JUMPING FINITE AUTOMATA: DETERMINISM

Juraj Hrubý
Bachelor Degree Programme (3), FIT BUT
E-mail: xhruby19@stud.fit.vutbr.cz

Supervised by: Alexander Meduna
E-mail: meduna@fit.vutbr.cz

Abstract: Determinism of jumping finite automata is required for their usage and implementation. Determinism based on restriction of the state control, meaning that for every configuration of the automaton there is just one or none next possible state, is not sufficient for implementation because unpredictable jumps on input tape still remain. This problem is solved by restriction of the binary jumping relation.

Keywords: jumping finite automata, determinism, usage

1 ÚVOD

Diskontuálne spracovanie informácii v modernej výpočtovej technike je popisované formálnymi modelmi (konečné automaty, zásobníkové automaty), ktoré spracovávajú vstupnú pásku zľava doprava. Tento mierny rozpor viedol k myšlienke profesora Medunu zaviesť formálny model, ktorý by nepracoval zaužívane zľava doprava, ale naopak, ktorý by mohol skákať po vstupnej páske. So svojím študentom, inžinierom Zemkom, definovali skákajúce konečné automaty a popísali niektoré vlastnosti [1], čím otvorili novú oblasť teoretickej informatiky a poukázali na viacero otvorených problémov, ktoré je potrebné riešiť. Jedným z nich je striktný determinizmus.

Pre plné porozumenie princípu a následne aj významu skákajúcich konečných automatov sa predpokladá znalosť základných pojmov a princípov formálnych jazykov, Chomského hierarchie jazykov, základných typov automatov, matematických pojmov a operácií s nimi spojených [2, 3].

2 SKÁKAJÚCI KONEČNÝ AUTOMAT

Pre prehľadnosť uvádzam definíciu zavedenú v článku [1], na ktorú sa budem odvolávať.

ε je prázdny reťazec.

Definícia 2.0.1. Skákajúci konečný automat (ďalej len **SKA**) M je pätica $(Q, \Sigma, \delta, s, F)$ kde

Q je konečná množina stavov,
Σ je konečná vstupná abeceda,
δ je zobrazenie $Q \times (\Sigma \cup \{\varepsilon\}) \to 2^Q$,
$s \in Q$ je počiatočný stav,
$F \subseteq Q$ je množina koncových stavov.

Konfigurácia **SKA** M je trojica (x, q, y) kde

$xy \in \Sigma^*$ je vstupný reťazec na vstupnej páske s najľavejším znakom reťazca y pod čítacou hlavou,
$q \in Q$ je stav, v ktorom sa automat nachádza.

Binárna skoková relácia nad konfiguráciami $\Sigma^* Q \Sigma^*$, symbolicky značená $\curvearrowright$, je definovaná nasledovne. Pre všetky $q, r \in Q$, $a \in \Sigma \cup \{\varepsilon\}$, $x', y', x, y \in \Sigma^*$, pre ktoré $\delta(q,a)$ obsahuje r a $x'y' = xy$, píšeme $(x,q,ay) \curvearrowright (x',r,y')$.

Môžeme rozšíriť $\curvearrowright$ *na* $\curvearrowright^m$, kde $m \geq 0$, $\curvearrowright^+$ je tranzitívny uzáver a $\curvearrowright^*$ je tranzitívno-reflexívny uzáver.

Vstupný reťazec xy je prijatý **SKA** M len ak $(x,s,y) \curvearrowright^* (\varepsilon, f, \varepsilon)$, kde s je počiatočný stav a $f \in F$.

Jazyk $L(M)$, prijatý **SKA** M, je množina reťazcov $\{w \mid w = xy, (x,s,y) \curvearrowright^* (\varepsilon, f, \varepsilon)$ pre nejaký $f \in F\}$.

2.1 DETERMINISTICKÝ SKA

Vyššie uvedená definícia 2.0.1 popisuje nedeterministický **SKA**. Pre zavedenie deterministického konečného skákajúceho automatu (ďalej len **DSKA**) je najprv potrebné previesť **SKA** na **SKA** bez ε-skokov, t.j. definičný obor zobrazenia δ bude obsahovať iba dvojice (q,a), kde $q \in Q$ a $a \in \Sigma$.

DSKA je taký **SKA** bez ε-skokov, v ktorom pre každý stav $q \in Q$ a nejaký znak $a \in \Sigma$ existuje najviac jeden stav $r \in Q$, do ktorého môže skočiť.

Avšak pre **DSKA** M tu vzniká problém. V binárnej skokovej relácii $(x,q,ay) \curvearrowright (x',r,y')$ nie je explicitne uvedené, ako sa zmení konfigurácia. Vieme len, že $xy = x'y'$ a $a \in \Sigma$. Nemáme žiadnu informáciu o najľavejšom znaku reťazca y', nad ktorým bude čítacia hlava v nasledujúcom stave. To znamená, že pre M existuje n možných konfigurácií, do ktorých môže skočiť, pričom $n = |x'y'|$.

2.2 STRIKTNE DETERMINISTICKÝ SKA

Problém takto oslabeného determinizmu, môžeme vyriešiť tak, že určíme znak, nad ktorý sa má čítacia hlava presunúť. Ktorý znak to bude?

Ak (x,q,ay) je konfigurácia M, $\delta(q,a)$ obsahuje r, potom pre každú dvojicu (q,a) existuje množina znakov, pre ktoré sa M v budúcom stave r nezasekne.

