REGULAR EXPRESSION

Dr. A. Beulah

AP/CSE

LEARNING OBJECTIVE

- To construct finite automata for any given pattern and find its equivalent regular expressions
  - To understand what is Regular Expression

INTRODUCTION

- Regular expressions describe regular languages
- ie the language accepted by a finite automata are easily described by regular expression.
- Many programming languages provide regular expression capabilities,
  - Built-in → Perl, JavaScript, Ruby, AWK, Tcl,
  - Standard library →.NET, Java, Python C++
- REs are widely supported in programming languages, text processing programs (particular lexers, lex, yacc), advanced text editors

INTRODUCTION

- Let Σ be a finite set of symbols.
- Let L_1, L_2 be set of strings in Σ^*.
- The concatenation of L_1 and L_2 denoted by $L_1 L_2$ is the set of all strings of the form *xy*, where $x \in L_1$ and $y \in L_2$.
- $L^0 = \{\varepsilon\}$
- $L^i = LL^{i-1}$ for $i \geq 1$.

$L_1 \cup L_2 = \{x, \; x \in L_1 \text{ or } x \in L_2\}$

$L_1 \cdot L_2 = \{xy \;\; x \in L_1 \;\; y \in L_2\}$

L_1^* $\quad L^* = \bigcup_{i=0}^{\infty} L^i$

INTRODUCTION

- Kleene Closure

$$L^* = \bigcup_{i=0}^{\infty} L^i = L^0 \cup L^1 \cup L^2 \cup \ldots$$

- Positive Closure

$$L^+ = \bigcup_{i=1}^{\infty} L^i = L^1 \cup L^2 \cup \ldots$$

EXAMPLE

Let $L_1 = \{10, 01\}$, $L_2 = \{11, 00\}$

Then $L_1L_2 = \{1011, 1000, 0111, 0100\}$

$L_1 \cup L_2 = \{00\ 11, 10, 01\}$

Let $L = \{10, 11\}$

Then $L^* = L_0 \cup L_1 \cup L_2 \cup \ldots\ldots$

$= \{\varepsilon\} \cup \{10, 11\} \cup \{1011, 1010, 1110, 1111\} \cup \ldots$

$= \{\varepsilon, 10, 11, 1011, 1010, 1110, 1111, \ldots\ldots\}$

OPERATORS OF RE

* * → L^*
* . → $L_1 . L_2$, L_1L_2
* / → $L_1 \cup L_2$

r_1 r_2 → RE

Union / or +

r_1 / r_2 (or) $r_1 + r_2$

Con . $r_1 r_2$

$r_1 . r_2$ (or) $r_1 r_2$

Clo *

r_1^*

DEFINITION OF REGULAR EXPRESSION

- Let Σ be an alphabet. The regular expressions over Σ and the sets that they denote are defined recursively as follows:

1. ϕ is a regular expression and denotes the empty set {}. ϕ
2. ε is a regular expression and denotes the set {ε} ε
3. For each a ∈ Σ, 'a' is a regular expression and denotes the set {a}.
4. If r and s are regular expressions denoting the languages R and S respectively then (r + s), (rs), (r)* are regular expressions that denotes the sets RUS, RS and R* respectively.

$r \rightarrow R$

$s \rightarrow S$

Σ = {a, b}

a	b	+	r/s	RUS → lan
RE	RE	.	r.s	RS
		*	r*	R*

PRECEDENCE OF RE OPERATORS

* → higher precedence

.

/ → Lower precedence

EXAMPLE

- (0/1)* = {ε,0,1,00, 01, 10, 11.....} = (0+1)* (i.e.) all strings of 0 and 1
- 01* = {0, 01, 011, 0111, ...}
- 0* = {ε, 0, 00, 000, ...}
- 1(1)* = {1, 11, 111, 1111, ...) = 1^+

$(0|1)^*$ {0, 1}

$0|1^*$ {0, 1, 11, ...}

$(0|1)^* = \{\varepsilon, 0, 1, 00, 11 \ldots\}$

0 = {0}

$1^* = \{\varepsilon, 1, 11, 111, \ldots\}$

$1\,1^*$

1 = {1}

$0\,1^* = \{0, 01, 011, 0111 \ldots\}$

$1\,1^* = \{1, 11, 111, \ldots\}$

EXAMPLE

$L = \{01, 11\}$ RE? $(0/1)1$

$L = \{0^n / n \text{ is } 5\}$

$L = \{00000\}$ 0 $L = \{0\}$

RE = ?

00000

$L = \{1^m / m \text{ is divisible by } 2 \text{ or } 7\}$

$L = \{\varepsilon, 11, 1111, 111111, 1111111, \ldots\}$ RE = ? $(11)^* / (1111111)^*$

EXAMPLE

- The set of all strings over {0,1} with three consecutive 0's.

L=? { 000, ~~000~~

RE=? $(0/1)^*000(0/1)^*$

- The set of all strings over {0,1} beginning with 00.

L=? L={ 00, 00.... }

RE=? $00(0/1)^*$

EXAMPLE

- {1, 11, 111,........}

L=?

RE=?

- The set of all strings over {0, 1} which has atmost two 0's.

L=?

RE=?

