

String Matching

15-750

3/15/10

v2

Input: Finite Alphabet $\Sigma = \{A, B, C, \dots\}$

Def $\Sigma^* \equiv$ All finite strings over Σ

$P \in \Sigma^*$: pattern & $S \in \Sigma^*$: text or subject

$|P| = m$

$|S| = n$

Goal: Does P occur in S ?

eg: $P \equiv$ A B A B C
 $\downarrow \downarrow \downarrow \downarrow \downarrow$
 $S \equiv$ A B A B A B C C A
 $\downarrow$
 A B C B C
 A B A B C $\leftarrow$ match

Applications: Text, Protein Seq, ...

Naive String Matching

$$P = P_1 \dots P_m \quad S = S_1 \dots S_n$$

Naive(P, S)

For $i = 0$ to $n - m$

Test if P occurs at i

$O(nm)$ time

worst case example.

Def $A^i \equiv \underbrace{A \dots A}_i$

$$P = A^{m-1}B \quad S = A^n$$

String Matching Using Finite Automaton

$$5\text{-tuple} \equiv (Q, q_0, A, \Sigma, \delta)$$

$Q \equiv$ finite set of states

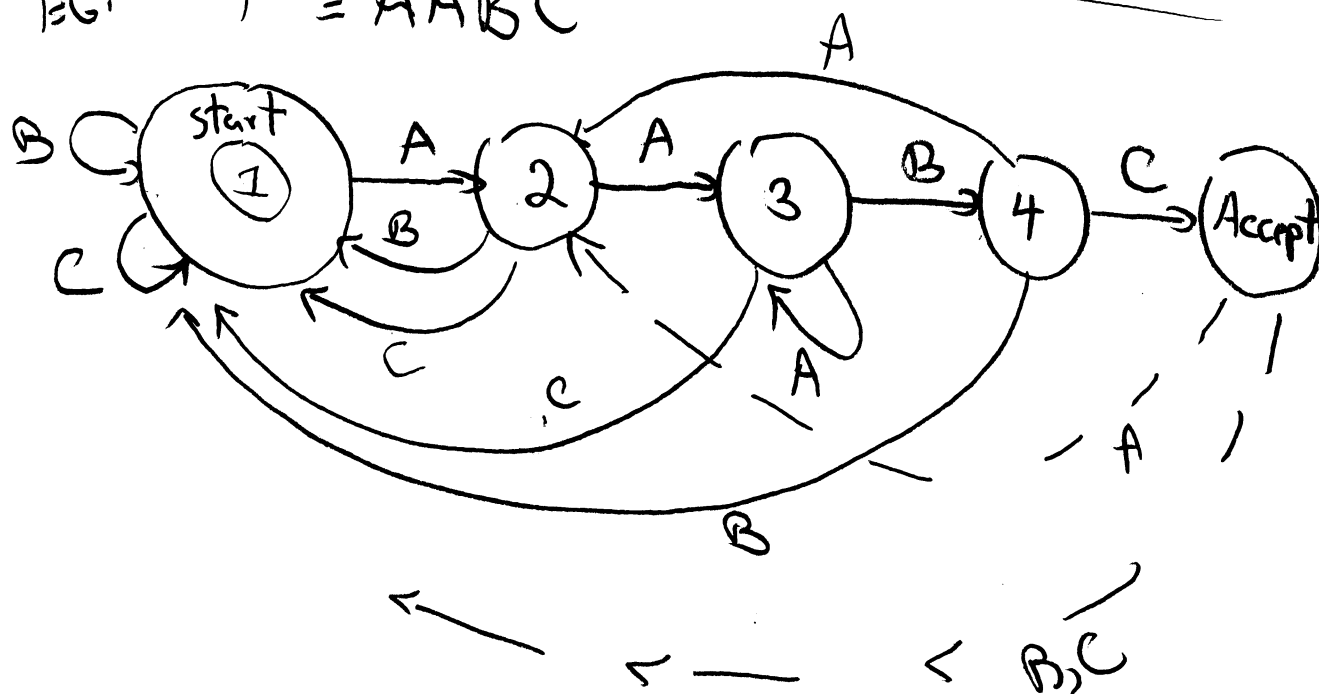
$q_0 \in Q$ start state

$A \subseteq Q$ accepting states

Σ alphabet

$$\delta: Q \times \Sigma \Rightarrow Q$$

EG: $P \equiv AABC$



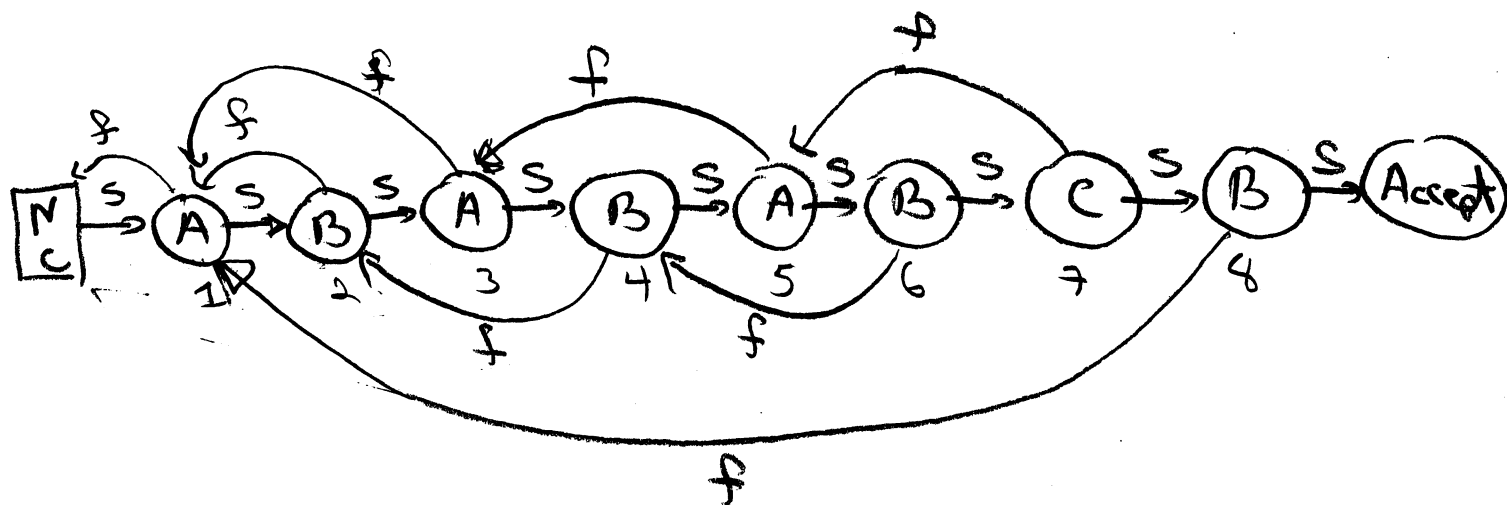
2 Questions:

- 1) How do we compute FA?
- 2) Size grows with alphabet size?

Answer Knuth-Morris-Pratt

FA: edges will be either $s \equiv$ character match
 $f \equiv$ character mismatch

$P \equiv ABABABCB$



es. $Flink(6) = 4$

ABABABCB
 ABABA

Output: Flink $\equiv$ array of failure links.

Properties: $\text{Flink}(i) = j$

1) $j < i$

