Praktický index pre hľadanie takmer najdlhších spoločných podreťazcov založený na bezkontextových gramatikách

autor: Zuzana Skubeňová, školiteľ: Mgr. Adrián Goga

26.6. 2024

Úvod do problematiky

Motivácia

reálny svet:

- obrovské množstvo dát, málo informácií
- vysoká repetitívnosť
- príklad: sekvencie DNA, siete teleskopov
- nové dátové štruktúry

Základné pojmy

Komprimované textové indexy

- dátové štruktúry
- efektívne vyhľadávanie
- rýchly prístup

Maximálne presné zhody (MEMs)

- maximálny podreťazec vzoru P, ktorý sa vyskytuje v indexovanom texte T
- vyhľadávanie pomocou Komprimovaných textových indexov *(Computing MEMs on repetitive text collections, G. Navarro, 2022)*

Základné pojmy

SLP

- špecifický typ bezkontextovej gramatiky
- generuje len jediný konečný reťazec S

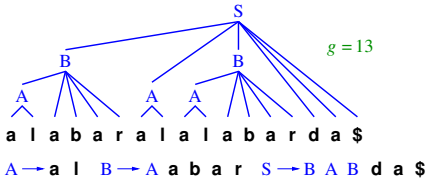

Figure: SLP gramatika

Cieľ práce

- implementovať zjednodušenú verziu indexovej dátovej štruktúry *(ALCS, Gagie a spol., 2023)*, vyhodnotiť na reálnych dátach
- ALCS umožňuje efektívne nájdenie časti najdlhšieho spoločného podreťazca vzoru P a indexovaného textu T
- používa iba priestor úmerný veľkosti komprimovaného textu T
- výsledky - zatiaľ len v teoretickej rovine

Rozdiel článku a našej implementácie

- konštrukcia indexu totožná
- **článok**: nájde približne najdlhší podreťazec vzoru P v texte T
- **naša práca**: nájde približne polovicu najdlhšieho podreťazca P v T

Implementácia

Konštrukcia indexu

- vstup: SLP gramatika textu T (BIGREPAIR), parameter $0 < \varepsilon < 1$
- Karb-Rabin hash: Mersennovo prvočíslo, náhodná konštanta *(Fingerprints in compressed strings, P. Bille a spol., 2017)*
- vypočítame množinu veľkostí blokov $\mathscr{L}$ pomocou parametra ε
- index - hash tabuľka - prefixové a sufixové bloky neterminálov
- uložíme index na disk

Konštrukcia indexu

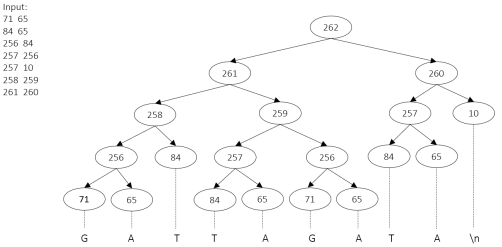

Figure: Príklad vstupu

(a) Prefixy neterminálov

256	[1,1]	G
256	[1,2]	GA
257	[4,4]	T
257	[4,5]	TA
258	[1,2]	GA
258	[1,3]	GAT
259	[4,5]	TA
259	[4,6]	TAG
259	[4,7]	TAGA
260	[8,9]	TA
260	[8,10]	TA\n
261	[1,3]	GAT
261	[1,4]	GATT
261	[1,6]	GATTAG
262	[1,8]	GATTAGAT

(b) Suffixy neterminálov

256	[2,2]	A
256	[1,2]	GA
257	[5,5]	A
257	[4,5]	TA
258	[3,3]	T
258	[2,3]	AT
258	[1,3]	GAT
259	[6,7]	GA
259	[5,7]	AGA
259	[4,7]	TAGA
260	[10,10]	\n
260	[9,10]	A\n
260	[8,10]	TA\n
261	[4,7]	TAGA
261	[2,7]	ATTAGA
262	[8,10]	TA\n
262	[7,10]	ATA\n
262	[5,10]	AGATA\n
262	[3,10]	TTAGATA\n

