460-4005/01: Teoretická informatika (TI) přednáška 2

prof. RNDr Petr Jančar, CSc.

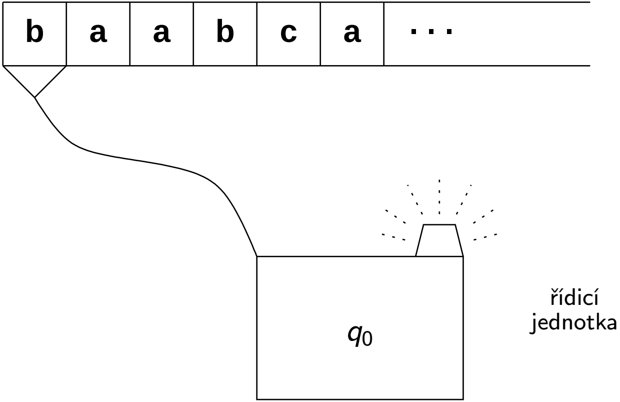
katedra informatiky FEI VŠB-TUO
www.cs.vsb.cz/jancar

LS 2010/2011

Připomenutí programu „sudý počet výskytů symbolu a“

```
#include <stdio.h>
int main(int argc, char **argv) {
    bool even_a = true;
    while(true) {
        int c = getchar();
        switch(c) {
            case 'a': even_a = !even_a; break;
            case EOF:
            case '\n':
                 if (even_a) { printf("Yes\n"); }
                       else { printf("No\n"); }
                 return 0;
        }
    }
}
```

Konečný automat (program s „malou“ pamětí) zvnějšku

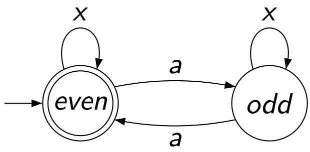


Program „sudý počet znaků a“ jinak

Á la Haskell

```
even_a [] = true
even_a ["a":rest] = ! even_a [rest]
even_a [_:rest] = even_a [rest]
```

Á la konečný automat



$x \neq "a"$

A	a	b	c	$\ldots$
$\leftrightarrow r_0$	r_1	r_0		
r_1	r_0	r_1		

$A = (Q, \Sigma, \delta, q_0, F) = (\{r_0, r_1\}, \{a, b, c, \ldots\}, \delta, r_0, \{r_0\})$

kde $\delta : Q \times \Sigma \rightarrow Q$ je konkrétně $\{((r_0, a), r_1), ((r_0, b), r_0), \ldots\}$
(tedy $\delta(r_0, a) = r_1$, $\delta(r_0, b) = r_0$, $\ldots$)

Konečný automat - (matematická) definice

Konečný automat (KA) je struktura (zde pětice)

$A = (Q, \Sigma, \delta, q_0, F)$, kde

Q ... konečná množina *stavů*

Σ ... konečná množina (vstupních) *symbolů* (*abeceda*)

$\delta : Q \times \Sigma \rightarrow Q$ ... *přechodová funkce*

$q_0 \in Q$ ... *počáteční stav*

$F \subseteq Q$ ... množina *přijímajících (koncových) stavů*

Abeceda, slovo, jazyk

abeceda .... označujeme $\Sigma, \Gamma, \Delta, \ldots$

prvky abecedy (*písmena*) ... označujeme $a, b, c, \ldots$

slovo .... $u, v, w, \ldots,\ u = a_1 a_2 \ldots a_n$

délka slova ... $|u|$
prázdné slovo ... ε ($|\varepsilon| = 0$)

zřetězení slov ... $u \cdot v$, stručněji uv

u je *podslovem* slova $v \quad \Leftrightarrow_{df}$ ex. slova w_1, w_2 tž. $v = w_1 u w_2$.
u je *prefixem* (*sufixem*) slova $v \quad \Leftrightarrow_{df} \quad \ldots$

značení $u^0 = \varepsilon,\ u^1 = u,\ u^2 = uu,\ \ldots\ u^n = uu..u$ (n-krát)

Σ^* ... množina všech slov v abecedě Σ

Jazyk nad abecedou Σ ... $L \subseteq \Sigma^*$

Příklad jazyka v abecedě $\{0,1\}$

$L = \{\, w \in \{0,1\}^* \mid bn(w)$ je dělitelné třemi a
w neobsahuje podřetězec $101\,\}$

kde $bn(w)$ znamená číslo, pro něž je w jeho binárním zápisem (např. $bn(0101) = 5$); klademe $bn(\varepsilon) = 0$.