2) $P_1 \dots P_{j-1} \equiv P_{i-(j-1)} \dots P_{i-1}$

3) j is largest integer satisfying 1) & 2)
in min shift

$$\text{Shift}(i) = i - \text{Flink}(i)$$

Note: $\text{Flink}(1) = 0$ & $\text{Flink}(i) \geq 1$ for $i > 1$.

KMP Flowchart

Input! $P \quad |P|=m$

Output! Array Flink

Alg: KMP Flow chart

1) $Flink(1) = 0$; $i = 2$

2) while $i \leq m$ do

$j = Flink(i-1)$

while $j \neq 0$ & $P_j \neq P_{i-1}$ (*)
do $j = Flink(j)$

$Flink(i) = j+1$

$i = i+1$

note (*) $P_j = P_{i-1}$ success

$P_j \neq P_{i-1}$ failure

KMP Scan Alg

Input: P, S $|P|=m$ & $|S|=n$

Flink

Alg: KMP Scan

1) $i=1$; $j=1$ i (subject) j (pattern)

2) while $i \leq n$ do

a) while $j \neq 0$ & $P_j \neq S_i$

do $j = \text{Flink}(j)$

b) if $j=m$ then return "Match"

else do $i=i+1$; $j=j+1$

3) return "No Match"

Correctness of KMP Flowchart

Claim: Flink satisfies 1), 2), 3)

pt induction on i

$i=1$ Flink(1) = 0 correct

assume True for $i' < i$

note: Shift(i) = Shift($i-1$) satisfies

all condition except $P_{i-1} = P_{i'-1}$

if $P_{j-1} = P_{i-1}$ done

else (we need more shifting)

"next best" $j = \text{Flink}(\text{Flink}(i-1))$

again need only check $P_{j-1} = P_{i-1}$?

Timing for Flowchart

Claim: $O(n)$ time

ie count # of comparisons
line (*)

successes $\leq m-1$

Each "failure" decrements j at least 1

Each "success" increments j exactly 1

successes $\geq$ # failures

total $\leq 2(m-1)$