ELMÉLETI INFORMATIKA

I. rész

Formális nyelvek és automaták

Reguláris kifejezések,
pumpáló lemma reguláris nyelvekre

4. előadás

Reguláris kifejezések

Az aritmetikában számok és műveletek segítségével kifejezéseket írhatunk fel, mint pl. $3.2 + 4.(1 + 6)$. Könnyen meghatározhatjuk, hogy ennek a kifejezésnek az értéke 34.

Hasonló módon a Σ ábécé elemei és a reguláris műveletek alkalmazásával ún. **reguláris kifejezés**eket tudunk felírni, amelyek Σ ábécé feletti nyelveket fognak leírni.

Ahogy az aritmetikában, úgy a reguláris kifejezéseknél is be kell tartani az ún. **precedencia-szabály**t. A reguláris műveleteknél a precedencia csökkenő sorrend szerint a következő: *iteráció*, *konkatenáció*, *egyesítés*. A zárójelek természetesen befolyásolják a precedenciát: először mindig a zárójelben található reguláris kifejezés által reprezentált nyelvet kell meghatározni.

4.1 definíció: (reguláris kifejezés)

Legyen Σ egy ábécé.

1) Az $\emptyset$, a λ és a (ahol $a \in \Sigma$) **elemi reguláris kifejezések**.
2) Ha R_1 és R_2 reguláris kifejezések, akkor $R_1 + R_2$, $R_1 R_2$, R_1^*, és (R_1) szintén reguláris kifejezések.
3) Egy $R \in \{\Sigma \cup \{\emptyset, \lambda, (,), +, *\}\}^*$ szimbólumlánc akkor és csakis akkor **reguláris kifejezés**, ha előállítható az elemi reguláris kifejezésekből a 2) pontban megadott szabályok véges számú alkalmazásával.

4.1 példa: Legyen $\Sigma = \{a, b, c\}$.

Az $\big(a + (a + bc)\big)^*(c + \emptyset)$ szimbólumlánc reguláris kifejezés.

Az $(a + b+)^*$ szimbólumlánc nem reguláris kifejezés.

4.2 definíció: (a reguláris kifejezés által reprezentált nyelv)

Legyen R egy Σ egy ábécé feletti reguláris kifejezés. Az R **reguláris kifejezés által reprezentált** $L(R)$ **nyelv** az alábbi módon határozható meg:

- Az $\emptyset$ reguláris kifejezés az üres nyelvet reprezentálja, vagyis $L(\emptyset) = \emptyset$.
- A λ reguláris kifejezés a $\{\lambda\}$ nyelvet reprezentálja, vagyis $L(\lambda) = \{\lambda\}$.
- Ha $a \in \Sigma$, akkor az a reguláris kifejezés az $\{a\}$ nyelvet reprezentálja, vagyis $L(a) = \{a\}$.
- Ha R_1 és R_2 reguláris kifejezések, akkor

$$L(R_1 + R_2) = L(R_1) \cup L(R_2),$$
$$L(R_1 R_2) = L(R_1)L(R_2),$$
$$L(R_1^*) = (L(R_1))^*,$$
$$L((R_1)) = L(R_1).$$

4.2 példa: Legyen $\Sigma = \{0, 1\}$.

$L(0) = \{0\}, \quad L(1) = \{1\}$

$L(0 + 1) = L(0) \cup L(1) = \{0\} \cup \{1\} = \{0, 1\}$

$L(01) = L(0)L(1) = \{0\}\{1\} = \{01\}$

$L(0^*) = \left(L(0)\right)^* = \{0\}^* = \{\, \lambda, 0, 00, 000, \ldots\}$

$L\left((0 + 1)^*\right) = \{0, 1\}^* = \{\, \lambda, 0, 1, 00, 01, 10, 11, 000 \ldots\}$

$L\left((0 + 1)^* 00\right) = \{0, 1\}^* \{00\} = \{\, w \mid w = v00, \; v \in \{0, 1\}^*\}$

$L\left((01)^*\right) = \{01\}^* = \{\, \lambda, 01, 0101, 010101, \ldots\}$

$L\left((0 + 1)^* 1 (0 + 1)^*\right) = \{\, w \mid w = u1v, \; u, v \in \{0, 1\}^*\}$

$L\left((01)^* 111 (01)^*\right) = \{\, w \mid w = u111v, \; u, v \in \{01\}^*\}$

4.3 definíció: (reguláris kifejezések ekvivalenciája)

Az R_1 és R_2 reguláris kifejezések, akkor és csakis akkor **ekvivalens**ek, ha $L(R_1) = L(R_2)$.

4.3 példa: Legyen $\Sigma = \{a, b\}$.

Az $R_1 = a + ab$ és az $R_2 = a\emptyset^* + ab$ reguláris kifejezések ekvivalensek, mivel mindkettő az $\{a, ab\}$ nyelvet reprezentálja.

A reguláris kifejezések tulajdonságai

$$R_1 + R_2 \equiv R_2 + R_1$$

$$(R_1 + R_2) + R_3 \equiv R_1 + (R_2 + R_3)$$

$$(R_1 R_2) R_3 \equiv R_1 (R_2 R_3)$$

$$(R_1 + R_2) R_3 \equiv R_1 R_3 + R_2 R_3$$

$$R_1 (R_2 + R_3) \equiv R_1 R_2 + R_1 R_3$$

$$(R_1 + R_2)^* \equiv (R_1^* + R_2)^* \equiv (R_1 + R_2^*)^* \equiv (R_1^* + R_2^*)^*$$

$$(R_1 + R_2)^* \equiv (R_1^* R_2^*)^*$$

$$(R_1^*)^* \equiv R_1^*$$

$$R_1^* R_1 \equiv R_1 R_1^*$$

$$R_1 R_1^* + \lambda \equiv R_1^*$$

4.1 tétel: A reguláris kifejezéssel reprezentálható nyelvek osztálya zárt az unió halmazműveletre és a reguláris műveletekre nézve.