$(0/1)0$ $\quad 1^*01^*01^* + 1^*01^* + 1^*$

IDENTITIES FOR REGULAR EXPRESSIONS

I1 $\phi + R = R$

I2 $\phi R = R\phi = \phi$

I3 $\lambda R = R\lambda = R$

I4 $\lambda^* = \lambda$

I5 $R + R = R$

I6 $R^*R^* = R^*$

I7 $RR^* = R^*R$

I8 $(R^*)^* = R^*$

I9 $\lambda + RR^* = R^* = \lambda + R^*R$

I10 $(PQ)^*P = P(QP)^*$

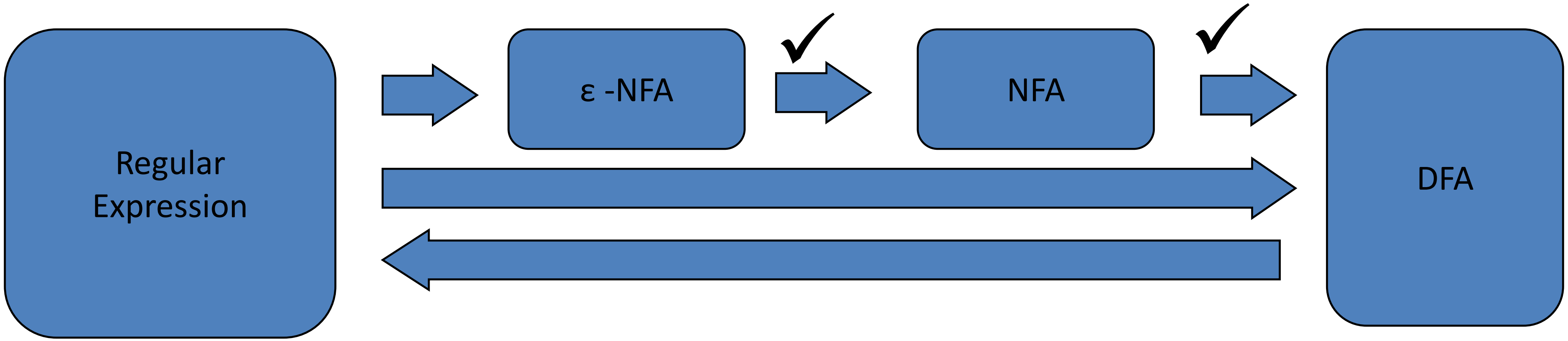
I11 $(P + Q)^* = (P^*Q^*)^* = (P^* + Q^*)^*$

