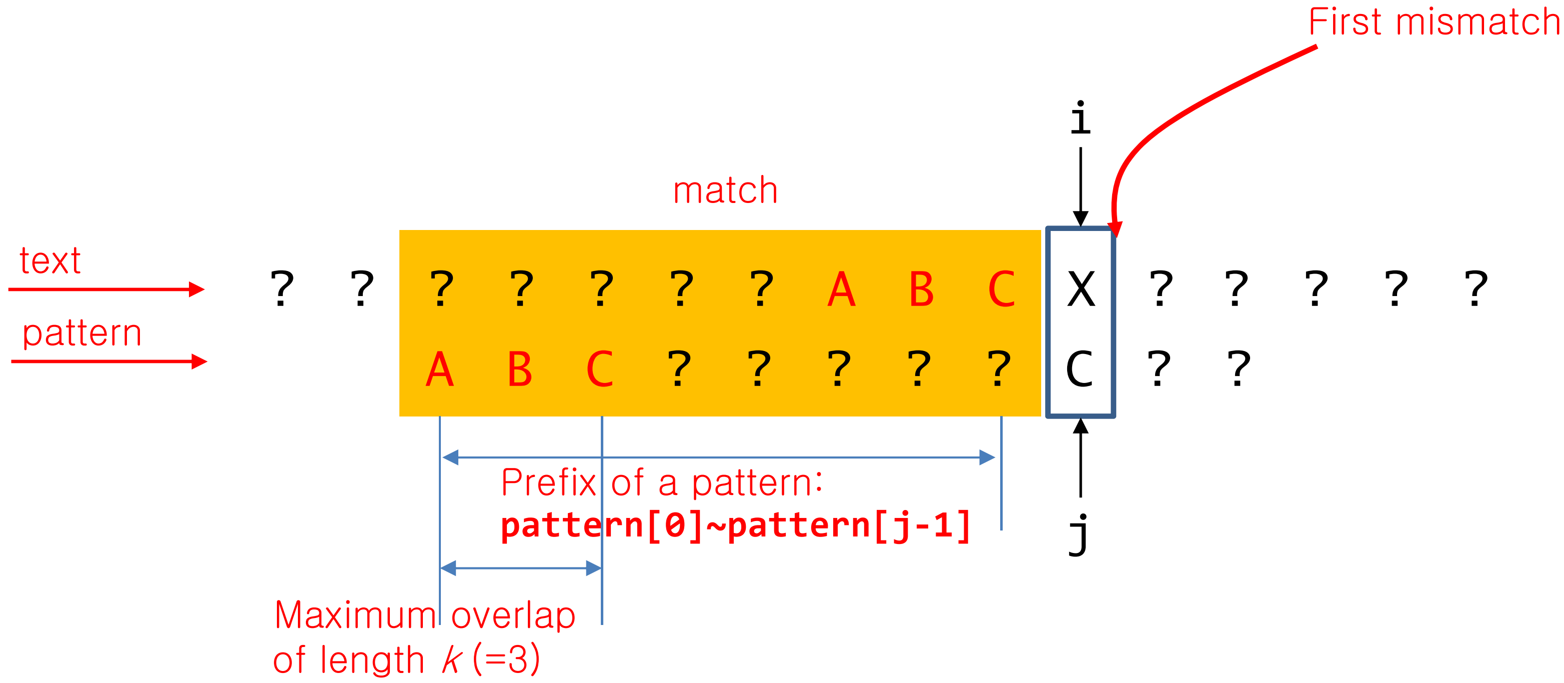
Prefix	Maximum Overlap
AAAAA	AAAA
AABA	A
AAAB	
ABABABAB	ABABAB