Všimněme si, že

$$L = L_1 \cap \overline{L_2}$$

kde
$L_1 = \{\, w \in \{0,1\}^* \mid bn(w)$ je dělitelné třemi $\}$
$L_2 = \{w \in \{0,1\}^* \mid w$ obsahuje podřetězec $101\}$

Slova přijímaná automatem

Mějme zadán $A = (Q, \Sigma, \delta, q_0, F)$
Zavádíme značení

$$q \xrightarrow{w} q'$$

(neboli ternární relaci, zde
podmnožinu kartézského součinu $Q \times \Sigma^* \times Q$)

Induktivní definice (co to vlastně je?)

1. $q \xrightarrow{\varepsilon} q$ (axiom),
2. $(\delta(q, a) = q' \wedge q' \xrightarrow{u} q'') \Longrightarrow q \xrightarrow{au} q''$ (odvozovací pravidlo).

Á la „Haskell“

```
succ[q][]=q
succ[q][a:rest]=succ[delta(q,a)][rest]
```

Zkratka:
pro $R \subseteq Q$, místo $\exists r \in R : q \xrightarrow{u} r$ píšeme stručněji $q \xrightarrow{u} R$

Slovo $w \in \Sigma^*$ je *přijímáno* automatem $A \quad \Leftrightarrow_{df} \quad q_0 \xrightarrow{w} F$.

Regulární jazyky

Jazykem rozpoznávaným (přijímaným) automatem A je jazyk
$L(A) = \{w \in \Sigma^* \mid \text{ slovo } w \text{ je přijímáno } A\} = \{w \mid q_0 \xrightarrow{w} F\}$

Jazyk L je *regulární* $\Leftrightarrow_{df}$ existuje KA A tž. $L(A) = L$.

Alespoň intuitivně chápeme:

Levý kvocient jazyka L podle slova u: $u \backslash L = \{v \mid uv \in L\}$.

Věta. Jazyk $L \subseteq \Sigma^*$ je regulární právě tehdy, když je množina kvocientů $\{w \backslash L \mid w \in \Sigma^*\}$ konečná.

(... čili když si při čtení slova stačí z dosud přečteného pamatovat omezenou informaci ...)

Schválně, poznáte, které jazyky jsou regulární?

pro $w \in \Sigma^*$,
$|w|$ označuje délku slova w,
$|w|_a$ označuje počet výskytů symbolu a ve w ("délka w v a-čkách"),

1. $\{w \in \{a,b\}^* \mid |w|_a \bmod 2 = 0\}$
2. $\{w \in \{a,b\}^* \mid w$ začíná nebo končí dvojicí stejných písmen $\}$
3. $\{w \in \{a,b\}^* \mid |w|_a < |w|_b\}$
4. $\{w \in \{a,b,c\}^* \mid$ jestliže w neobsahuje podřetězec abc, pak končí řetězcem $bca\}$
5. $\{w \in \{a,b\}^* \mid |w|_a > |w|_b$ nebo w končí $baa\}$
6. $\{w \in \{a,b\}^* \mid |w|_a > |w|_b$ nebo $|w|_b \geq 2\}$

(Modulární) konstrukce automatu přijímajícího daný jazyk

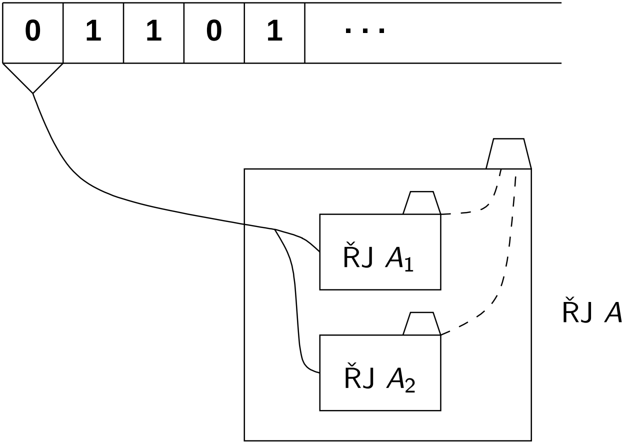
$L = \{ w \in \{0,1\}^* \mid bn(w)$ je dělitelné třemi a
w neobsahuje podřetězec $101\}$