I12 $(P + Q)R = PR + QR$ and $R(P + Q) = RP + RQ$

ROAD MAP

DFA
NFA
ε-NFA
RE

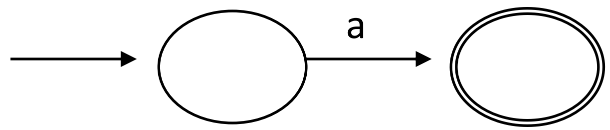
NFA → DFA



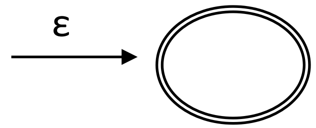
THOMPSON'S CONSTRUCTION

- Basis

R=a

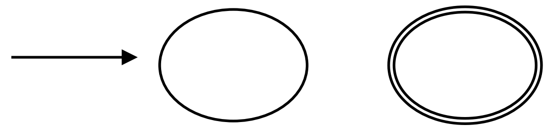


R=ε



R= ϕ

{ }

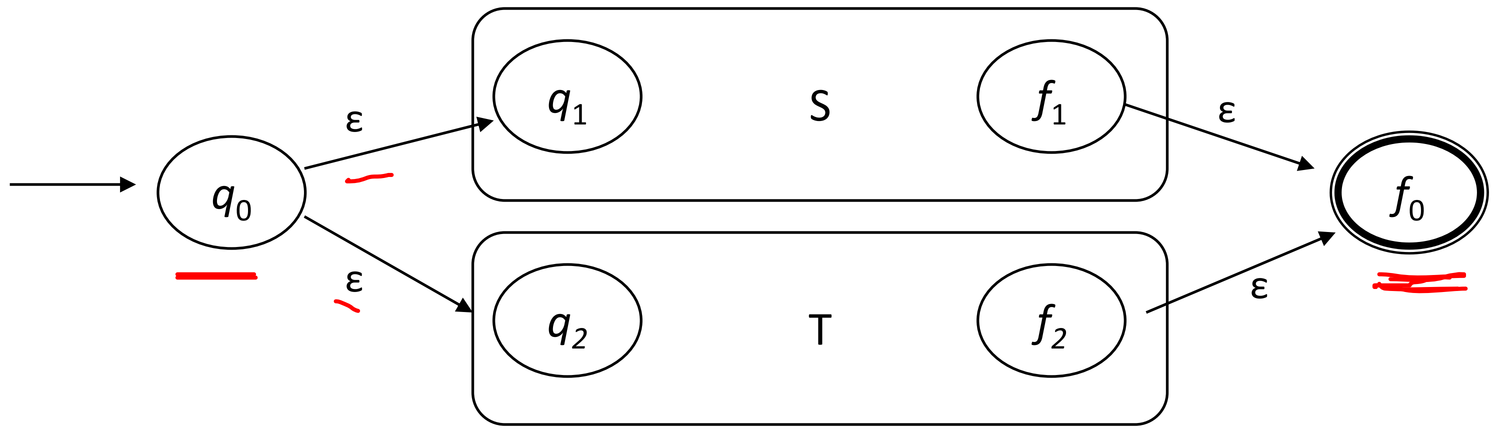


THOMPSON'S CONSTRUCTION

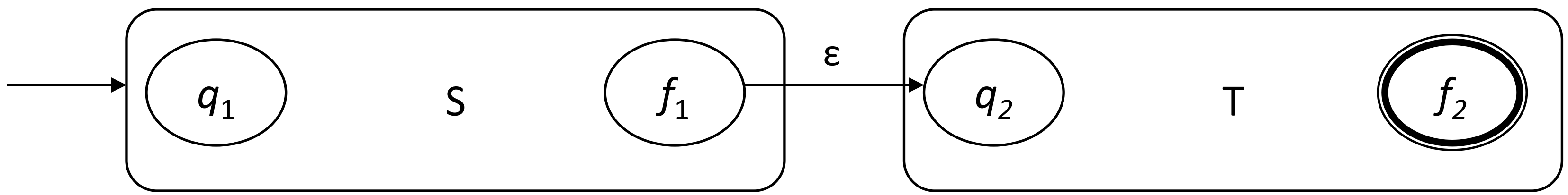
4 ii) ST

S/T

R=S+T

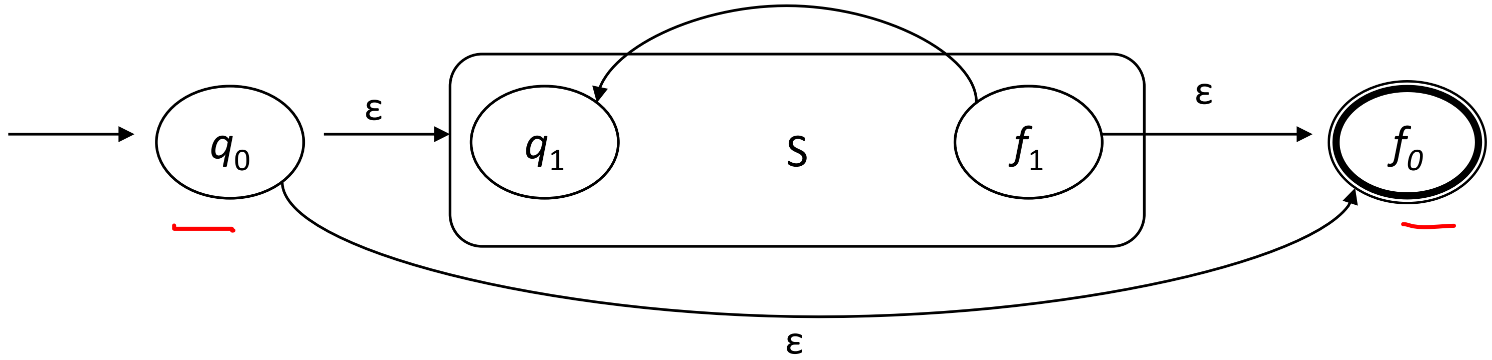


R=ST



THOMPSON'S CONSTRUCTION

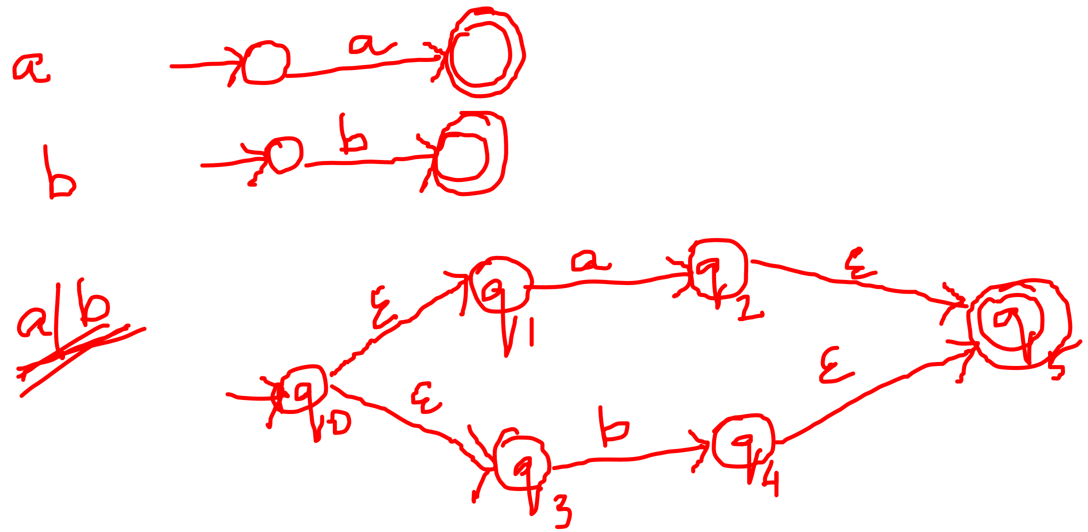
$R=S^*$



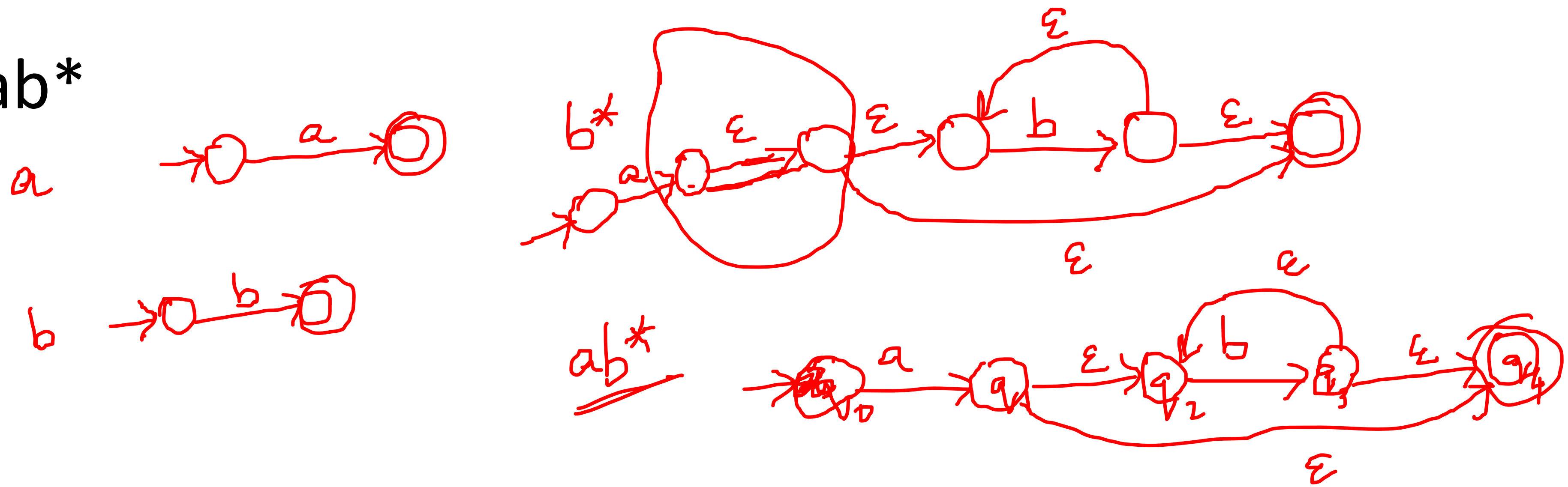
THOMPSON'S CONSTRUCTION

- a/b

a, b.