not **AAAAA**

NULL String

KMP Algorithm

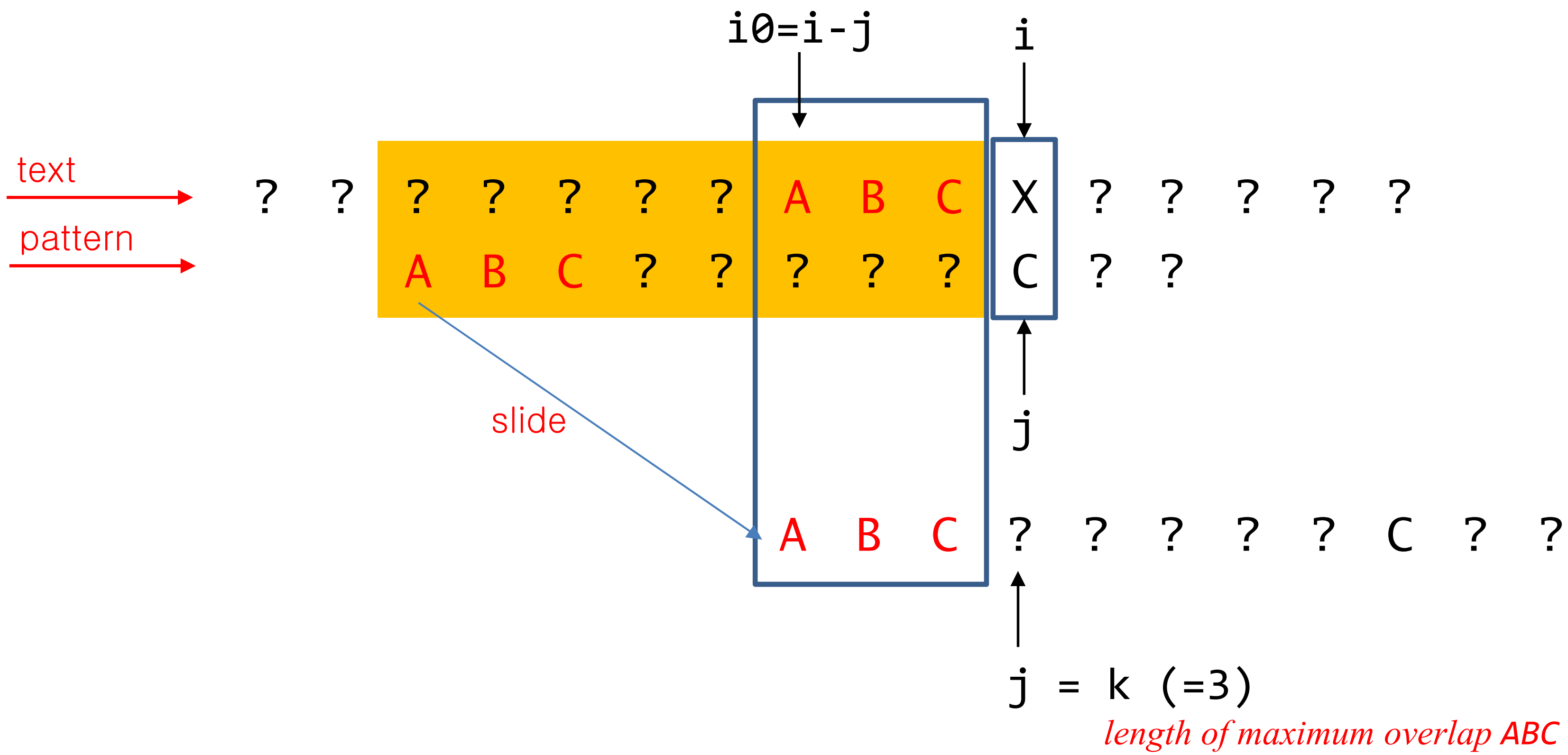
- Reuse of prefix information when there is a mismatch
  - Mismatch at `text[i]` and `pattern[j]`



Note that if the mismatched *location* is `pattern[j]`, then *prefix* is: `pattern[0]~pattern[j-1]`

KMP Algorithm

– Then we can slide the pattern so that the *suffix and prefix aligns without missing out on a match*:



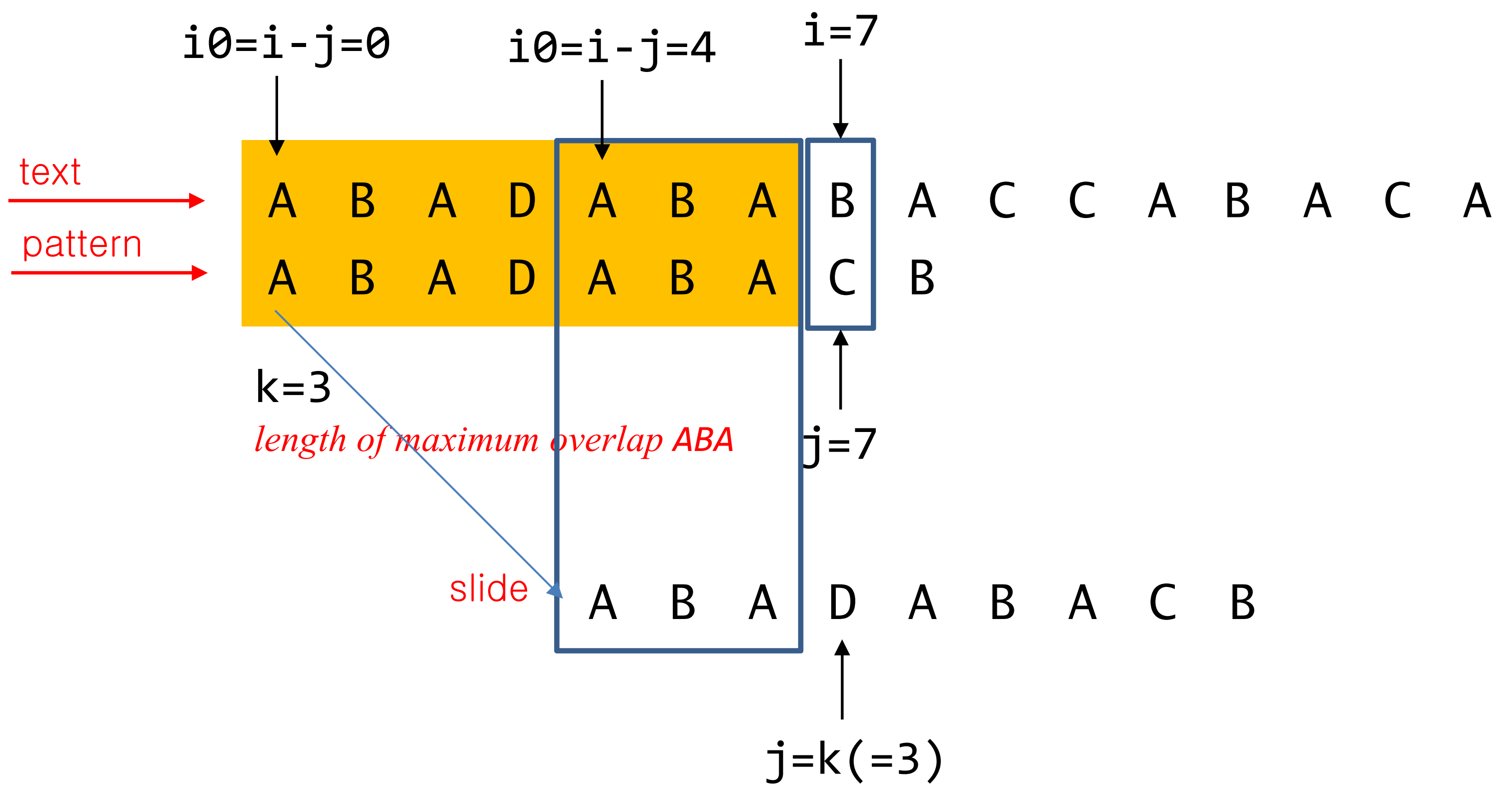
KMP Algorithm

- Fast sliding algorithm: 어느정도 가서 매칭하는 알고리즘
  - Psuedo program:

```
// mismatch found at text[i], pattern[j]

prefix = pattern[0] ~ pattern[j-1];

k = Length of maximum overlap of prefix;

j = k;
// i is unchanged !

// Matched position i0 in text starts from (i - j);
i0 = i – j;
```

KMP Algorithm

- Fast sliding algorithm:
  - Example:



KMP Algorithm

- Failure function:
  - m : the length of a pattern

페일펑션 :

  - For $0 < k < m$, the failure function $fail(k)$ is the length of maximum overlap of a prefix `pattern[0]` ~ pattern[k]
    - Note that $fail(0) = 0$

banabana

k	prefix	$fail(k)$
0	b	0
1	ba	0
2	ban	0
3	bana	0
4	banab	1
5	banaba	2
6	banaban	~~0~~ => 3
7	banabana	4