A_1	0	1
$\leftrightarrow q_1$	q_1	q_2
q_2	q_3	q_1
q_3	q_2	q_3

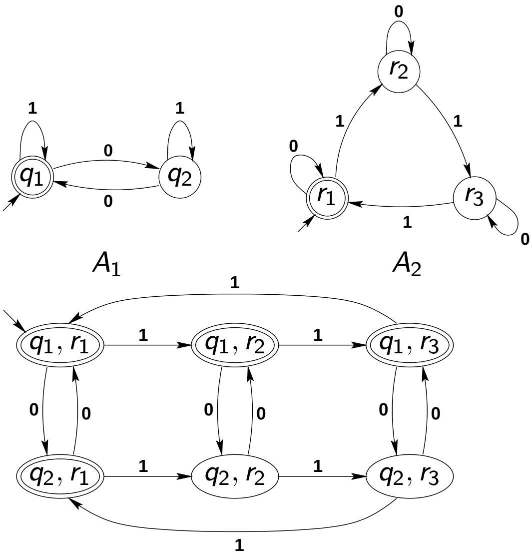
A_2	0	1
$\leftrightarrow r_1$	r_1	r_2
$\leftarrow r_2$	r_3	r_2
$\leftarrow r_3$	r_1	r_4
r_4	r_4	r_4

A	0	1
$\leftrightarrow(q_1, r_1)$	(q_1, r_1)	(q_2, r_2)
$\leftarrow(q_1, r_2)$	(q_1, r_3)	(q_2, r_2)
$\leftarrow(q_1, r_3)$	(q_1, r_1)	(q_2, r_4)
(q_1, r_4)	(q_1, r_4)	(q_2, r_4)
(q_2, r_1)	(q_3, r_1)	(q_1, r_2)
(q_2, r_2)	(q_3, r_3)	(q_1, r_2)
(q_2, r_3)	(q_3, r_1)	(q_1, r_4)
(q_2, r_4)	(q_3, r_4)	(q_1, r_4)
(q_3, r_1)	(q_2, r_1)	(q_3, r_2)
(q_3, r_2)	(q_2, r_3)	(q_3, r_2)
(q_3, r_3)	(q_2, r_1)	(q_3, r_4)
(q_3, r_4)	(q_2, r_4)	(q_3, r_4)

KA A pro jazyk $L(A_1) \cap L(A_2)$ či $L(A_1) \cup L(A_2)$



Konstrukce KA pro rozpoznávání $L(A_1) \cup L(A_2)$



Nejelegantnější (exaktní) popis konstrukce (... že jo?)

Mějme dva konečné automaty
$A_1 = (Q_1, \Sigma, \delta_1, q_{01}, F_1)$, $A_2 = (Q_2, \Sigma, \delta_2, q_{02}, F_2)$.
Definujme $A = (Q, \Sigma, \delta, q_0, F)$ takto:

- $Q = Q_1 \times Q_2$
- pro každé $q_1 \in Q_1$, $q_2 \in Q_2$, $a \in \Sigma$:
  $\delta(\,(q_1, q_2), a\,) = (\,\delta_1(q_1, a), \delta_2(q_2, a)\,)$
- $q_0 = (q_{01}, q_{02})$
- $F = (F_1 \times F_2)$

Induktivně jde snadno dokázat toto tvrzení (kdo to umí?):

Pro vš. $q_1, q_1' \in Q_1$, $q_2, q_2' \in Q_2$, $w \in \Sigma^*$:

$$(q_1, q_2) \xrightarrow{w}_A (q_1', q_2') \quad \Leftrightarrow \quad (q_1 \xrightarrow{w}_{A_1} q_1') \wedge (q_2 \xrightarrow{w}_{A_2} q_2').$$

Tedy $L(A) = L(A_1) \cap L(A_2)$,
neboť $(q_{01}, q_{02}) \xrightarrow{w}_A F \Leftrightarrow (q_{01} \xrightarrow{w}_{A_1} F_1 \wedge q_{02} \xrightarrow{w}_{A_2} F_2)$.

Když dáme $F = (F_1 \times Q_2) \cup (Q_1 \times F_2)$, tak $L(A) = L(A_1) \cup L(A_2)$.

(Některé) uzávěrové vlastnosti třídy REG

Věta.
Jestliže $L \subseteq \Sigma^*$ je regulární jazyk, pak

- doplněk $\overline{L} = \Sigma^* - L$ je regulární jazyk.