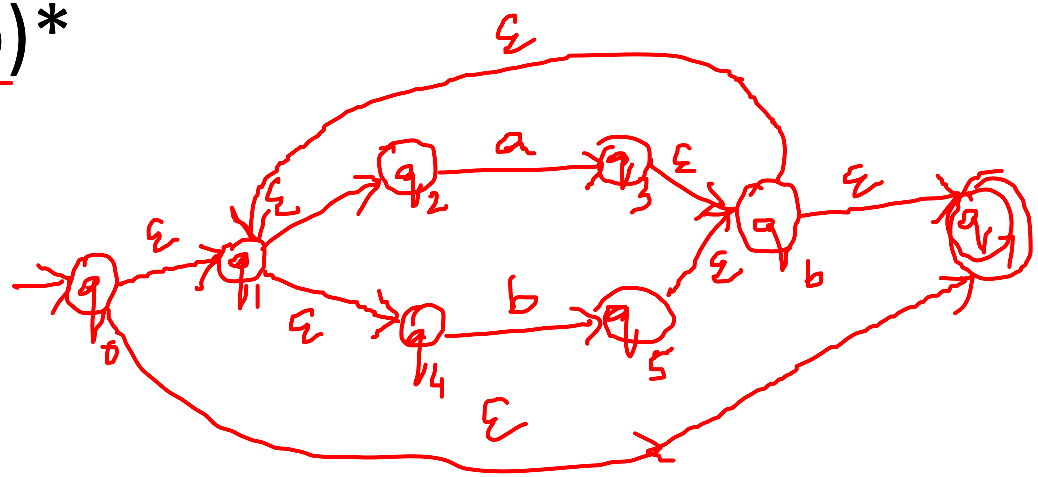


- ab*



THOMPSON'S CONSTRUCTION

- (a/b)*



$L = \{\varepsilon, a, b, ab, ba, bba \dots\}$

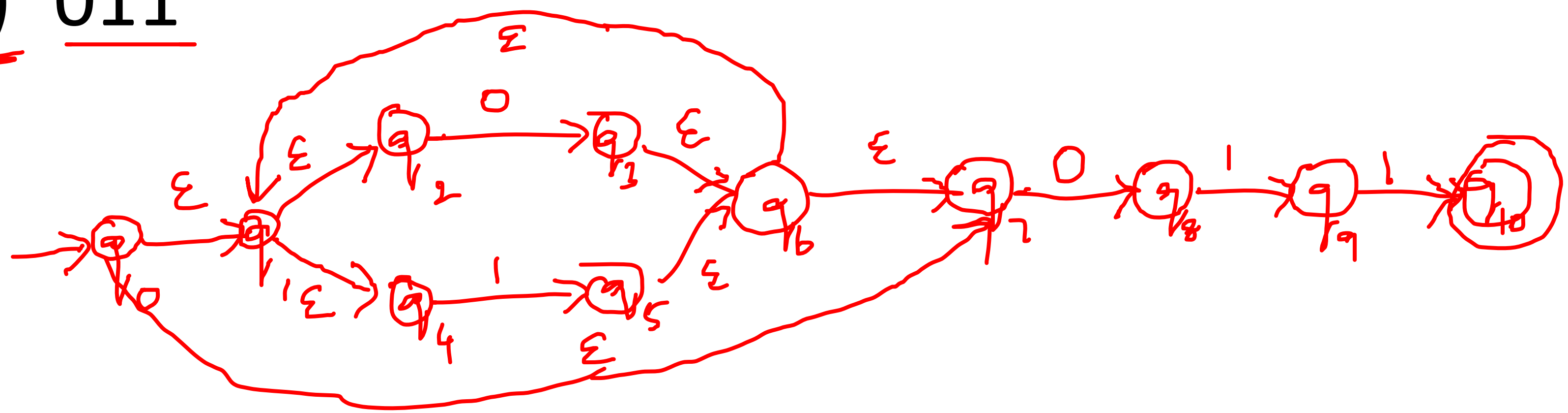
$a^* \; \{\varepsilon, a, aa$

$(0/1)^* (0/1)^*$

$10(01/0)^*$

$(11/00)(0/1)^*$

- (0/1)*011



THEOREM

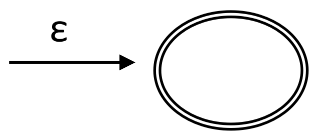
For every regular expression *r* there exists a NFA with ε-transitions that accepts L(*r*)

- Proof
  - **Basis step (Zero operators)**

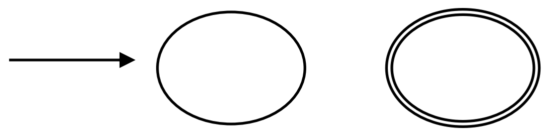
Suppose *r* is ε, ϕ or *a* for some *a* ∈ ∑.

Then the equivalent NFA's are:

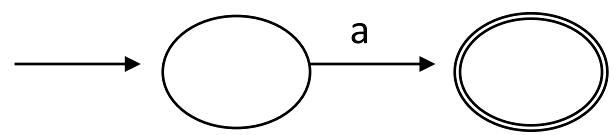
R=ε



R= ϕ



R=a



INDUCTION CASE I

- $r = r_1 + r_2$
- $M_1 = (Q_1, \sum_1, \delta_1, q_1, \{f_1\})$ $L(M_1) = L(r_1)$
- $M_2 = (Q_2, \sum_2, \delta_2, q_2 \{f_2\})$ $L(M_2) = L(r_2)$.
- Assume Q_1 and Q_2 are disjoint.
- Let q_0, f_0 be a new initial and final state respectively.

CASE I

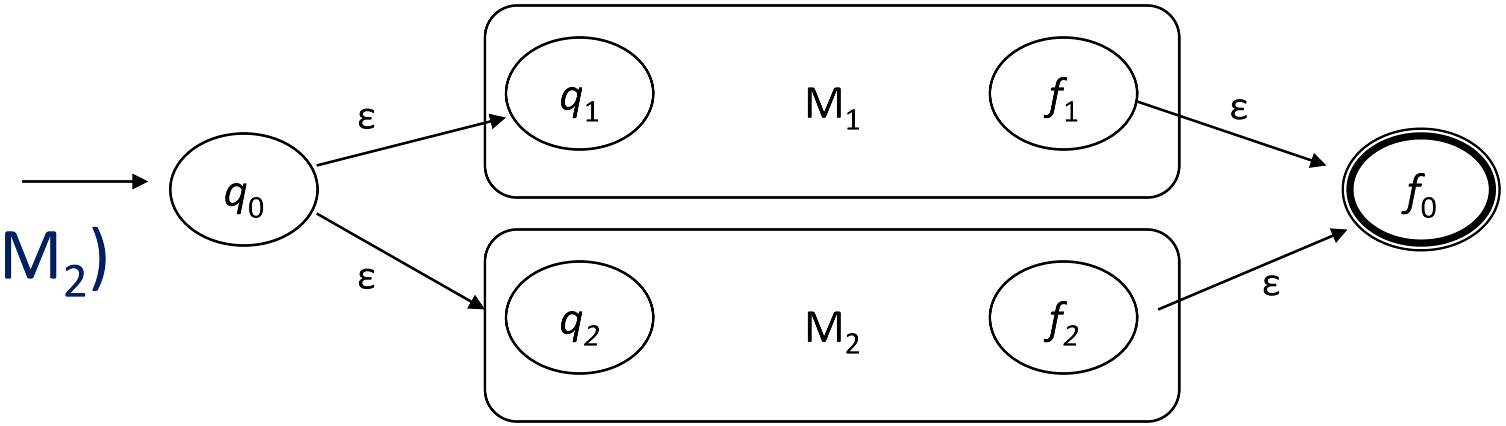
- $M = (Q_1 \cup Q_2 \cup \{q_0, f_0\}, \Sigma_1 \cup \Sigma_2, \delta, q_0, \{f_0\})$ where δ is defined by