Bizonyítás:

Legyenek L_1, $L_2 \subseteq \Sigma^*$ reguláris kifejezéssel reprezentálható nyelvek. Megmutatjuk, hogy az $L_1 \cup L_2$, $L_1 L_2$ és L_1^* nyelvek is reprezentálhatók reguláris kifejezéssel.

Mivel L_1 és L_2 reguláris kifejezéssel reprezentálható nyelvek, ezért léteznek olyan R_1 és R_2 reguláris kifejezések, melyekre teljesül, hogy $L(R_1) = L_1$ és $L(R_2) = L_2$.

A **4.2 definíció** értelmében elmondhatjuk, hogy

$$L_1 \cup L_2 = L(R_1) \cup L(R_2) = L(R_1 + R_2),$$
$$L_1 L_2 = L(R_1)L(R_2) = L(R_1 R_2),$$
$$L_1^* = (L(R_1))^* = L(R_1^*),$$

vagyis mindhárom nyelv reprezentálható reguláris kifejezéssel. ■

4.2 tétel: Minden véges nyelv reprezentálható reguláris kifejezéssel.

Bizonyítás:

Legyen Σ egy ábécé és $L \subseteq \Sigma^*$ egy véges nyelv. Megmutatjuk, hogy az $L_1 \cup L_2$, $L_1 L_2$ és L_1^* nyelvek is reprezentálhatók reguláris kifejezéssel.

- Ha $L = \emptyset$, akkor az L nyelv reprezentálható az $R = \emptyset$ reguláris kifejezéssel.
- Ha $L \neq \emptyset$, akkor az $L = \{x_1, x_2, \dots, x_n\}$ ahol $n \geq 1$ és $x_1, x_2, \dots, x_n \in \Sigma^*$. Ekkor $L = \{x_1\} \cup \{x_2\} \cup \cdots \cup \{x_n\}$ és mivel a reguláris kifejezéssel reprezentálható nyelvek osztálya zárt az egyesítés műveletére nézve, elegendő igazolni, hogy egy $\{x\}$ alakú nyelv, ahol $x \in \Sigma^*$, reprezentálható reguláris kifejezéssel.
  - ha $x = \lambda$, akkor ez a nyelv reprezentálható pl. az $R = \emptyset^*$ reguláris kifejezéssel,

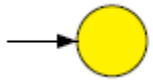
– ha $x \neq \lambda$, akkor $x = a_1 a_2 \ldots a_k$, ahol $k \geq 1$ és $a_1, a_2, \ldots, a_n \in \Sigma$. Ekkor $\{x\} = \{a_1\}\{a_2\} \ldots \{a_k\}$. A **4.2 definíció** értelmében az $\{a_1\}, \{a_2\}, \ldots, \{a_k\}$ nyelvek reprezentálhatók rendre az a_1, a_2,…, a_k reguláris kifejezésekkel. Mivel a reguláris kifejezéssel reprezentálható nyelvek osztálya zárt a konkatenáció műveletére nézve, ezért az $\{x\}$ nyelv is reprezentálható lesz reguláris kifejezéssel. ■

4.3 tétel: Tetszőleges Σ ábécé feletti reguláris kifejezéssel reprezentálható nyelv reguláris (*felismerhető véges automatával*).

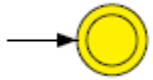
Bizonyítás:

Legyen Σ egy ábécé és $L \subseteq \Sigma^*$ egy reguláris kifejezéssel reprezentálható nyelv. A bizonyítást az L nyelvet reprezentáló R reguláris kifejezés struktúrája szerinti indukcióval végezzük.

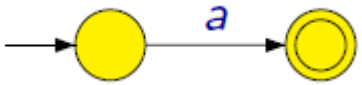
i. Legyen $R = \emptyset$. Ekkor $\mathrm{L}(R) = \emptyset$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:



ii. Legyen $R = \lambda$. Ekkor $\mathrm{L}(R) = \{\lambda\}$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:

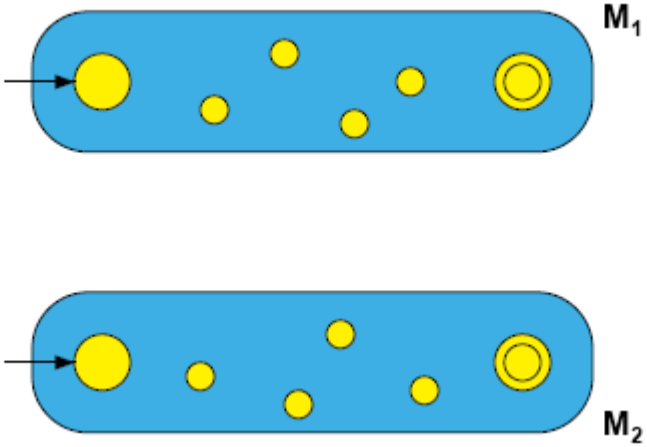


iii. Legyen $R = a$. Ekkor $\mathrm{L}(R) = \{a\}$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:

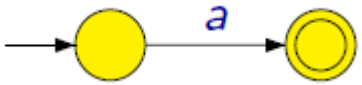


iv. a) Legyen $R = R_1 + R_2$ és tételezzük fel, hogy az R_1 és R_2 reguláris kifejezések által reprezentált $L(R_1)$ és $L(R_2)$ nyelvek felismerhetők az M_1 és M_2 véges automatákkal.