Jestliže L_1, L_2 jsou regulární jazyky, pak

- $(L_1 \cup L_2)$ je regulární jazyk,
- $(L_1 \cap L_2)$ je regulární jazyk,
- $(L_1 - L_2)$ je regulární jazyk.

Jinými slovy
Věta.
Třída regulárních jazyků, označme ji REG, je uzavřena vůči booleovským množinovým operacím (... a sice konstruktivně ...).

(Izomorfní) úprava automatu

A	0	1
$\leftrightarrow(q_1, r_1)$	(q_1, r_1)	(q_2, r_2)
$\leftarrow(q_1, r_2)$	(q_1, r_3)	(q_2, r_2)
$\leftarrow(q_1, r_3)$	(q_1, r_1)	(q_2, r_4)
(q_1, r_4)	(q_1, r_4)	(q_2, r_4)
(q_2, r_1)	(q_3, r_1)	(q_1, r_2)
(q_2, r_2)	(q_3, r_3)	(q_1, r_2)
(q_2, r_3)	(q_3, r_1)	(q_1, r_4)
(q_2, r_4)	(q_3, r_4)	(q_1, r_4)
(q_3, r_1)	(q_2, r_1)	(q_3, r_2)
(q_3, r_2)	(q_2, r_3)	(q_3, r_2)
(q_3, r_3)	(q_2, r_1)	(q_3, r_4)
(q_3, r_4)	(q_2, r_4)	(q_3, r_4)

A'	0	1	
$\leftrightarrow s_1$	s_1	s_2	$s_1 = (q_1, r_1)$
s_2	s_3	s_4	$s_2 = (q_2, r_2)$
s_3	s_5	s_6	$s_3 = (q_3, r_3)$
$\leftarrow s_4$	s_7	s_2	$s_4 = (q_1, r_2)$
s_5	s_8	s_4	$s_5 = (q_2, r_1)$
s_6	s_9	s_6	$s_6 = (q_3, r_4)$
$\leftarrow s_7$	s_1	s_9	$s_7 = (q_1, r_3)$
s_8	s_5	s_{10}	$s_8 = (q_3, r_1)$
s_9	s_6	s_{11}	$s_9 = (q_2, r_4)$
s_{10}	s_{12}	s_{10}	$s_{10} = (q_3, r_2)$
s_{11}	s_{11}	s_9	$s_{11} = (q_1, r_4)$
s_{12}	s_8	s_{11}	$s_{12} = (q_2, r_3)$

Co je to ten izomorfismus?

Dva konečné automaty
$A_1 = (Q_1, \Sigma, \delta_1, q_{01}, F_1)$, $A_2 = (Q_2, \Sigma, \delta_2, q_{02}, F_2)$
jsou *izomorfní*, jestliže

existuje zobrazení (funkce) $f : Q_1 \rightarrow Q_2$ takové, že

- f je bijekce
  ("prosté": $q \neq q' \Rightarrow f(q) \neq f(q')$,
  "na": $\forall r \in Q_2 \, \exists q \in Q_1 : f(q) = r$)
- $f(q_{01}) = f(q_{02})$
- $q \in F_1 \iff f(q) \in F_2$
- $q \xrightarrow{a} q' \iff f(q) \xrightarrow{a} f(q')$
  neboli $f(\delta_1(q, a)) = \delta_2(f(q), a)$

Tedy: Izomorfní automaty A_1, A_2 jsou "stejné, až na pojmenování stavů".

Jaká je to množina?

Induktivní definice:

Množina $X \subseteq Q$ automatu $A = (Q, \Sigma, \delta, q_0, F)$ je induktivně definována takto:
1/ $q_0 \in X$,
2/ $(q \in X \wedge \exists a : q \xrightarrow{a} q'\ ) \Longrightarrow q' \in X$.

Ano, je to tak, jedná se o množinu dosažitelných stavů automatu $A = (Q, \Sigma, \delta, q_0, F)$, tedy

$$Reach(A) = \{\, q \mid q_0 \longrightarrow^* q \,\}$$

Pozor, žádnou relaci $\longrightarrow^* \subseteq Q \times Q$ jsme ještě nedefinovali!
Ale je to jasné, ne?

- relace $\longrightarrow$ je definována takto: $q \longrightarrow q' \Leftrightarrow_{df} \exists a : q \xrightarrow{a} q'$,
- relace $\longrightarrow^*$ je reflexivním a tranzitivním uzávěrem relace $\longrightarrow$.

