String Matching

15-750
3/16/09

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

Def Σ^* ≡ All finite strings over Σ

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

$|P| = m$ $\quad\quad$ $|S| = n$

Goal: Does P occur in S?

eg: $P = ABABC$

$S = ABABCCA$

$\quad\quad ABCBC$

$\quad\quad\quad ABABC$ ← match

Applications: Text, Protein Seq, ...

Naive String Matching

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

Naive(P, S)
  For $i = 0$ to $n-m$
    Test if P occurs at i

$O(n \cdot m)$ time

worst case example.

Def $A^i \equiv \underbrace{A \cdots A}_{i}$

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

String Matching Using Finite Automaton

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

$Q \equiv$ finite set of states

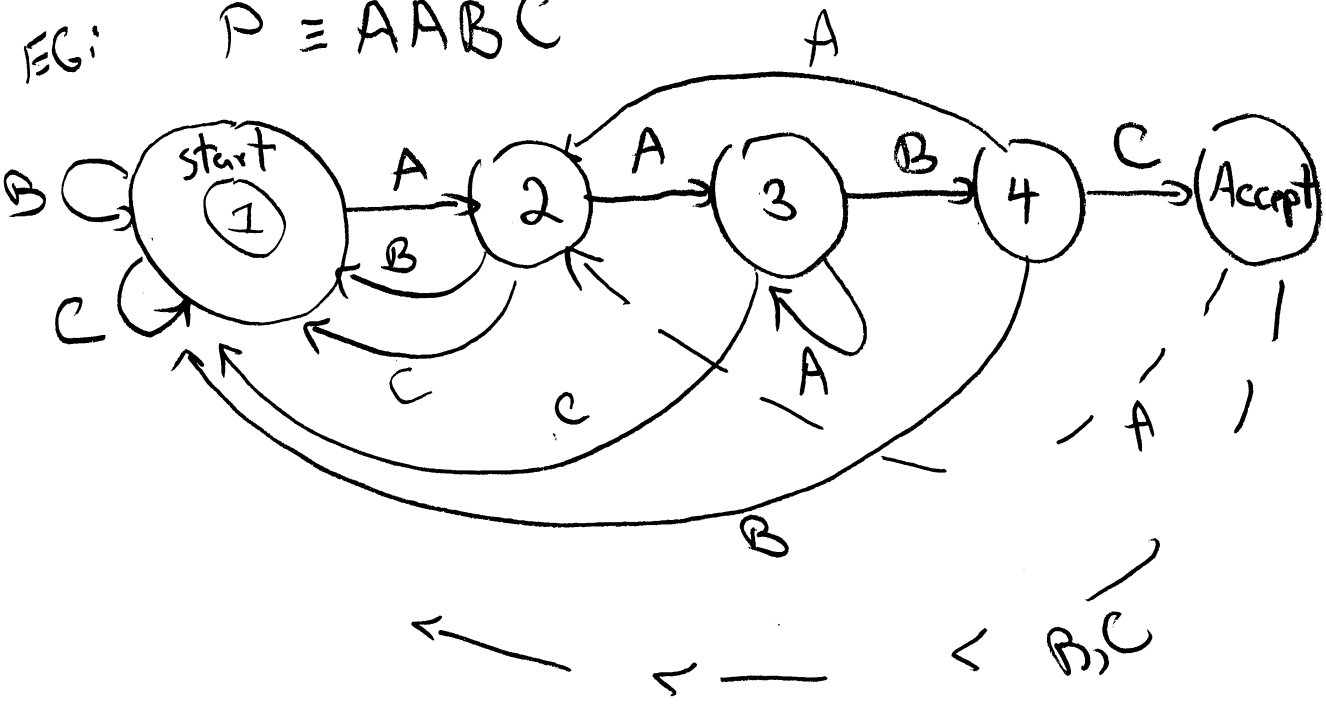
$q_0 \in Q$ start state

$A \subseteq Q$ accepting states

Σ alphebet

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

EG: $P \equiv AABC$



2 Questions'

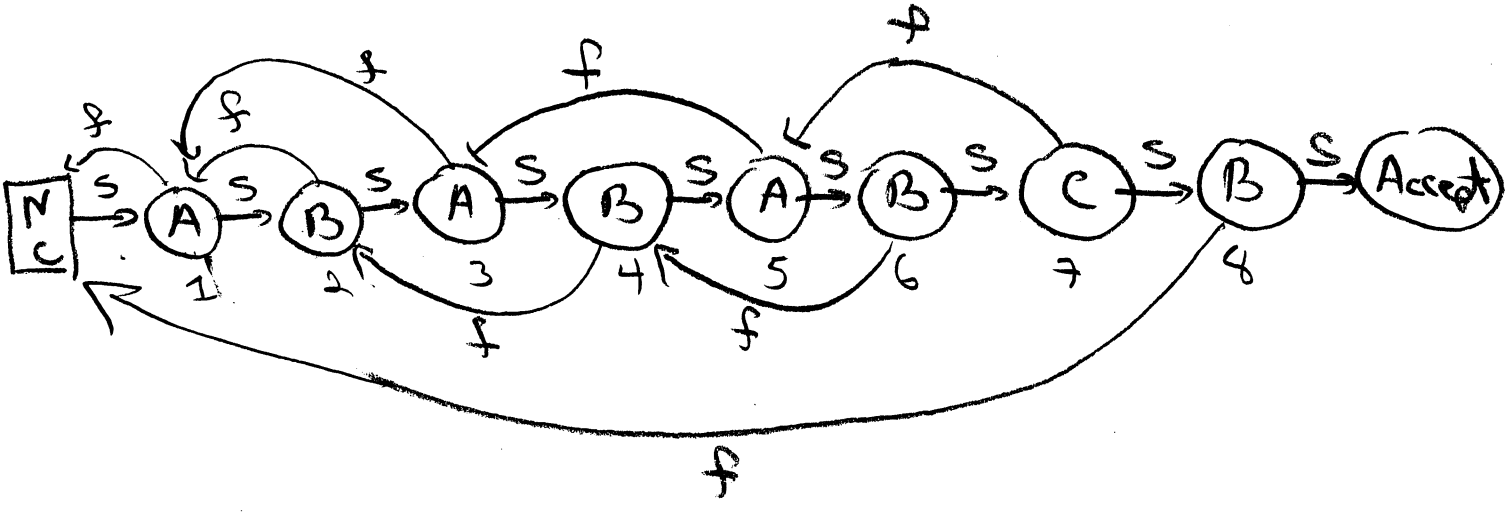
1) How do we compute FA?

2) Size grows with alphapet size?

Answer Knuth-Morris-Pratt

FA: edges will be either s ≡ character match
f ≡ character mismatch

P ≡ ABABABCB



eg. Flink(6) = 4

ABABABC
AB AB A

Output: Flink ≡ array of failure links.

Properties; $Flink(i) = j$

1) $j < i$

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

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

$$shift(i) = i - Flink(i)$$

KMP Flowchart

Input! P $|P| = m$

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 = Flink(j)

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

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

3) return "Failure"

Correctness of KMP Flowchart

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

pf induction on i

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

assume True for $i' < i$

note: Shift(i) = Shift(i-1) satisfies all condition except $P_{j-1} = P_{i-1}$

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

else (we need more shifting)

"next best" j = Flink(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 i exactly 1

successes $\geq$ # failures

total $\leq 2(m-1)$