Ekkor $L(R) = L(R_1) \cup L(R_2)$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:

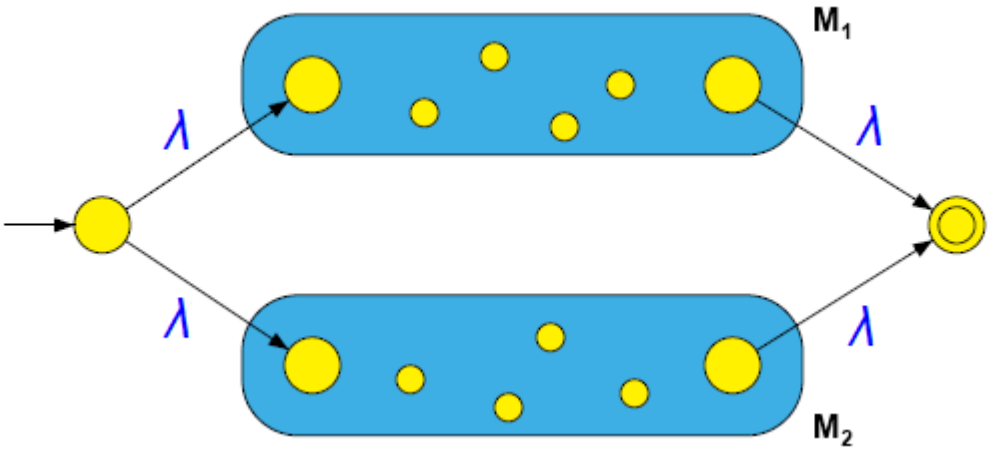


iii. Legyen $R = a$. Ekkor $\mathrm{L}(R) = \{a\}$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:



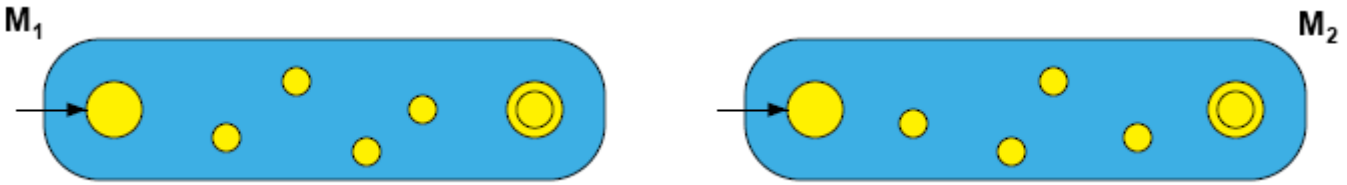
iv. a) Legyen $R = R_1 + R_2$ és tételezzük fel, hogy az R_1 és R_2 reguláris kifejezések által reprezentált $L(R_1)$ és $L(R_2)$ nyelvek felismerhetők az M_1 és M_2 véges automatákkal.

Ekkor $L(R) = L(R_1) \cup L(R_2)$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:



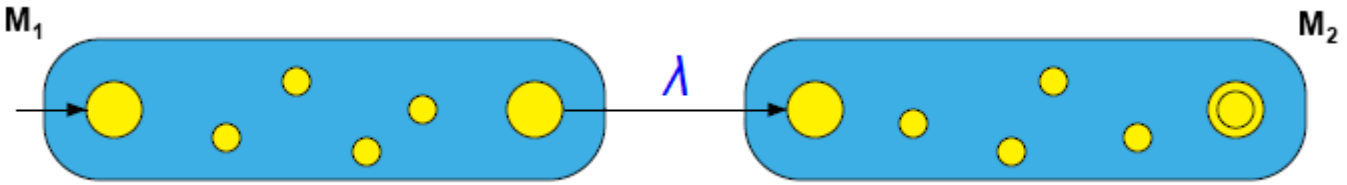
iv. b) Legyen $R = R_1 R_2$ és tételezzük fel, hogy az R_1 és R_2 reguláris kifejezések által reprezentált $L(R_1)$ és $L(R_2)$ nyelvek felismerhetők az M_1 és M_2 véges automatákkal.

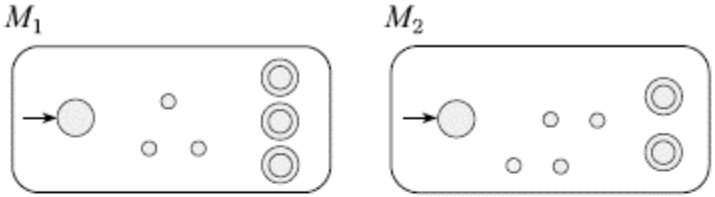
Ekkor $L(R) = L(R_1)L(R_2)$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:

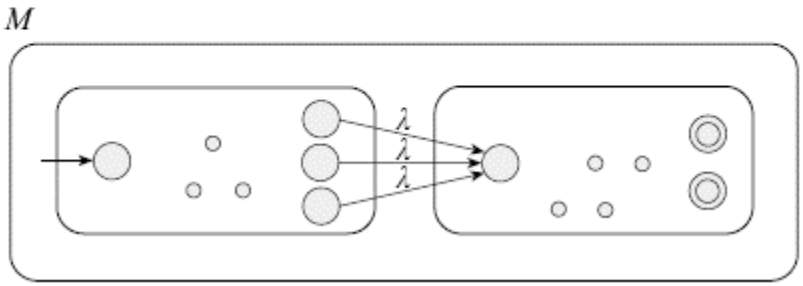


iv. b) Legyen $R = R_1 R_2$ és tételezzük fel, hogy az R_1 és R_2 reguláris kifejezések által reprezentált $L(R_1)$ és $L(R_2)$ nyelvek felismerhetők az M_1 és M_2 véges automatákkal.