$\delta(q_0, \varepsilon) = \{q_1, q_2\}$

$\delta(q, a) = \delta_1(q, a)$ if $q \in Q_1 - \{f_1\}$, $a \in \Sigma_1 \cup \{\varepsilon\}$

$\delta(q, a) = \delta_2(q, a)$ if $q \in Q_2 - \{f_2\}$, $a \in \Sigma_2 \cup \{\varepsilon\}$

$\delta_1(f_1, \varepsilon) = \delta_2(f_2, \varepsilon) = \{f_0\}$



- $L(M) = L(M_1) \cup L(M_2)$

CASE II

- $r = r_1 \; . \; r_2$
- $M_1 = (Q_1, \sum_1, \delta_1, q_1, \{f_1\})$ $L(M_1) = L(r_1)$
- $M_2 = (Q_2, \sum_2, \delta_2, q_2 \{f_2\})$ $L(M_2) = L(r_2)$
- $M = (Q_1 \cup Q_2, \sum_1 \cup \sum_2, \delta, \{q_1\}, \{f_2\})$, where δ is given by:

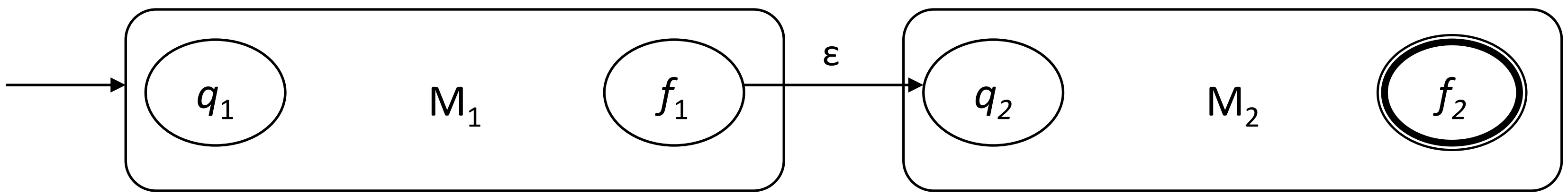
$\delta(q,a) = \delta_1(q,a)$ for q in $Q_1 - \{f_1\}$ and a in $\sum_1 \cup \{\varepsilon\}$

$\delta(f_1,\varepsilon) = \{q_2\}$

$\delta(q,a) = \delta_2(q,a)$ for q in Q_2 and a in $\sum_2 \cup \{\varepsilon\}$

CASE II

- $L(M) = \{xy \mid x \text{ is in } L(M_1) \text{ and } y \text{ is in } L(M_2)\}$
- $L(M) = L(M_1) \cdot L(M_2)$.

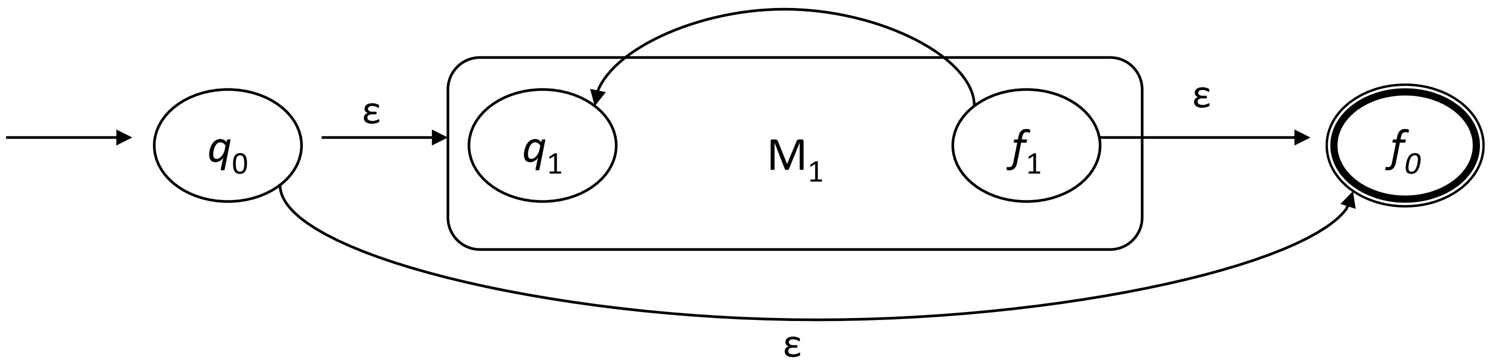


CASE III

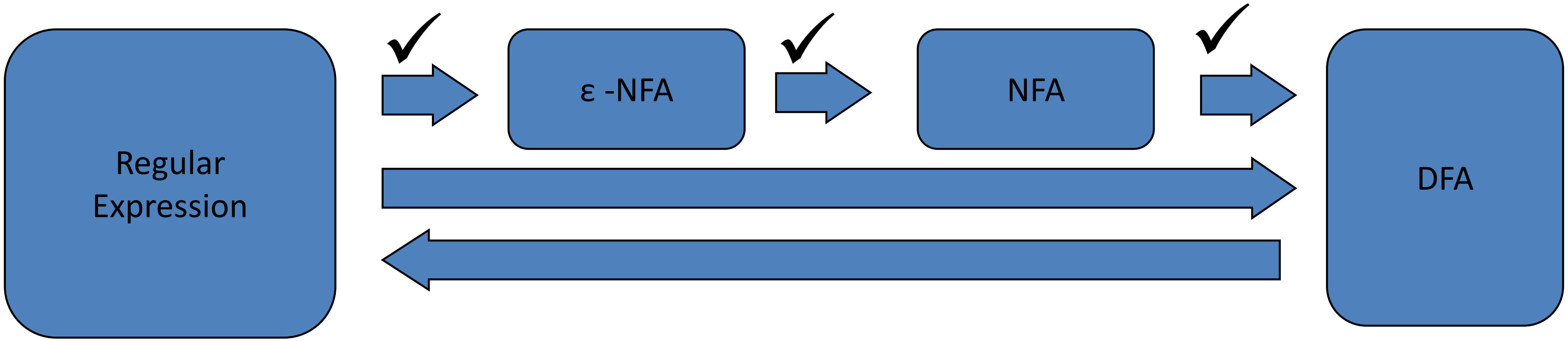
- $r = r_1^{\,*}$
- $M_1 = (Q_1, \sum_1, \delta_1, q_1, \{f_1\})$ $L(M_1) = r_1$
- $M = (Q_1 \cup \{q_0, f_0\}, \sum_1, \delta, q_0, \{f_0\})$, where δ is given by:

$\delta(q_0, \varepsilon) = \delta(f_1, \varepsilon) = \{q_1, f_0\}$

$\delta(q, a) = \delta_1(q, a)$ for q in $Q_1 - \{f_1\}$ and a in $\sum_1 \cup \{\varepsilon\}$



ROAD MAP



CONVERSION OF E-NFA TO DFA

1. Find the ε-CLOSURE of the state q_0 from the constructed ε-NFA (i.e) from state q_0, ε transition to other states are identified as well as ε transitions from other states are also identified and combined as one set (new state).