Table 1: Príklad indexu pre $T = GATTAGATA\backslash n$

Dotazy

- vstup: index, vzor P z disku a parameter L
- hľadáme všetky zhody približne polovičnej požadovanej dĺžky L textu T a vzoru P
- nájdeme najbližšieho predchodcu $\lceil L/2 \rceil$ v $\mathscr{L} = k$
- posuvné okno dĺžky k vzoru P (hash), nájdené zhody s indexom zaznamenáme

Experimenty

Experiment 1

- vplyv parametra $0 < \varepsilon < 1$ na počet blokov indexu
- pre rôzne ε sme skonštruovali index na rôznej dĺžke sekvencií covidu a sledovali sme jeho veľkosť

Experiment 1

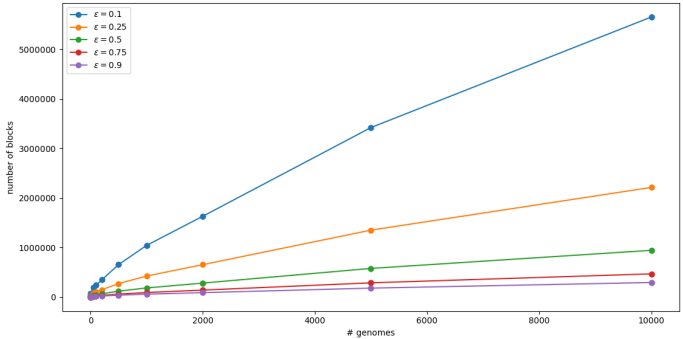

Figure: Vplyv parametra ε a dĺžky sekvencií genómu na výsledný počet blokov

Experiment 2

- vplyv parametra L na nájdené zhody medzi blokmi indexu a 10 rôznymi genómami
- postavený index: $\varepsilon = 0.25$, text: 10k sekvencií covid genómu (BigRePair)

Experiment 2

L	genom1	genom2	genom3	genom4	genom5	genom6	genom7	genom8	genom9	genom10
2	5,227,008	7,136,410	6,346,491	7,095,061	7,134,029	7,081,064	7,084,795	7,133,494	7,130,078	7,010,883
5	55	68	294	135	68	88	294	68	68	294
10	12	16	114	183	16	19	114	16	16	114
25	0	0	2	1	0	2	2	0	0	2
50	0	0	4	1	0	3	4	0	0	4
100	0	0	2	0	0	0	2	0	0	2
200	0	0	2	0	0	0	2	0	0	2

Table: Vplyv veľkosti parametra L na počet nájdených zhôd pre 10 rôznych genómov.

Záver, Budúci vývoj

Záver, Budúci vývoj

Záver

- implementovali sme verziu dátovej štruktúry ALCS, ktorá nájde približne polovicu najdlhšieho podreťazca P v T
- otestovaná na reálnych dátach

Budúci vývoj

- hľadať všetky zhody takmer dĺžky L textu T a vzoru P

Ďakujem za pozornosť.

Odpovede na otázky od oponenta

1. otázka:
   - δ - konštanta, využíva sa v pôvodnom článku ALCS, v našej implementácii sa nepoužíva
2. otázka:
   - ”This index, when given a pattern $P[1 \dots m]$, can find, with high probability, a substring of length $\ell = (1/(1-\epsilon))^k$ of P that occurs in T in $O(m \log^{\delta} g)$ time.”
   - ”We show that, on α-balanced grammars, this can be solved with high probability in time $O(m \log^{\delta} g)$ for any fixed constant $\delta > 0$.” (ALCS, T. Gagie a spol., 2023)
3. otázka:
   - index (databáza): text T - 10k genómov covidu
   - vzor P: 1 genóm z indexu, $|P|$: 26927 pre $L = 2 \cdot |P|$
   - výstup: ” l 20670 ” - veľkost bloku tiež 20670 $> \frac{2}{3} \cdot |P|$