Ekkor $L(R) = L(R_1)L(R_2)$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:

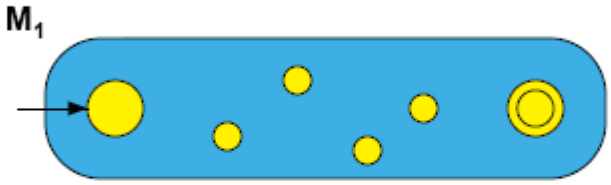






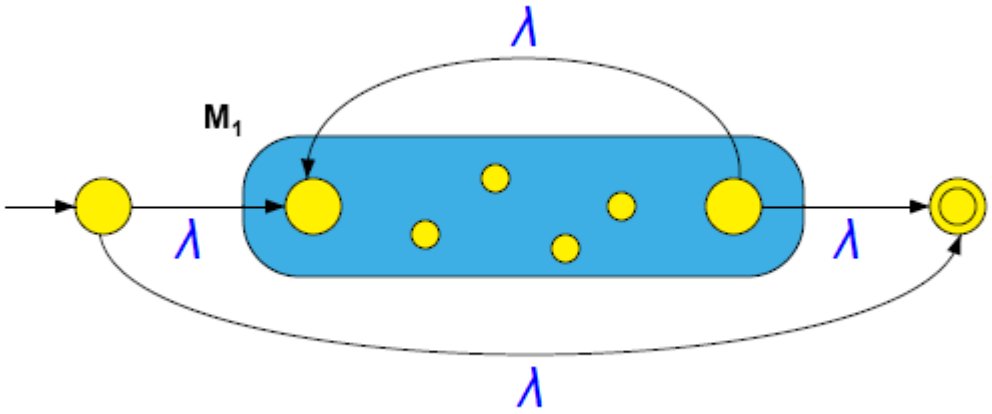
iv. c) Legyen $R = R_1^*$ és tételezzük fel, hogy az R_1 reguláris kifejezés által reprezentált $L(R_1)$ nyelv felismerhető az M_1 véges automatával.

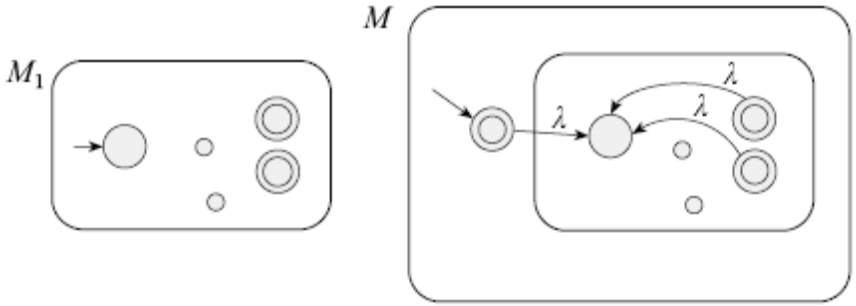
Ekkor $L(R) = L(R_1^*)$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:



iv. c) Legyen $R = R_1^*$ és tételezzük fel, hogy az R_1 reguláris kifejezés által reprezentált $L(R_1)$ nyelv felismerhető az M_1 véges automatával.

Ekkor $L(R) = L(R_1^*)$, és ez a nyelv felismerhető az alábbi átmenetdiagrammal megadott véges automatával:

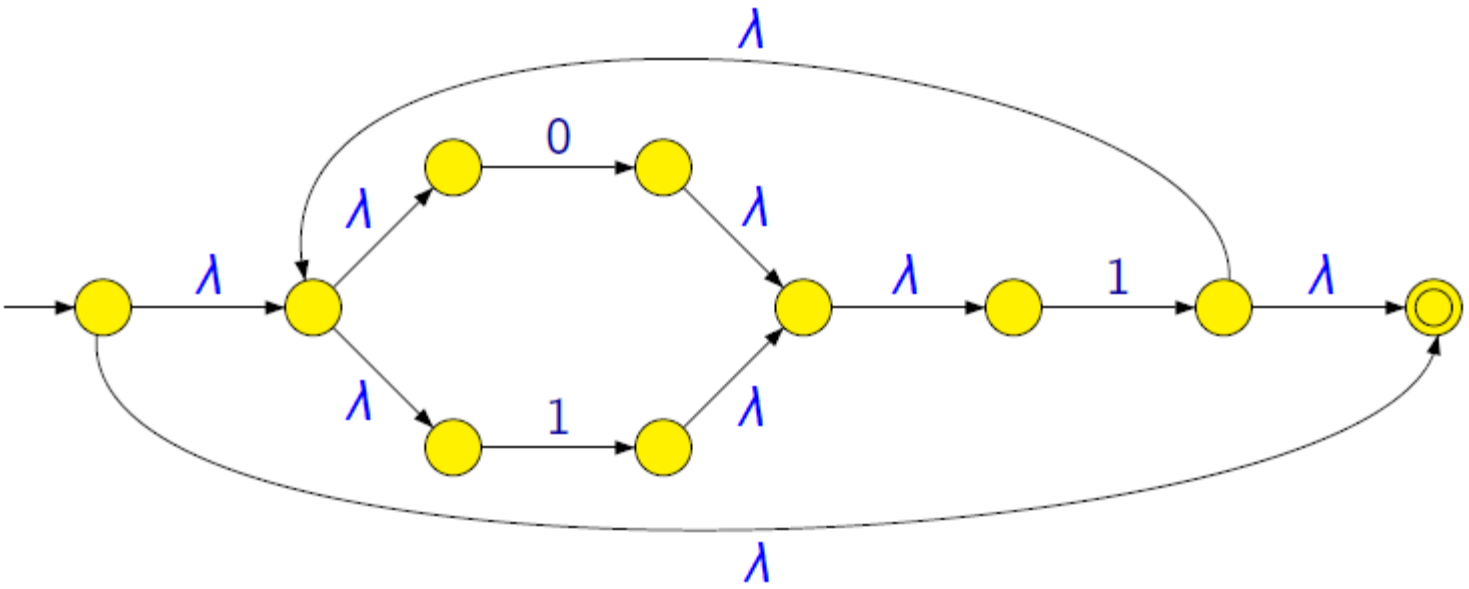




■

Selye János Egyetem
GIK, Informatika Tanszék

4.4 példa: Legyen adott az $R = ((\mathbf{0} + \mathbf{1})\mathbf{1})^*$ reguláris kifejezés. Szerkesztünk olyan M véges automatát, amelyre $L(M) = L(R)$.



