Plánovanie tunelov v Burrows-Wheelerovej transformácii pomocou celočíselného lineárneho programovania

Klára Sládečková
Školiteľ: Andrej Baláž, MSc.

Burrows-Wheelerova transformácia

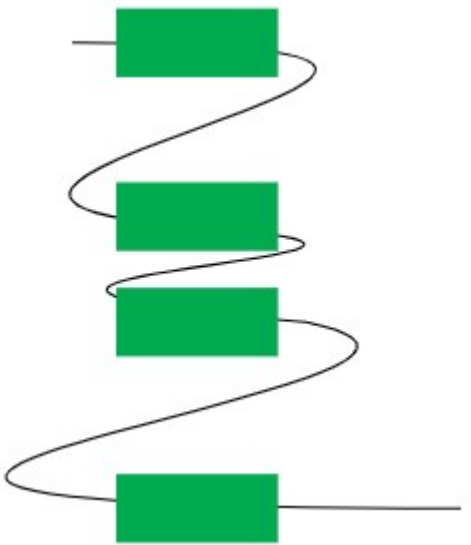
- Posledný stĺpec matice zoradených cyklických permutácií
- Bezstratová transformácia textu
- Komprimačné vlastnosti - Tunelovanie

$readysteadygo\$$

i	SA[i]	$F[i]$	$S[SA[i]]...S[13]S[0]...S[SA[i]-1]$	$L[i]$
0	13	$	$ readysteadygo	o
1	8	a	adygo$readyste	e
2	2	a	adysteadygo$re	e
3	9	d	dygo$readystea	a
4	3	d	dysteadygo$rea	a
5	7	e	eadygo$readyst	t
6	1	e	eadysteadygo$r	r
7	11	g	go$readysteady	y
8	12	o	o$readysteadyg	g
9	0	r	readysteadygo$	$
10	5	s	steadygo$ready	y
11	6	t	teadygo$readys	s
12	10	y	ygo$readystead	d
13	4	y	ysteadygo$read	d

Bloky v BWT

- Opakujúci sa vzor
- Stĺpce pozostávajú z práve jedného písmena (riadky sú ekvivalentné)

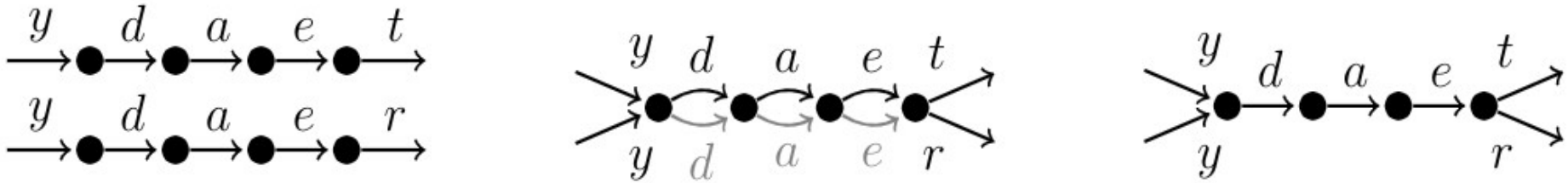


Príklad bloku

$readysteadygo\$$

i	SA[i]	$F[i]$	$S[SA[i]]...S[13]S[0]...S[SA[i]-1]$	$L[i]$
0	13	$	$ readysteadygo	o
1	8	a	adygo$readyste	e
2	2	a	adysteadygo$re	e
3	9	d	dygo$readystea	a
4	3	d	dysteadygo$rea	a
5	7	e	eadygo$readyst	t
6	1	e	eadysteadygo$r	r
7	11	g	go$readysteady	y
8	12	o	o$readysteadyg	g
9	0	r	readysteadygo$	$
10	5	s	steadygo$ready	y
11	6	t	teadygo$readys	s
12	10	y	ygo$readystead	d
13	4	y	ysteadygo$read	d