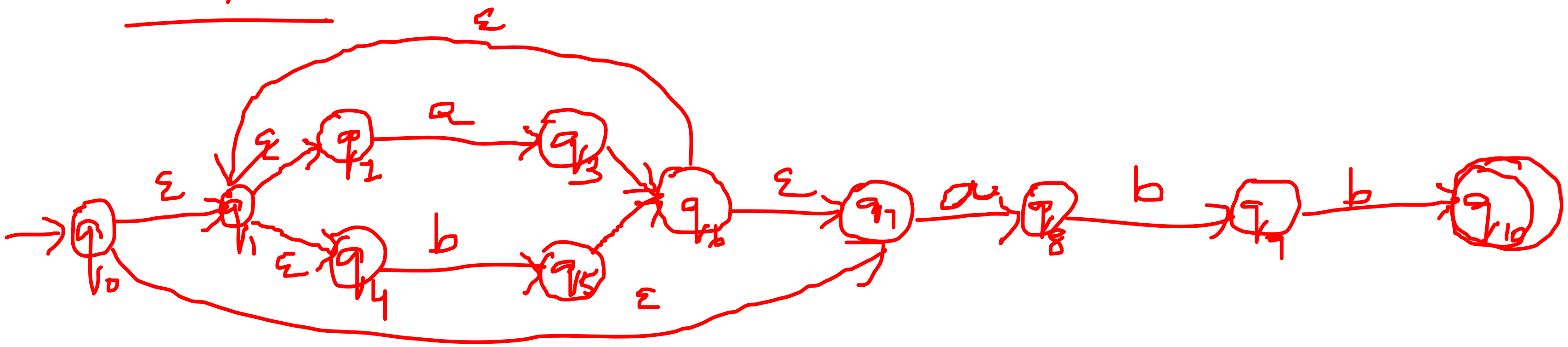
CONVERSION OF E-NFA TO DFA

2. Perform the following steps until there are no more new states as been constructed.
   i. Find the transition of the given regular expression symbols over ∑ from the new state (i.e) move (new state, symbol)
   ii. Find the ε-CLOSURE of move (new state, symbol).