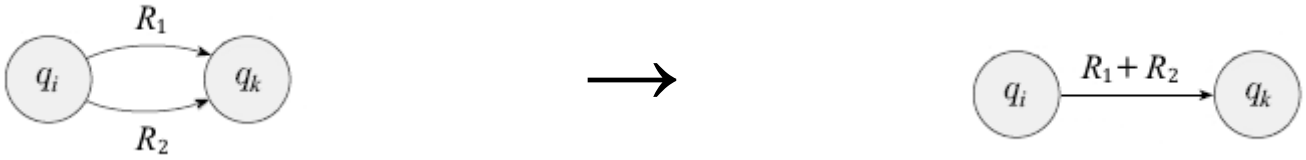
4.4 tétel: Tetszőleges Σ ábécé feletti reguláris nyelv reprezentálható reguláris kifejezéssel.

Bizonyítás:

Legyen Σ egy ábécé és $L \subseteq \Sigma^*$ egy reguláris nyelv. Ekkor létezik olyan M véges automata, melyre $L(M) = L$. Feltételezhetjük, hogy ennek az M automatának csak egyetlen végállapota van és $q_0 \notin F$.

Az M véges automatát módosítani fogjuk úgy, hogy végül csak egy kezdő- és egy végállapotot tartalmazzon. Az állapotokat fokozatosan fogjuk eltávolítani, és az eltávolítás utáni átmeneteket reguláris kifejezésekkel fogjuk jelölni.

A módosított automatán a köv. ekvivalens átalakítások végezhetők:



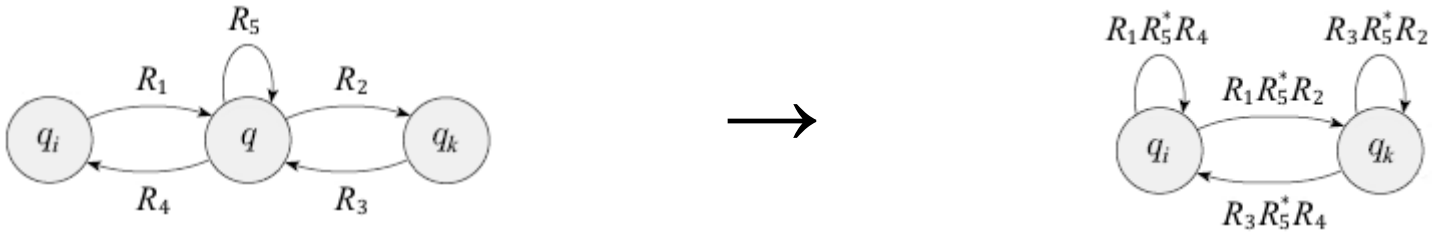
4.4 tétel: Tetszőleges Σ ábécé feletti reguláris nyelv reprezentálható reguláris kifejezéssel.

Bizonyítás:

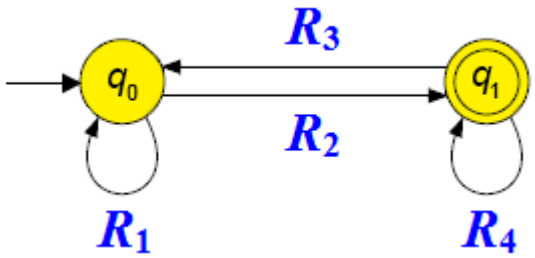
Legyen Σ egy ábécé és $L \subseteq \Sigma^*$ egy reguláris nyelv. Ekkor létezik olyan M véges automata, melyre $L(M) = L$. Feltételezhetjük, hogy ennek az M automatának csak egyetlen végállapota van és $q_0 \notin F$.

Az M véges automatát módosítani fogjuk úgy, hogy végül csak egy kezdő- és egy végállapotot tartalmazzon. Az állapotokat fokozatosan fogjuk eltávolítani, és az eltávolítás utáni átmeneteket reguláris kifejezésekkel fogjuk jelölni.

A módosított automatán a köv. ekvivalens átalakítások végezhetők:



Az állapotok eltávolítása után az alábbi véges automatát kapjuk:

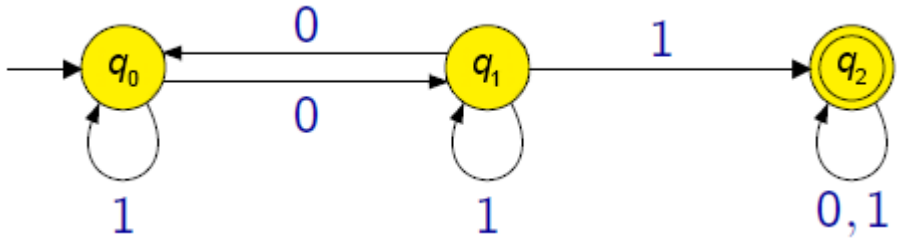


Ez a véges automata egy L nyelvet ismer fel, amely a következő reguláris kifejezéssel reprezentálható:

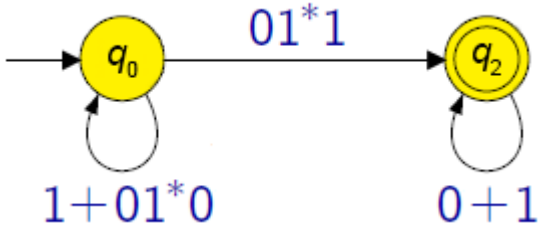
$$R = R_1^* R_2 (R_4 + R_3 R_1^* R_2)^*$$

Mivel az így kapott véges automata ekvivalens az M véges automatával, ezért érvényes, hogy $L(R) = L(M)$. ■

4.5 példa: Legyen adott az alábbi átmenetdiagramon látható M véges automata. Megadunk egy R reguláris kifejezést, amelyre $L(R) = L(M)$.