Tunelovanie bloku

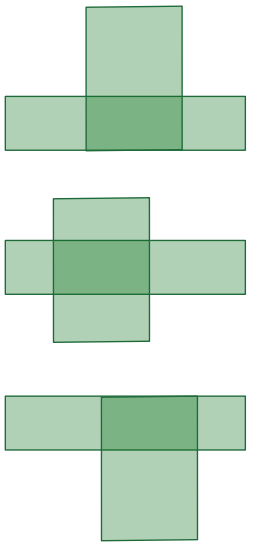


- Skóre bloku:
  - Číslo označujúce veľkosť benefitu tunelovania daného bloku
  - Napríklad počet pozícií, ktoré sa tunelovaním vymažú z BWT

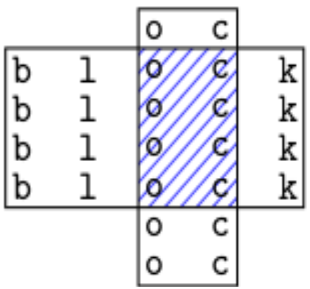
Kolízie blokov

Bloky kolidujú ak zdieľajú spolu nejakú pozíciu.
Kolidujú:

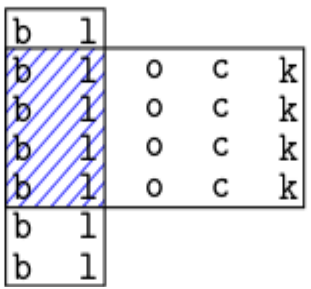
- Kompenzovateľne, ak:
  - Nie sú rovnako široké
  - Prvý a posledný stĺpec širšieho bloku nie je zdieľaný
  - Užší blok je vyšší ako širší a nezdieľa aspoň jeden riadok so širším blokom
- Kriticky, inak



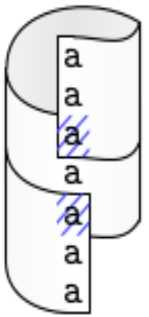
Kolízie blokov



(a) kompenzovateľná kolízia



(b) kritická kolízia



(c) kritická samo-kolízia

Problém plánovania tunelov

- Žiadne kritické kolízie a maximalizovaná kompresia
- NP-ťažký problém
- Rôzne heuristiky obmedzené na “run-based” bloky
- → redukcia na ILP

Postup

- VSTUP: textový reťazec
- BWT
- Výpočet blokov, skóre blokov, kolízie
- Redukcia
- Výpočet pomocou ILP solvera
- Tunelovanie
- VÝSTUP: tunelovaná BWT

Celočíselné lineárne programovanie

- NP-optimalizačný problém
- VSTUP:
  - Cieľ : *min* alebo *max*
  - Hlavná suma tvaru $\sum_{k=1}^{n} a_n x_n$
  - Konečne veľa podmienok tvaru $\sum_{j=1}^{m} b_j x_{i_j} \leq C$
  - Obmedzenia ohraničujúce premenné $x_1, x_2, ..., x_n$
- VÝSTUP:
  - Hodnota (celočíselná) premenných $x_1, x_2, ..., x_n$, pri ktorých hlavná suma nadobúda maximum/minimum a všetky podmienky sú dodržané

ILP - príklad

- VSTUP:
  - Cieľ : min
  - Hlavná suma: $-2x_1 + 0.5x_2 - 200x_3$
  - Podmienky: $x_1 + 2x_2 \leq 3$
  - Obmedzenia: $-5 \leq x_1 \leq 5, \quad 0 \leq x_2, \quad x_3 \leq -1$
- VÝSTUP:
  - Minimálna hodnota je 194 pri ohodnotení

$$x_1 = 3, \quad x_2 = 0, \quad x_3 = -1$$

Redukcia

Bloky, skóre blokov, dvojice kolidujúcich blokov a typ kolízií

- Maximalizácia celkového skóre
- Kritické kolízie
- Správne skóre pre kompenzovateľné kolízie