EXAMPLE

(a/b)*abb

ε-NFA

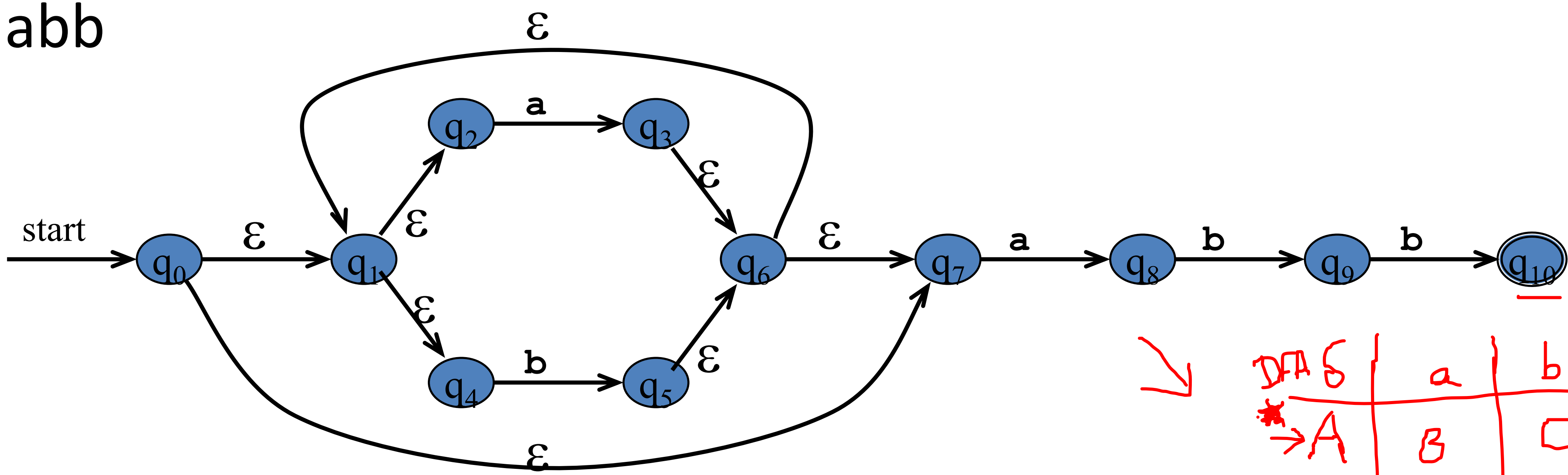


ε-NFA → DFA subset construction

ε-NFA → NFA → DFA

EXAMPLE

(a/b)*abb



$\varepsilon\text{-closure}(q_0) = \{q_0, q_1, q_2, q_4, q_7\} = A$

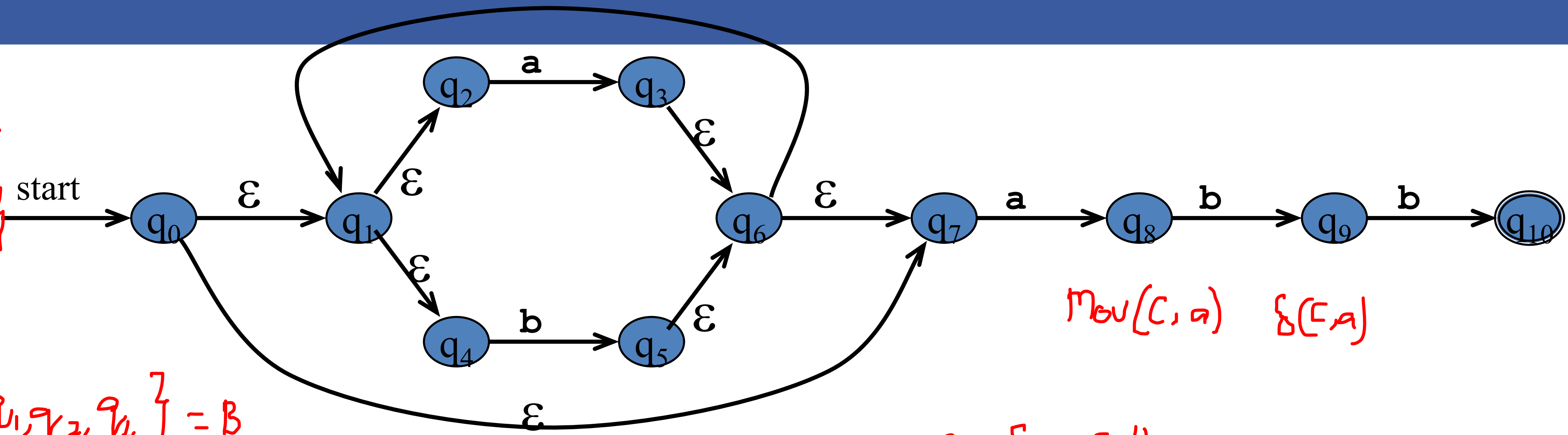
$Mov(A, a) = \{q_3, q_8\}$ | $\varepsilon\text{-closure}(mov(A,a)) = \{q_3, q_8, q_6, q_1, q_7, q_2, q_4\} = B$

$Mov(A, b) = \{q_5\}$ | $\varepsilon\text{-closure}(mov(A,b)) = \{q_5, q_6, q_1, q_7, q_2, q_4\} = C$

DFA δ	a	b
$\rightarrow A$	B	C
B	B	D
C	B	C

	a	b
D	B	E
*E	B	C

EXAMPLE



$Mov(B, a) = \{q_3, q_8\}$

$Mov(B, b) = \{q_5, q_9\}$

$\varepsilon\text{-closure}(\underline{Mov(B, a)})$

$= \{q_3, q_8, q_6, q_7, q_1, q_2, q_4\} = B$

$\varepsilon\text{-closure}(Mov(B, b)) = \{q_5, q_9, q_6, q_7, q_1, q_2, q_4\} = D$