Reflexivní a tranzitivní uzávěr relace

relace ρ *na množině* M ... $\rho \subseteq M \times M$

píšeme často $x \rho y$ místo $(x, y) \in \rho$

ρ je *reflexivní* $\Leftrightarrow_{df}$ $\forall x \in M : x \rho x$

ρ je *tranzitivní* $\Leftrightarrow_{df}$
$\forall x, y, z \in M : (x \rho y \ \wedge \ y \rho z) \Longrightarrow x \rho z$

Reflexivním a tranzitivním uzávěrem relace ρ
je nejmenší relace, která obsahuje ρ a je reflexivní i tranzitivní;
často se označuje ρ^*.

Odstranění nedosažitelných stavů

Tvrzení. Ke každému KA A lze zkonstruovat KA A', v němž každý stav je dosažitelný a $L(A') = L(A)$.

A	0	1
$\leftrightarrow s_1$	s_1	s_2
s_2	s_3	s_4
s_3	s_5	s_6
$\leftarrow s_4$	s_7	s_2
s_5	s_8	s_4
s_6	s_9	s_6
$\leftarrow s_7$	s_1	s_9
s_8	s_5	s_{10}
s_9	s_6	s_{11}
s_{10}	s_{12}	s_{10}
s_{11}	s_{11}	s_9
s_{12}	s_8	s_{11}

1/ $q_0 \in Reach(A)$, 2/ $(q \in Reach(A) \wedge q \longrightarrow q'\ ) \Longrightarrow q' \in Reach(A)$.

Podílový automat (faktor-automat)

K $A = (Q_\sim, \Sigma, \delta, q_0, F)$ definujme ekvivalenci $\sim$ na množině Q takto:

$$q \sim q' \iff_{df} L_q^{toAcc} = L_{q'}^{toAcc}$$

kde $L_q^{toAcc} = \{w \in \Sigma^* \mid q \xrightarrow{w} F\}$.

Podílový automat podle ekvivalence $\sim$, označený $A_\sim$, definujeme takto (píšeme stručněji $[q]$ místo $[q]_\sim$):

$A_\sim = (Q_\sim, \Sigma, \delta_\sim, [q_0], F_\sim)$, kde
$Q_\sim = \{\ [q] \mid q \in Q\ \}$,
$F_\sim = \{\ [q] \mid q \in F\ \}$
$\delta_\sim([q], a) = [\delta(q, a)]$

Korektnost definice (co to je?) plyne z
$\big(p \sim q \Rightarrow (p \in F \Leftrightarrow q \in F)\big)$ a z $\big(p \sim q \Rightarrow \delta(p, a) \sim \delta(q, a)\big)$.

Jak ukážeme, že $L(A) = L(A_\sim)$?

Stačí (induktivně dle délky slova) dokázat
$q \xrightarrow{w}_A q' \ \Leftrightarrow\ [q] \xrightarrow{w}_{A_\sim} [q']$
.... stačí, ale platí to ? (Pokud ne, zkuste tvrzení upravit ...)

Připomenutí: relace ekvivalence

Relace ρ na množině M je
ekvivalence $\Leftrightarrow_{df}$ ρ je:

- reflexivní ($\forall x \in M : x \rho x$),
- symetrická
  ($\forall x, y \in M : x \rho y \Rightarrow y \rho x$),
- tranzitivní ($\forall x, y, z \in M :$
  $(x \rho y \ \wedge \ y \rho z) \Longrightarrow x \rho z$).

Ekvivalence ρ definuje na M *rozklad*

$$\{ [x]_\rho \mid x \in M \}$$

tj. systém vzájemně disjunktních množin, zvaných třídy ekvivalence, jejichž sjednocením je M, kde

$$[x]_\rho = \{ y \mid x \rho y \}.$$

Jak ale zjistit, zda $L_q^{toAcc} = L_{q'}^{toAcc}$?

A	0	1
$\leftrightarrow s_1$	s_1	s_2
s_2	s_3	s_4
s_3	s_5	s_6
$\leftarrow s_4$	s_7	s_2
s_5	s_8	s_4
s_6	s_9	s_6
$\leftarrow s_7$	s_1	s_9
s_8	s_5	s_{10}
s_9	s_6	s_{11}
s_{10}	s_{12}	s_{10}
s_{11}	s_{11}	s_9
s_{12}	s_8	s_{11}

$\sim_0, \sim_1, \sim_2, \cdots$