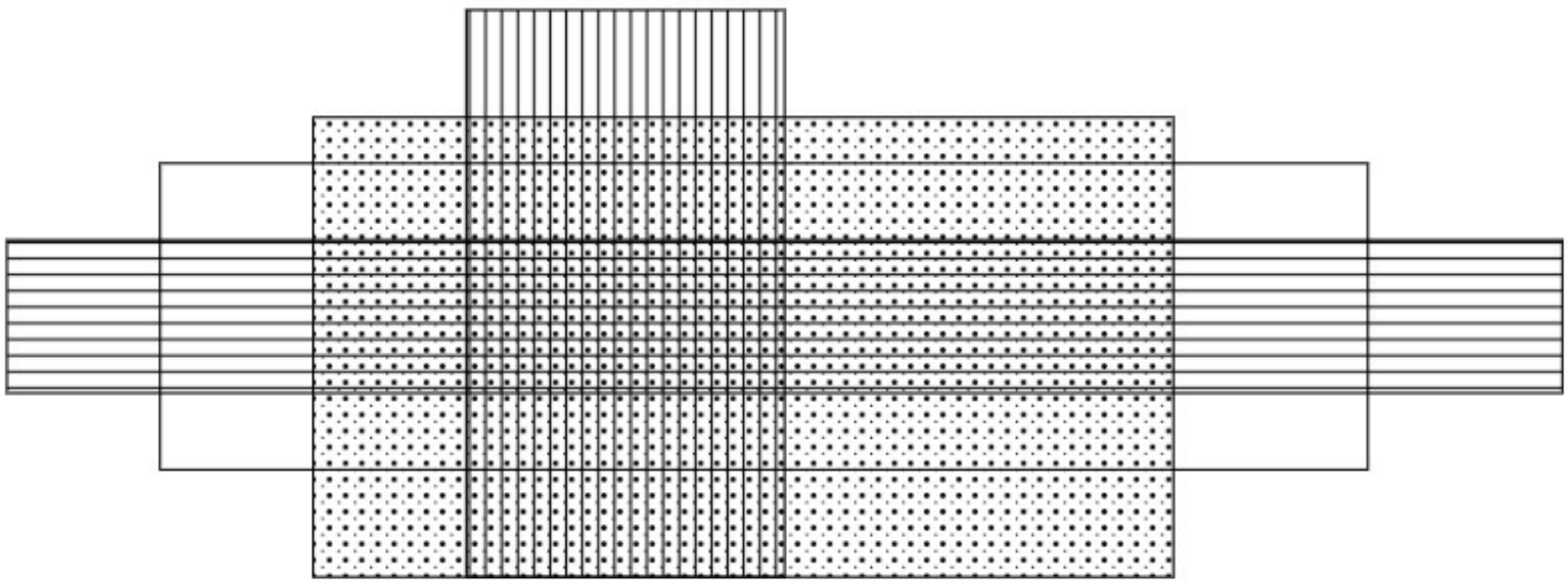
Skóre (prvá verzia)

- Maximalizácia skóre $p_0x_0 + p_1x_1 + \ldots + p_kx_k$
- k blokov
- p_i → cena i-teho bloku a
- x_i → bin premenná (true ↔ blok i naplánovaný na tunelovanie)