A q_1 állapot eltávolítása után a következő átmenetdiagramot kapjuk:

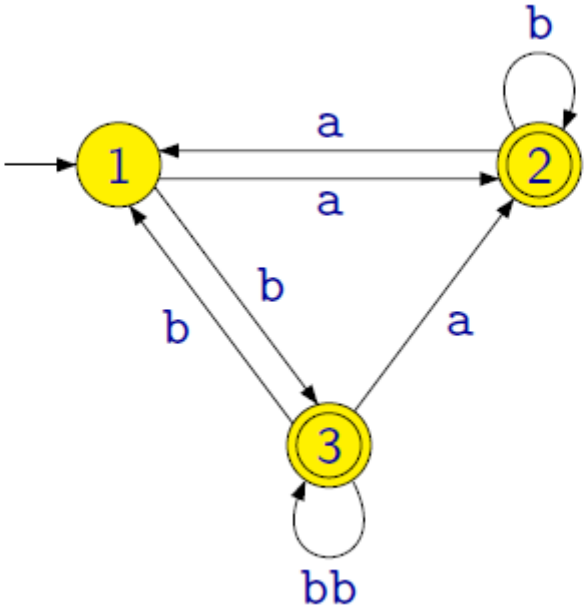


Az így kapott átmenetdiagramhoz reguláris kifejezés tartozik:

$$R = (\mathbf{1} + \mathbf{01^*0})^*\mathbf{01^*1}(\mathbf{0} + \mathbf{1})^*$$

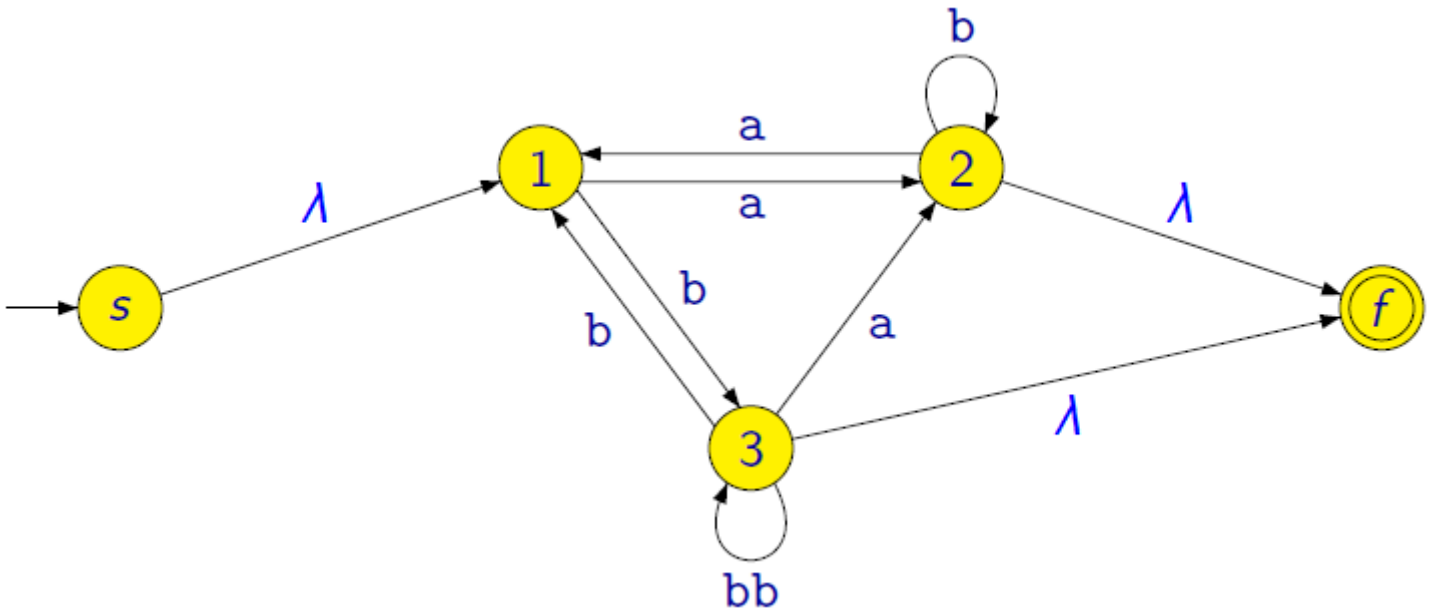
Selye János Egyetem
GIK, Informatika Tanszék

4.6 példa: Legyen adott az alábbi átmenetdiagramon látható M véges automata. Megadunk egy R reguláris kifejezést, amelyre $L(R) = L(M)$.