Pre všetky dvojice (q,a) v definičnom obore δ definujeme zobrazenie $\alpha : Q \times \Sigma \to 2^{\Sigma}$, pričom $\alpha(q,a) = \{b \mid (\delta(q,a),b)$ je v definičnom obore $\delta\}$.

Pre dosiahnutie striktného determinizmu je potrebné určiť jeden znak na vstupnej páske, nad ktorý sa presunie čítacia hlava.

Nech (x,q,ay) je konfigurácia M.

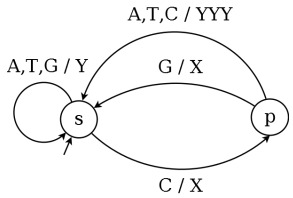
Idea
Čítacia hlava sa presunie nad prvý výskyt znaku $b \in \alpha(q,a)$ v reťazci xy zľava. V prípade, že reťazec xy neobsahuje žiadny znak $b \in \alpha(q,a)$, M sa zasekne.

Definícia 2.2.1. Deterministická binárna skoková relácia nad konfiguráciami, značená $\curvearrowright_d$, je definovaná nasledovne: (*x, q, ay*) $\curvearrowright_d$ (*x', r, y'*), kde $x, y, x', y' \in \Sigma^*$, $a \in \Sigma$, $q, r \in Q$, *xy*=*x'y'*, $\delta(q,a)$ obsahuje r a najľavejší znak reťazca *y'* je prvým výskytom znaku $b \in \alpha(q,a)$ v pôvodnom reťazci xy zľava, čiže reťazec *x'* neobsahuje žiadny znak $b \in \alpha(q,a)$.

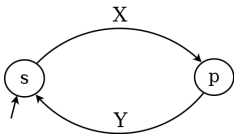
Dôsledok
Pomocou deterministickej binárnej skokovej relácie je možné definovať striktne deterministický konečný skákajúci automat (ďalej len **SDSKA**).

3 VYUŽITIE

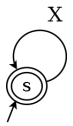
Praktické využitie **SDSKA**, ako som už spomínal, je predmetom výskumu, pričom jednou z možných oblastí využitia je genetika.
CpG ostrovček je definovaný ako oblasť s viac ako 200 bázovými dvojicami, kde výskyt dvojíc *CG* je vyšší ako 50 % a pomer pozorované/očakávané (obs/exp) vyšší ako 60 % [4]. Vlastnosť, že výskyt dvojíc *CG* v danom segmente reťazca je vyšší ako 50 % je možné overiť pomocou konečného prevodníka (Obr. 1), **SDSKA** (Obr. 2) a konečného automatu (Obr. 3), pričom výstup jedného automatu/prevodníka je vstupom nasledovného.



Obr. 1: Konečný prevodník na predspracovanie reťazca DNA.



Obr. 2: SDSKA na pokračovanie spracovania.



Obr. 3: Konečný automat na dokončenie spracovania.

Spracovávaný je reťazec nad abecedou $\Sigma = \{C, G, A, T\}$, kedže v DNA sa iné nukleotidy nenachádzajú. Prvým krokom je preklad konečným prevodníkom (Obr. 1) do novej, jednoduchšej abecedy $\Delta = \{X, Y\}$ eliminovaním dvojíc *CG*. Nasleduje spracovanie **SDSKA** (Obr. 2) a na záver konečný automat (Obr. 3). Reťazec je prijatý, ak automat (Obr. 3) prijme celý reťazec, čím je podmienka pre výskyt dvojíc *CG* (po preklade výskyt znakov *X*) vyšší ako 50 % splnená.

4 ZÁVER

Využitie skákajúcich konečných automatov v praxi je predmetom výskumu, pričom pre ich implementáciu je striktný determinizmus kľúčový. Nevýhodou riešenia zavedenia deterministickej binárnej skokovej relácie je nutnosť vyhľadávania znakov na vstupnej páske, čím sa zvyšuje časová náročnosť algoritmu, ktorým je **SDSKA** popísaný. Otvára sa tým nová otázka, či neefektívnosť jedného automatu bude možné kompenzovať použitím viacerých skákajúcich konečných automatov, ktoré budú spracovávať vstupnú pásku paralelne.

POĎAKOVANIE

Chcel by som sa poďakovať predovšetkým Prof. RNDr. Medunovi, CSc. za konzúltácie a smerovanie pri riešení problémov. Taktiež by som sa chcel poďakovať za konzlultácie Ing. Burgetovej, Ph.D., Ing. Zemkovi a Ing. Ryšavému, Ph.D.

LITERATÚRA

[1] Meduna, A., Zemek, P.: Jumping Finite Automata, In: International Journal of Foundations of Computer Science, roč. 23, č. 7, 2012, SG, s. 1555-1578, ISSN 0129-0541

[2] Martin, John C.: Introduction to languages and the theory of computation, [4. vydanie]. New York, US: McGraw-Hill, 2011. s. 1-298, ISBN 978-0-07-319146-1

[3] Aho, Alfred V. a Ullman, Jeffrey D.: The Theory of Parsing, Translation and Compiling, Volume I: Parsing. US: Prentice-Hall, Inc., 1972. ISBN 0-13-914556-7

[4] `http://en.wikipedia.org/wiki/CpG_island`