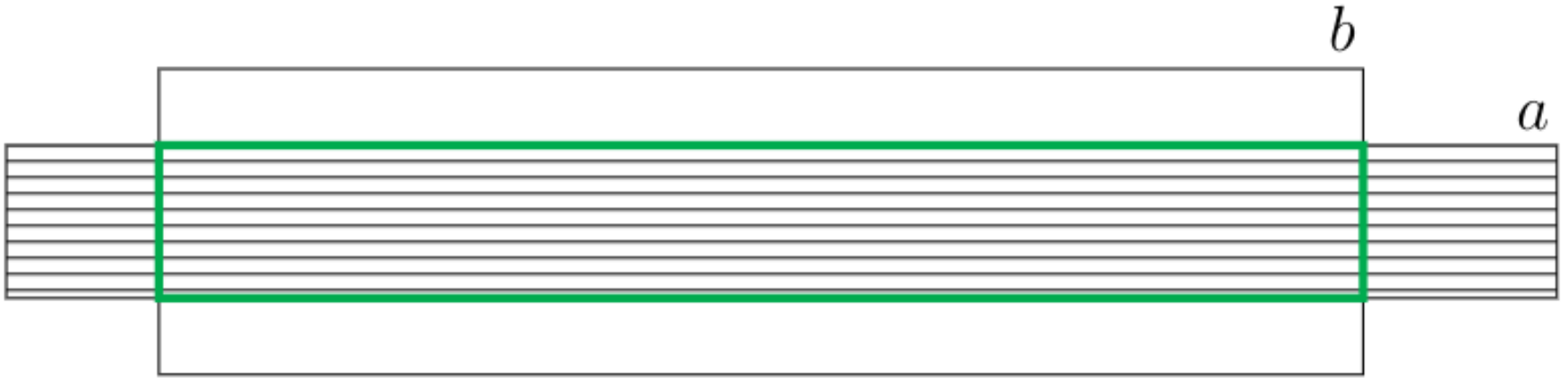
Kritické kolízie

$$x_i + x_j \leq 1$$

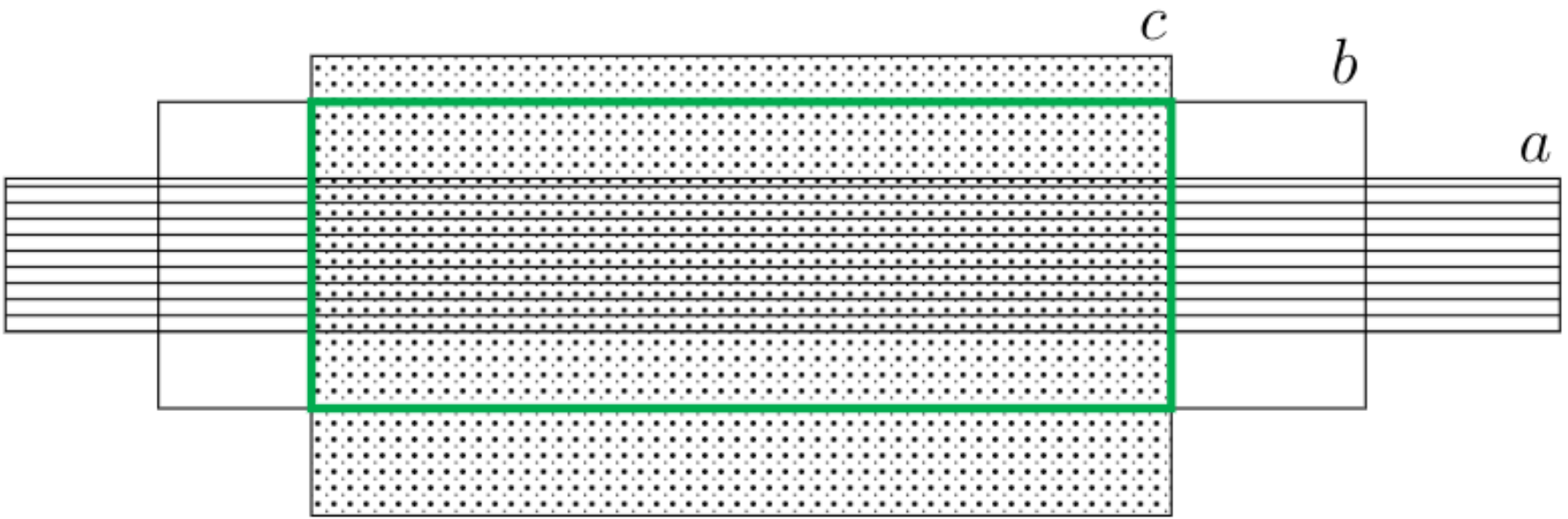
Kompenzovateľné kolízie



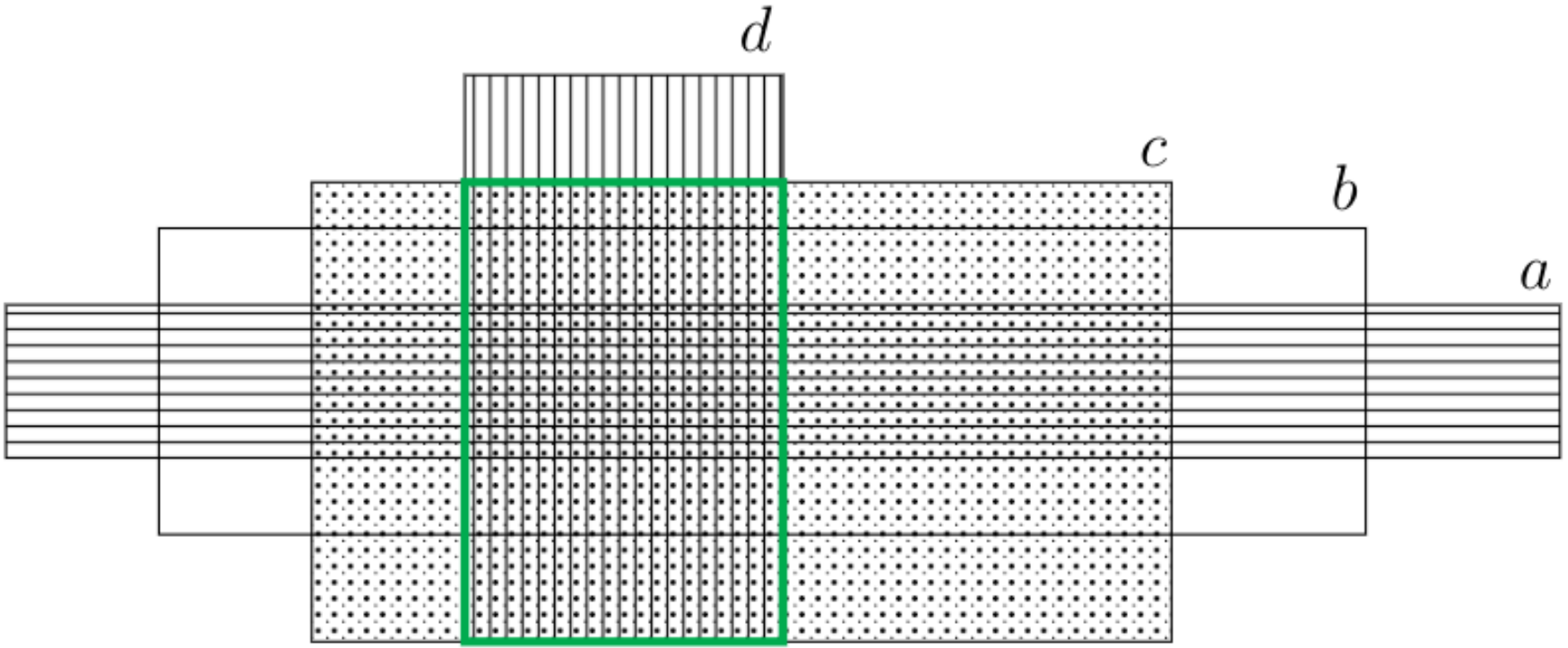
Kompenzovateľné kolízie



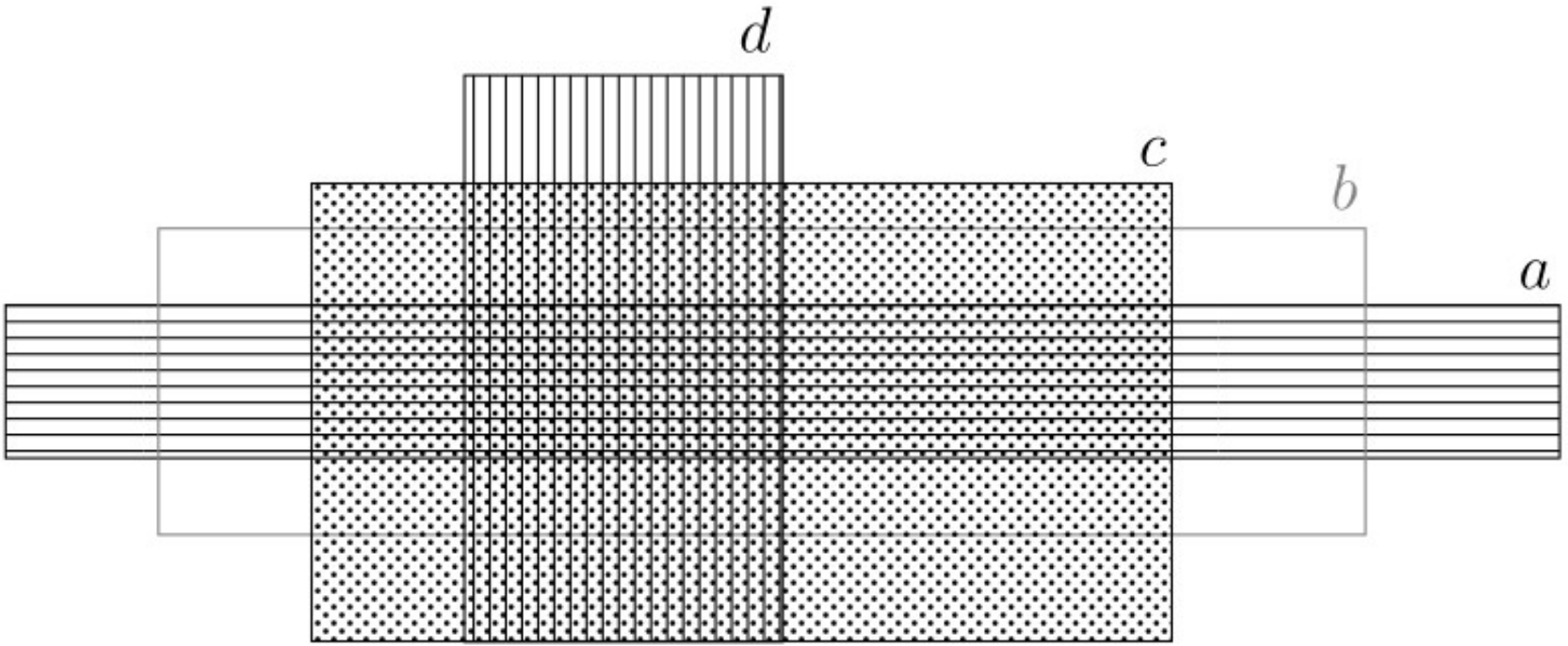
Kompenzovateľné kolízie



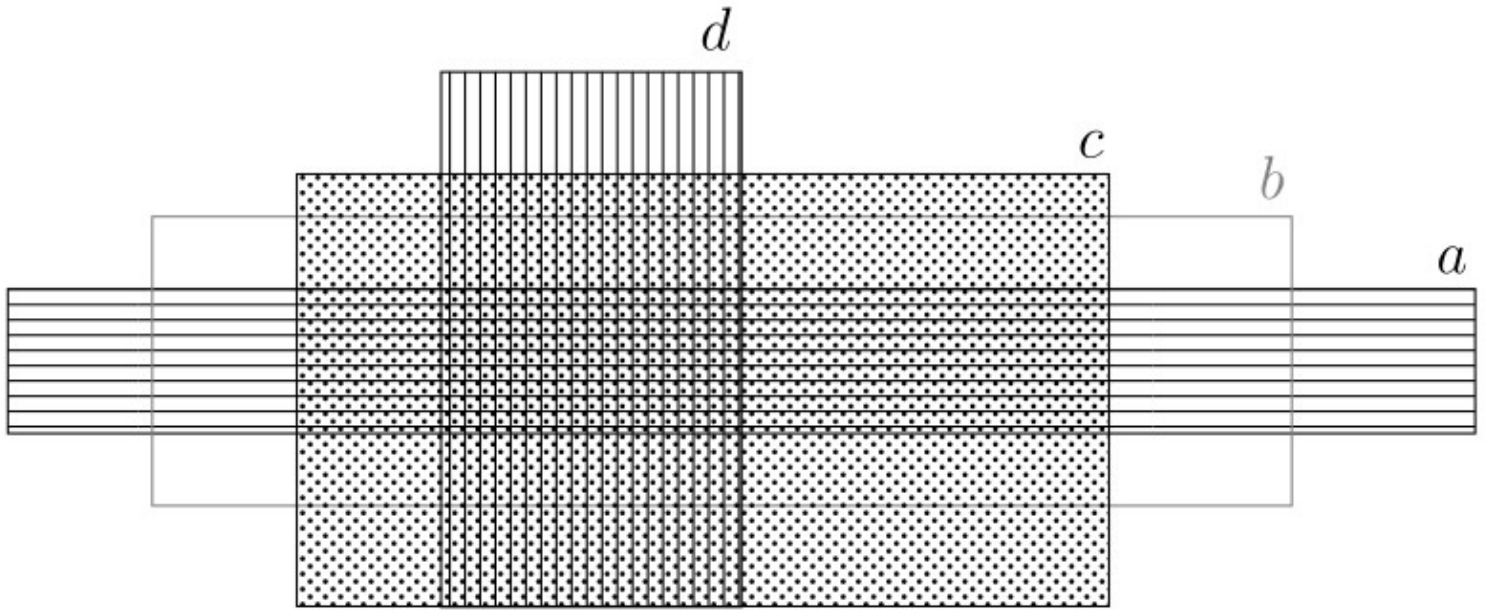
Kompenzovateľné kolízie



Kompenzovateľné kolízie



Kompenzovateľné kolízie



$$\texttt{score}(a) + \texttt{score}(c) + \texttt{score}(d) - \texttt{score}(a \cap c) - \texttt{score}(c \cap d)$$

Kompenzovateľné kolízie