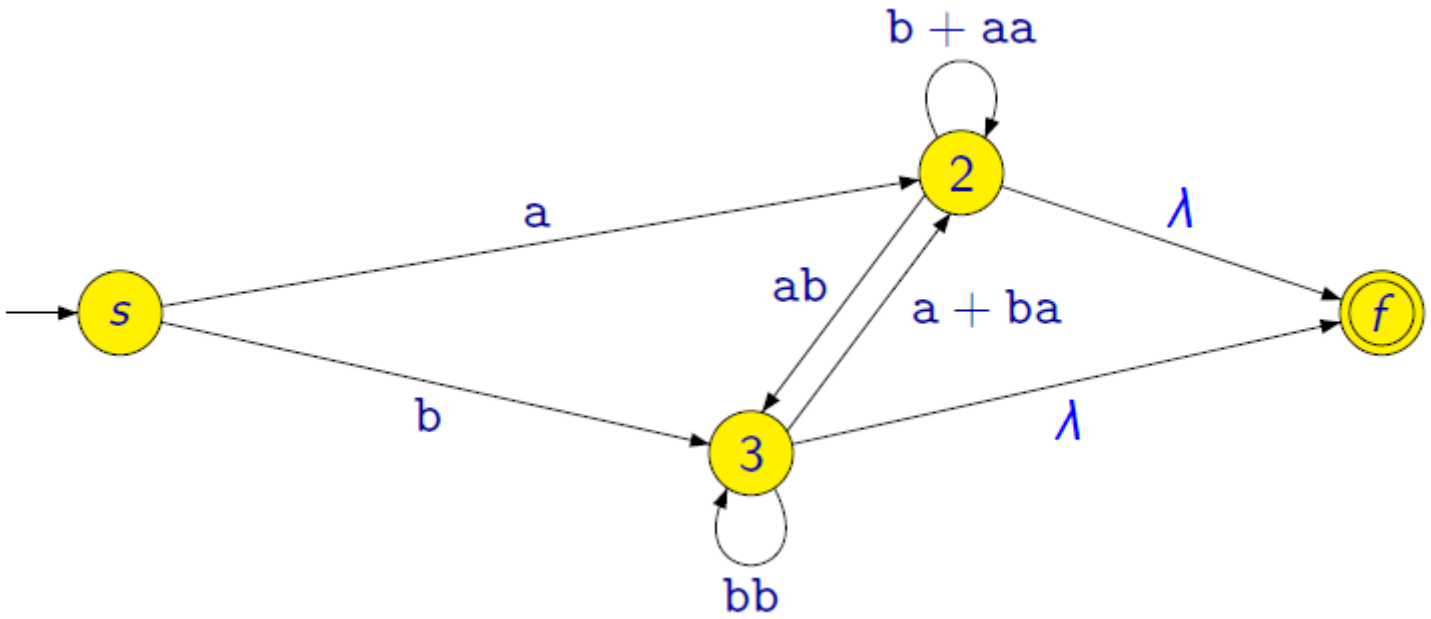
Selye János Egyetem
GIK, Informatika Tanszék

4.6 példa: Legyen adott az alábbi átmenetdiagramon látható M véges automata. Megadunk egy R reguláris kifejezést, amelyre $L(R) = L(M)$.



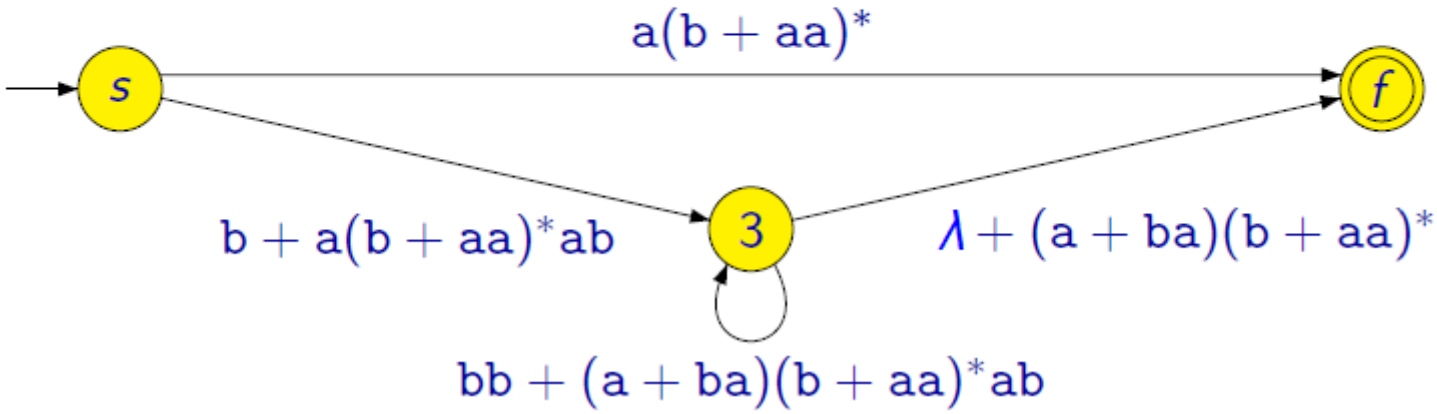
Selye János Egyetem
GIK, Informatika Tanszék

4.6 példa: Legyen adott az alábbi átmenetdiagramon látható M véges automata. Megadunk egy R reguláris kifejezést, amelyre $L(R) = L(M)$.

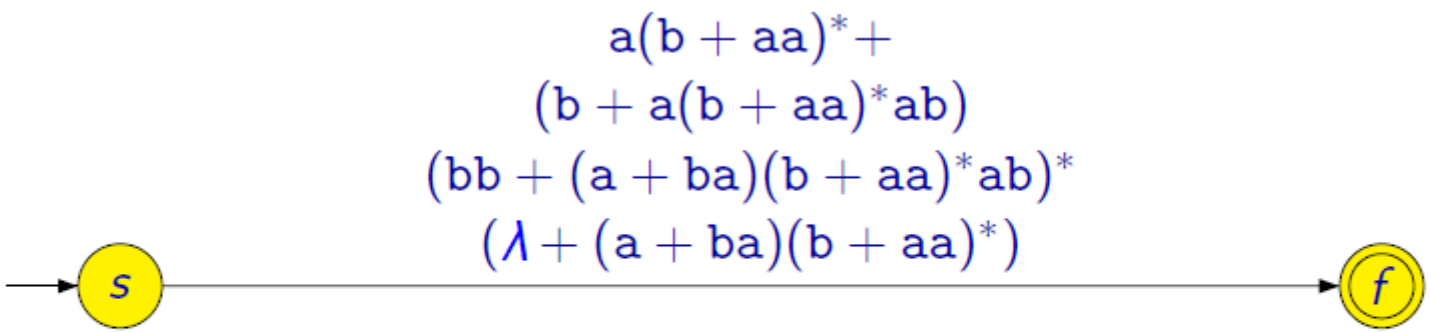


Selye János Egyetem
GIK, Informatika Tanszék

4.6 példa: Legyen adott az alábbi átmenetdiagramon látható M véges automata. Megadunk egy R reguláris kifejezést, amelyre $L(R) = L(M)$.