KMP Algorithm

- Knuth-Morris-Pratt(KMP) Algorithm

```
vector<int> kmp(string text, string pattern)
{
    vector<int> ans;
    fail = getFail(pattern);         // failure function
    int n = (int) text.size(), m = (int) pattern.size();
    int j =0;                        // j : index of pattern

    for(int i = 0 ; i < n ; i++)     // i : index of text
    {
        while(j>0 && text[i] != pattern[j])
            j = fail[j-1];
        if(text[i] == pattern[j])
            if(j==m-1)               // pattern matching is found
            {
                ans.push_back(i-j); // save the matched position
                j = pi[j];
            }
            else
                j++;
    }
    return ans;
}
```

KMP Algorithm

– Example:

i=0

text → A B A B A B A C A C A B A C A B

pattern → A B A C A B

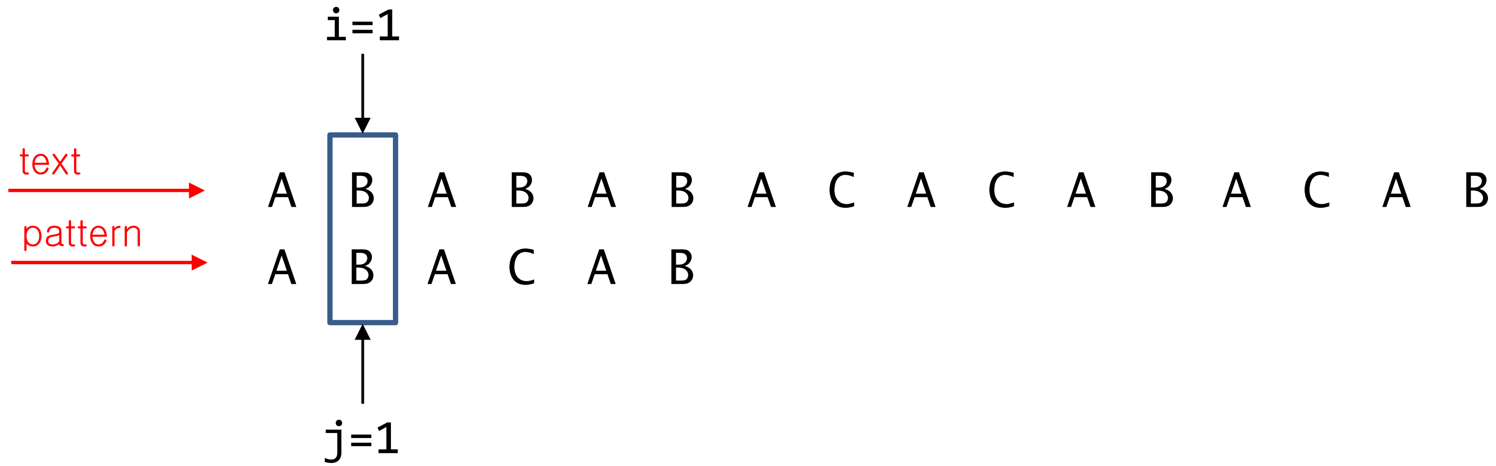
j=0

```
text[i] == pattern[j]

➔ i++, j++
```

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:




```
text[i] == pattern[j]

➔ i++, j++
```

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:

i=2

text: A B $\boxed{A}$ B A B A C A C A B A C A B

pattern: A B $\boxed{A}$ C A B

j=2

```
text[i] == pattern[j]

➔ i++, j++
```

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:

i=3

text → A B A B A B A C A C A B A C A B

pattern → A B A C A B

j=3

틀림!!! 그럼 앞에것들이 프리픽스... j를 1로...

```
text[i] != pattern[j]

➔ j = fail[j-1]
```

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:

i=3

text A B A B A B A C A C A B A C A B

pattern A B A C A B

j=1

`text[i] == pattern[j]`

➔ `i++, j++`

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:

i=4

text A B A B A B A C A C A B A C A B

pattern A B A C A B

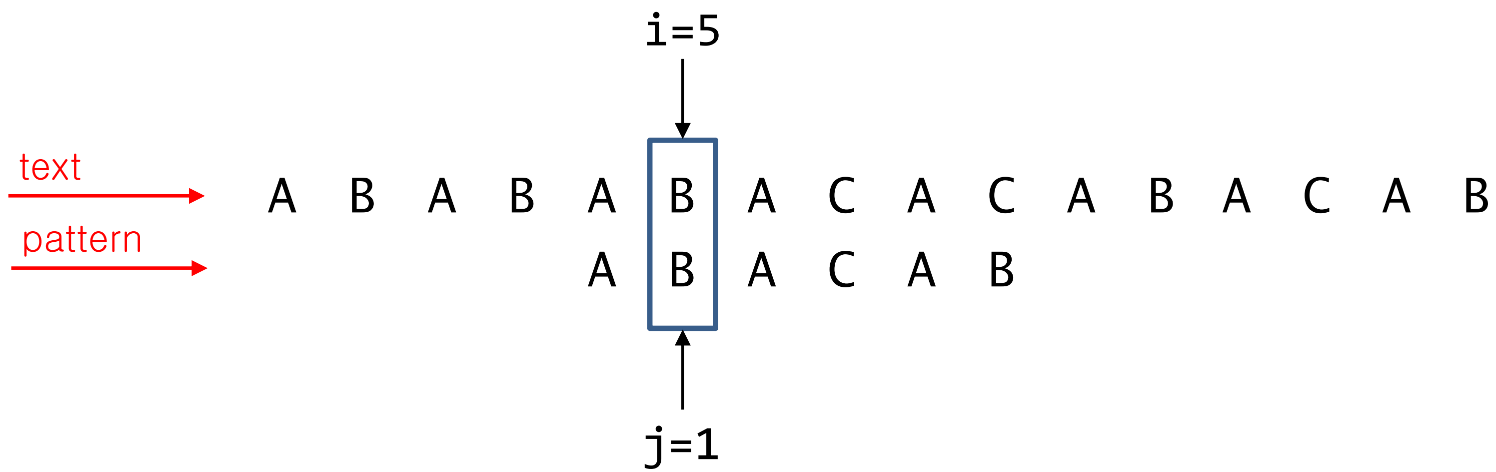
j=2

```
text[i] == pattern[j]

➔ i++, j++
```

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:



text[i] == pattern[j]

➔ i++, j++

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:

i=9

text A B A B A B A C A C A B A C A B

pattern A B A C A B

j=5

```
text[i] == pattern[j]

➔ i++, j++
```

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:

i=9

text → A B A B A B A C A $\boxed{C}$ A B A C A B

pattern → A $\boxed{B}$ A C A B

j=1

```
text[i] != pattern[j]

➔ j = fail[j-1]
```

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMU 국민대학교 KOOKMIN UNIVERSITY

Algorithmic Engineering Laboratory

KMP Algorithm

– Example:

i=9

text A B A B A B A C A C A B A C A B

pattern A B A C A B

j=0

`j==0 && text[i] != pattern[j]`

➔ `i++`

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:

i=10

text: A B A B A B A C A C A B A C A B

pattern: A B A C A B

j=0

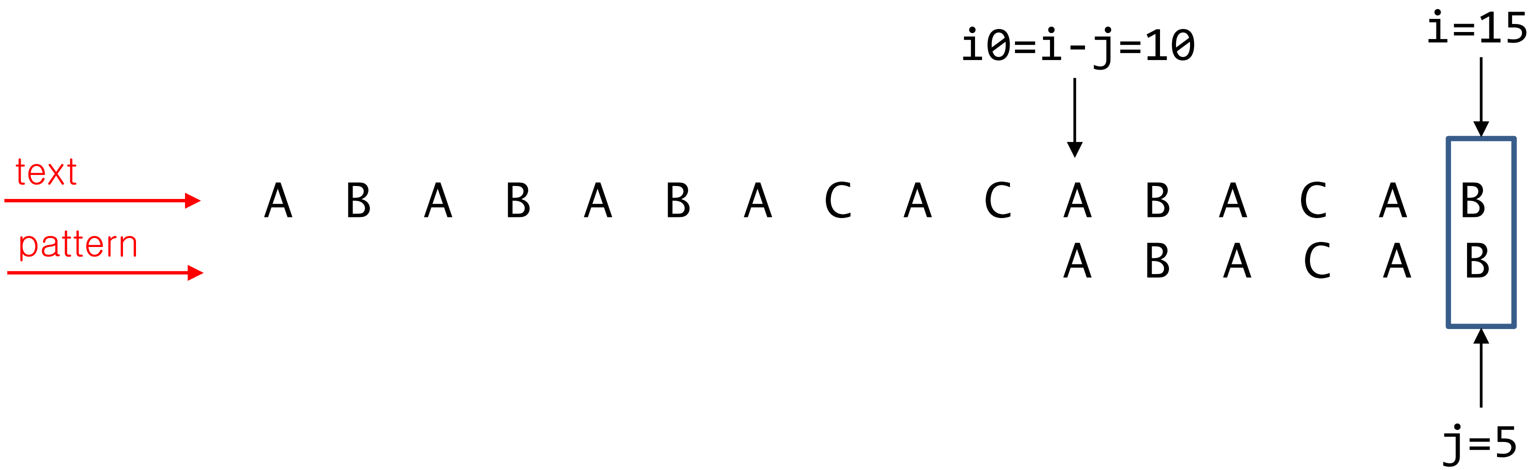
`text[i] == pattern[j]`

➔ `i++, j++`

i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

– Example:



i	0	1	2	3	4	5
fail(i)	0	0	1	0	1	2

KMP Algorithm

- Knuth-Morris-Pratt(KMP) Algorithm
  - Why?

```
while(j>0 && text[i] != pattern[j])
        j = pi[j-1];
```

i=0

text	A	B	A	B	A	B	A	C	A	B	A	B	A	B	A	B
pattern	A	B	A	B	A	B	A	B								

j=0

text[i] == pattern[j]

➔ i++, j++

i	0	1	2	3	4	5	6	7
fail(i)	0	0	1	2	3	4	5	~~0~~ => 6임

KMP Algorithm

- Knuth-Morris-Pratt(KMP) Algorithm
  - Why?

```
while(j>0 && text[i] != pattern[j])
        j = pi[j-1];
```

i=7

text → A B A B A B A C A B A B A B A B

pattern → A B A B A B A B

j=7

6(j-1인듯) 에 해당하는 5를 가져옴
패턴의 5번(0부터)을 C에 매칭

text[i] != pattern[j]

➔ j = fail[j-1]

i	0	1	2	3	4	5	6	7
fail(i)	0	0	1	2	3	4	5	0

KMP Algorithm

- Knuth-Morris-Pratt(KMP) Algorithm
  - Why?

```
while(j>0 && text[i] != pattern[j])
        j = pi[j-1];
```

i=7

text → A B A B A B A C A B A B A B A B

pattern → A B A B A B A B

j=5

text[i] != pattern[j]

➔ j = fail[j-1]

i	0	1	2	3	4	5	6	7
fail(i)	0	0	1	2	3	4	5	0

KMP Algorithm

- Knuth-Morris-Pratt(KMP) Algorithm
  - Why?

```
while(j>0 && text[i] != pattern[j])
            j = pi[j-1];
```

i=7

text → A B A B A B A C A B A B A B A B

pattern → A B A B A B A B

j=3

text[i] != pattern[j]

➔ j = fail[j-1]

i	0	1	2	3	4	5	6	7
fail(i)	0	0	1	2	3	4	5	0

KMP Algorithm

- Knuth-Morris-Pratt(KMP) Algorithm
  - Why?

```
while(j>0 && text[i] != pattern[j])
          j = pi[j-1];
```

i=7

text → A B A B A B A C A B A B A B A B

pattern → A B A B A B A B (aligned with text from i=6)

j=1

text[i] != pattern[j]

➔ j = fail[j-1]

i	0	1	2	3	4	5	6	7
fail(i)	0	0	1	2	3	4	5	0

KMP Algorithm

- Knuth-Morris-Pratt(KMP) Algorithm
  - Why?

```
while(j>0 && text[i] != pattern[j])
          j = pi[j-1];
```

i=7

text A B A B A B A C A B A B A B A B

pattern A B A B A B A B

j=0

text[i] != pattern[j]

➔ j = fail[j-1]

i	0	1	2	3	4	5	6	7
fail(i)	0	0	1	2	3	4	5	0

KMP Algorithm

- Knuth-Morris-Pratt(KMP) Algorithm
  - Why?

곁이는 부분이 많은 것들을 계산!!?!?!?!?!

```
while(j>0 && text[i] != pattern[j])
            j = pi[j-1];
```

i=8

text	A	B	A	B	A	B	A	C	A	B	A	B	A	B	A	B
pattern									A	B	A	B	A	B	A	B

j=0

j==0 && text[i] != pattern[j]

➔ i++

i	0	1	2	3	4	5	6	7
fail(i)	0	0	1	2	3	4	5	0