SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Implementacija FM-indeks algoritma

Ivan Borko, Sofia Čolaković, Florijan Stamenković
Voditelj: *doc. dr. sc. Mirjana Domazet Lošo*

Zagreb, siječanj 2015.

SADRŽAJ

1. Uvod i problematika

Pretraživanje teksta česta je praktična potreba mnogih informacijskih sustava. Pod terminom "pretraživanje teksta" podrazumijevamo pronalazak svih pojavljivanja nekog niza znakova Q (engl. *query*) unutar drugog niza znakova S (engl. *string*). Tipično se rezultat pretraživanja R formulira kao niz indeksa (rednog broja znaka) unutar niza S na kojem počinje pojavljivanje niza Q. Primjerice, za niz $S =$ *"Žuti pas je opasan kad je opasan remenom oko pasa"* i niz $Q =$ *"pas"* rezultati pretraživanja su $R = \{6, 14, 28, 46\}$.

U području bioinformatike pretraživanje teksta koristi se u za pronalazak specifičnih sekvenci unutar zadanog genoma. Definicija pretraživanja je jednaka. Specifičnost bioinformatičkog pretraživanja jest da su nizovi koji se pretražuju iznimno velike duljine. Primjerice, ljudski genom tipično sadrži oko 3.3×10^9 znakova, što bi otisnuto na A4 stranice fontom veličine 10pt rezultiralo s otprilike milijun stranica.

Postoje mnogi algoritmi pretraživanja teksta koji na jednostavan način ispunjavaju definirane zahtjeve. Iz perspektive računalne složenosti algoritama, jednostavno je implementirati pretraživanje teksta koje radi u linearnom vremenu[1]. Nažalost, za nizove vrlo velike duljine linearno vrijeme znači praktično predugo trajanje pretraživanja. Otud potreba, pogotovo u području bioinformatike, za vremenski sub-linearnim algoritmima pretraživanja.

U ovom projektu razmatramo implementaciju pretraživanja teksta koja se bazira na konceptu FM-indeksa. Konkretna implementacija bazira se na binarnim stablima valića (engl. *wavelet-trees*). Teoretsko razmatranje i praktično testiranje pokazuju da ovakva implementacija pretraživanja ima vremenski sub-linearnu složenost.

[1] Ako nije drukčije navedeno pri razmatranju složenosti pretraživanja uvijek govorimo o složenosti s obzirom na duljinu pretraživanog niza S.

2. FM-indeks algoritam

"Indeksiranje" teksta označava generiranje struktura podataka koje su podrška efikasnom pretraživanju. Za velike tekstove poželjno je da indeks bude memorijski efikasan. FM-indeks [1] pristup je indeksiranju koji ispunjava zahtjeve memorijske efikasnosti i sub-linearnog vremena pretraživanja. Prije nego definiramo FM-indeks, potrebno je razmotriti podatkovne strukture i algoritme koji ga sačinjavaju.

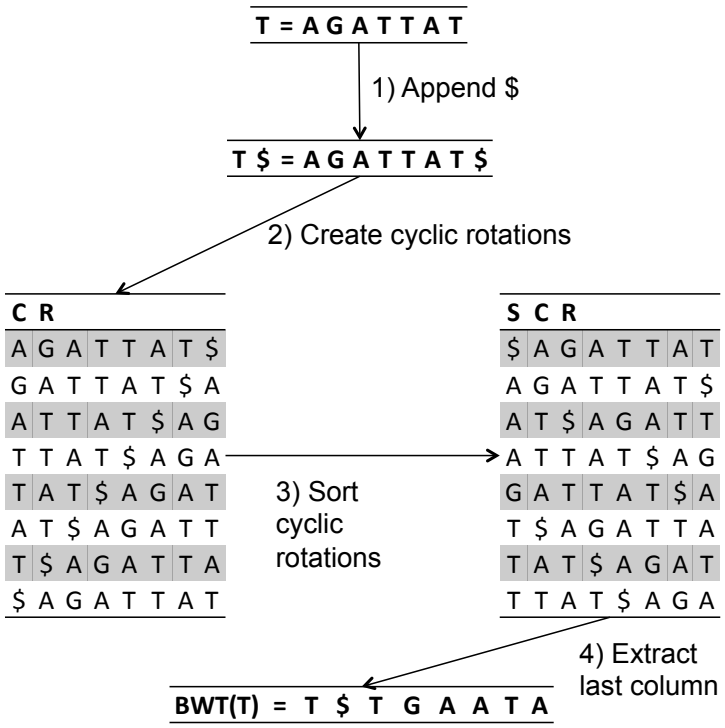
2.1. Burrows-Wheeler transformacija (BWT)

Burrows-Wheeler transformacija [2] transformira niz znakova na način koji će omogućiti efikasnu pohranu i brzo ptreživanje. BWT transformirani niz originalnog teksta T označavati ćemo sa T^{BTW}. Transformacija se provodi sljedećim koracima:

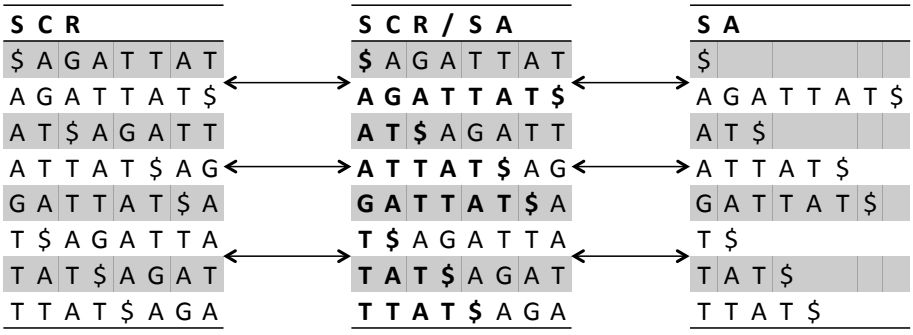
1. Poseban znak \$ koji je leksikografski manji od svih ostalih znakova se dodaje na kraj niza T
2. Cikličkim rotacijama niza T dobiva sa se skup nizova koji čini tablicu CR (engl. *cyclic rotation*)
3. Tablica CR se leksikografski sortira u tablicu SRC
4. Posljednji stupac tablice SCR se ektrahira u rezultat T^{BTW}

Opisani postupak ilustriran je za niz *T = "AGATTAT"* na slici 2.1, preuzetoj iz rada [3].

Bitno je primjetiti kako je BWT transformacija niza srodna sufiksnoj listi SA (engl. *suffix array*). Sufiksna lista je struktura podataka koja se često koristi u algoritmima nad tekstom. Njenu formulaciju nećemo detaljno objašnjavati, materijali na temu su široko dostupni. Sličnost između BWT transformacije i SCR tablice korištene u BWT transformaciji ilustrirana je slikom 2.2.



Slika 2.1: Primjer algoritma BWT transormacije niza *T* = *AGATTAT*



Slika 2.2: SCR tablica BWT tranfsformacije i sufiksna lista za niz *T* = *AGATTAT*

3. Implementacija i testiranje

4. Zaključak

Zaključak.

5. Literatura

[1] Paolo Ferragina and Giovanni Manzini. Opportunistic data structures with applications. pages 390–398, 2000.

[2] M. Burrows, D. J. Wheeler, M. Burrows, and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical report, 1994.

[3] J. Singer. A wawelet tree based fm-index for biological sequences in seqan. Master's thesis, 2012.

6. Sažetak

Sažetak.