4.6 példa: Legyen adott az alábbi átmenetdiagramon látható M véges automata. Megadunk egy R reguláris kifejezést, amelyre $L(R) = L(M)$.



4.5 tétel: Tetszőleges Σ ábécé feletti reguláris kifejezéssel reprezentálható nyelv generálható 3-típusú nyelvtannal.

A továbbiakban azt a nyelvosztályt, amelynek a három fontos jellemzését (a Chomsky-féle besorolásban 3-típusú, véges automatával felismerhető és reguláris kifejezéssel reprezentálható) is megadtuk, **a reguláris nyelvek osztályá**nak fogjuk nevezni, elemeit pedig **reguláris nyelvek**nek.

Tehát ha egy L nyelv reguláris, akkor elmondható róla, hogy felismerhető véges automatával, generálható 3-típusú nyelvtannal és reprezentálható reguláris kifejezéssel.

4.6 tétel: (pumpáló lemma reguláris nyelvekre, kis Bar-Hillel lemma)

Legyen L tetszőleges reguláris nyelv. Ekkor megadható olyan, csak az L nyelvtől függő $k \geq 1$ természetes szám, hogy az L nyelv bármely legalább k hosszúságú w szava felírható $w = xyz$ alakban úgy, hogy teljesül az alábbi három feltétel:

1) $|xy| \leq k$,
2) $y \neq \lambda$,
3) minden $i = 0,1,2, \ldots$ számra teljesül, hogy $xy^i z \in L$.

A pumpáló lemma segítségével egy nyelvről bebizonyítható, hogy nem reguláris.

4.7 példa: Az $L = \{a^n b^n \mid n \geq 1\}$ nyelv nem reguláris.

Bizonyítás: (ellentmondással)

Tételezzük fel, hogy az L nyelv reguláris. Ekkor a pumpáló lemma szerint létezik olyan $k \geq 1$ természetes szám, hogy minden $w \in L$ szóra melynek hossza legalább k, teljesülnek a lemmában szereplő 1) – 3) feltételek.

Tekintsük a $w = a^k b^k \in L$ szót, melynek hossza nyilván nagyobb, mint k. Ekkor a pumpáló lemma alapján a w szó részszavakra bontható, azaz $w = a^k b^k = xyz$.

Mivel a lemmában szereplő 1) feltétel alapján $|xy| \leq k$, ezért az y részszó csak a szimbólumot tartalmazhat. Mivel a lemmában szereplő 2) feltétel alapján $y \neq \lambda$, ezért az y legalább egy a szimbólumot biztosan tartalmaz.

Legyen $y = a^r$, ahol $1 \leq r \leq k$. A pumpáló lemma 3) feltétele alapján $xy^0z = xz = a^{k-r}b^k \in L$.

Legyen $y = a^r$, ahol $1 \leq r \leq k$. A pumpáló lemma 3) feltétele alapján $xy^0z = xz = a^{k-r}b^k \in L$.

Azonban ez nem lehetséges, mivel ez a szó r-rel kevesebb a szimbólumot tartalmaz, mint b szimbólumot. Ellentmondást kaptunk tehát azzal, hogy $xy^0z \in L$. Ezért a kezdeti feltételezésünk, mely szerint az L nyelv reguláris, nem helyes. ■

4.8 példa: Az $L = \{a^p \mid p \text{ prímszám}\}$ nyelv nem reguláris.

Bizonyítás: (ellentmondással)

Tételezzük fel, hogy az L nyelv reguláris. Ekkor a pumpáló lemma szerint létezik olyan $k \geq 1$ természetes szám, hogy minden $w \in L$ szóra melynek hossza legalább k, teljesülnek a lemmában szereplő 1) – 3) feltételek.

Legyen q egy k-nál nagyobb prímszám (ilyen prímszám biztosan létezik, mert végtelen sok prímszám van).

Tekintsük a $w = a^q \in L$ szót, melynek hossza nyilván nagyobb, mint k. Ekkor a pumpáló lemma alapján a w szó részszavakra bontható, azaz $w = a^q = xyz$.

Mivel a lemmában szereplő 1) feltétel alapján $|xy| \leq k$, valamint a 2) feltétel alapján $y \neq \lambda$, ezért az y részszó legalább egy a szimbólumot biztosan tartalmaz.

Legyen $y = a^r$, ahol $1 \leq r \leq k$. A pumpáló lemma 3) feltétele alapján $xy^{q+1}z = xyy^qz = a^q a^{rq} = a^{q(1+r)} \in L$.

Azonban ez nem lehetséges, mivel a $q(1+r)$ szorzat nem prímszám, ugyanis mindkét tényezője nagyobb, mint 1. Ellentmondást kaptunk tehát azzal, hogy $xy^{q+1}z \in L$. Ezért a kezdeti feltételezésünk, mely szerint az L nyelv reguláris, nem helyes. ■

4.1 következmény: Érvényes, hogy $\mathcal{L}_3 \subset \mathcal{L}_2$.

Bizonyítás:

A **4.7 példá**ban a pumpáló lemma segítségével bebizonyítottuk, hogy az $L = \{a^n b^n \mid n \geq 1\}$ nyelv nem reguláris, azaz nincs benne az $\mathcal{L}_3$ nyelvosztályban.

Az $\mathcal{L}_3 \subset \mathcal{L}_2$ valódi tartalmazás igazolásához elegendő megadni egy olyan környezetfüggetlen nyelvtant, amely az L nyelvet generálja. Legyen $G = (N, \Sigma, P, S)$, ahol $N = \{S\}$, $\Sigma = \{a, b\}$ és $P = \{S \longrightarrow aSb,\ S \longrightarrow ab\}$. Könnyen ellenőrizhető, hogy a $G = (N, \Sigma, P, S)$ nyelvtan környezetfüggetlen és éppen az L nyelvet generálja. ■