$Mov(C, a) = \{q_3, q_8\}$ | $\varepsilon\text{-closure}(Mov(C, a)) = B$

$Mov(C, b) = \{q_5\}$ | $\varepsilon\text{-closure}(Mov(C, b)) = C$

$Mov(D, a) = \{q_3, q_8\}$ | $\varepsilon\text{-closure}(Mov(D, a)) = B$

$Mov(D, b) = \{q_5, q_{10}\}$ | $\varepsilon\text{-closure}(Mov(D, b)) = \{$

$\{q_5, q_{10}, q_6, q_1, q_7, q_2, q_4\} = E*$

$Mov(C, a)$ $\delta(C, a)$

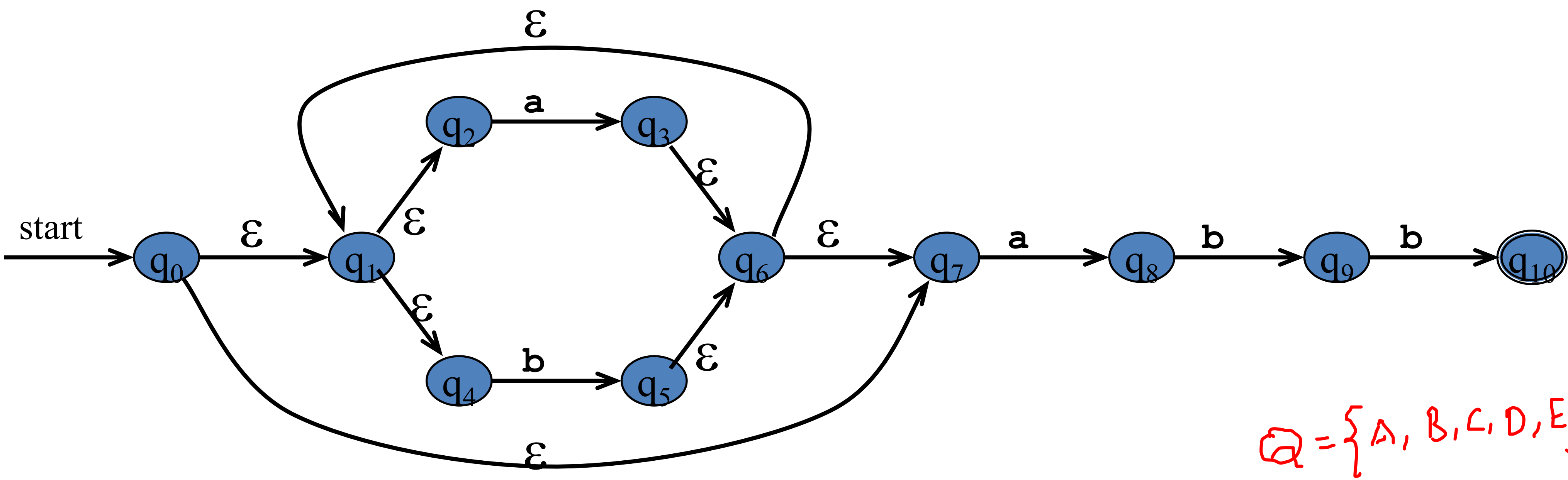
$Mov(E, a) = \{q_3, q_8\}$

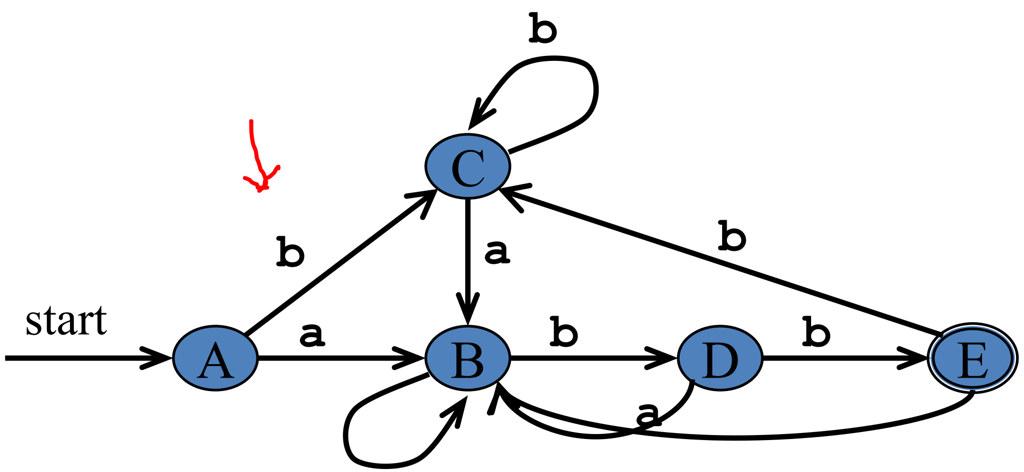
$Mov(E, b) = \{q_5\}$

$\varepsilon\text{-closure}(Mov(E, a)) = B$

$\varepsilon\text{-closure}(Mov(E, b)) = C$

EXAMPLE





Dstates

A = {0,1,2,4,7}

B = {1,2,3,4,6,7,8}

C = {1,2,4,5,6,7}

D = {1,2,4,5,6,7,9}

E = {1,2,4,5,6,7,10}

Q = {A, B, C, D, E}

Σ = {a, b}

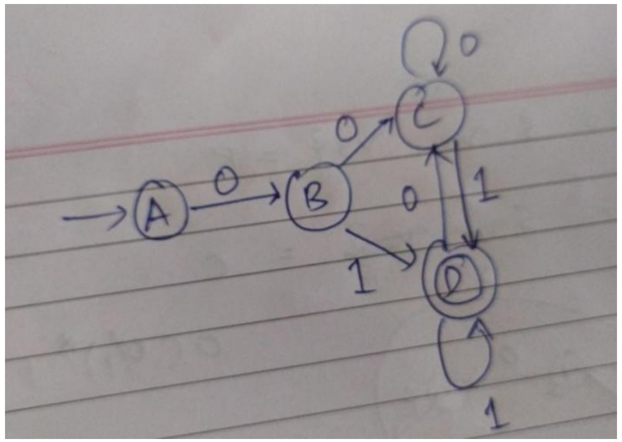
A → IS

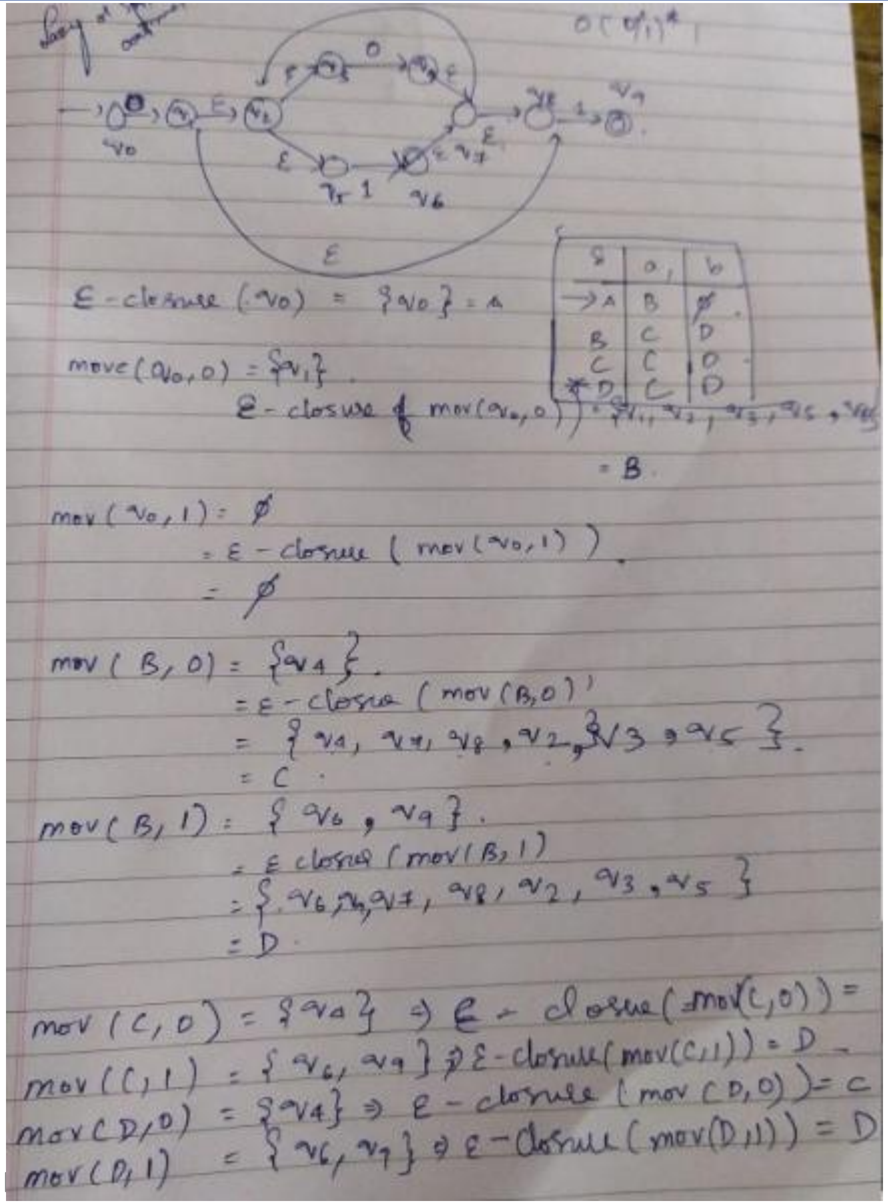
F = {E}

δ

EXAMPLE

- 0(0/1)*1





SUMMARY

- Definition of RE
- Precedence, identities, properties of RE.
- Thomson’s construction to convert RE to NFA and then to DFA

TEST YOUR KNOWLEDGE

- Which of the following does not represents the given language?
  Language: {0,01}
  a) 0+01
  b) {0} U {01}
  c) {0} U {0}{1}
  d) {0} ^ {01}

TEST YOUR KNOWLEDGE

- Regular Expression R and the language it describes can be represented as:
  a) R, R(L)
  b) L(R), R(L)
  c) R, L(R)
  d) All of the mentioned

LEARNING OUTCOME

On successful completion of this topic, the student will be able to:

- Understand the concepts of RE (K3)

REFERENCE

- Hopcroft J.E., Motwani R. and Ullman J.D, “Introduction to Automata Theory, Languages and Computations”, Second Edition, Pearson Education, 2008