- y_i → binárna premenná

  → true ↔ tunelujú sa oba bloky i-teho páru a tieto dva bloky tvoria bajbližší pár

- $\{z_1^i, z_2^i, \ldots, z_l^i\}$ → bloky "medzi" blokmi i-teho páru s_i a r_i

$$x_{r_i} + x_{s_i} + \sum_{j=1}^{l}(1 - x_{z_j^i}) \geq l + 2 \iff y_i \geq 1$$

Skóre (finálna verzia)

- Maximalizácia skóre

$$p_0x_0 + p_1x_1 + \ldots + p_kx_k - shp_0y_0 - shp_1y_1 - \ldots - shp_my_m$$

- p_i , x_i ako predtým
- shp_i → cena prekryvu i-teho páru komp. kol. blokov
- y_i → binárna premenná

  → true ↔ tunelujú sa oba bloky i-teho páru a tieto dva bloky tvoria “najbližší” pár

Kompenzovateľné kolízie

$$x_{r_i} + x_{s_i} + \sum_{j=1}^{l}(1 - x_{z_j^i}) \geq l + 2 \iff y_i \geq 1$$

- $x_{r_i} + x_{s_i} + \sum_{j=1}^{l}(1 - x_{z_j}) - (l + 2) < (l + 2) \cdot y_i \quad \Rightarrow$
- $x_{r_i} + x_{s_i} + \sum_{j=1}^{l}(1 - x_{z_j}) \geq (l + 2) \cdot y_i \quad \Leftarrow$

Experimentálne výsledky

Input file	Initial size	~Opt	Tunneled size	Size ratio	Time	Strategy
example.txt	18	10	10	0.56	0.1	ILPR
			12	0.67	0.1	dBGEM
protein.fasta	5109	5019	5019	0.98	0.36	ILPR
			5076	0.99	0.06	dBGEM
zinc_fingers.fa	10345	9031	9050	0.87	1111	ILPR
			10029	0.97	0.1	dBGEM
bacteriophage.fasta	34041	29796	29796	0.88	7.5 hours	ILPR
			33343	0.98	0.1	dBGEM
S-cereale.fasta	6837	5567	5569	0.81	730	ILPR
			6288	0.92	0.06	dBGEM
human_alphoid.fasta	2993	1678	1944	0.65	360	ILPR
			2527	0.84	0.06	dBGEM
huYchr.fasta	3693	2166	2370	0.64	520	ILPR
			3518	0.95	0.06	dBGEM
chrom21_rep.fasta	20001	2829	4431	0.22	15 hours	ILPR
			5022	0.25	0.06	dBGEM
repetitive.txt	3019	-4257	622	0.21	880	ILPR
			1881	0.62	0.06	dBGEM

ILPR = ILP reduction

dBGEM = de Bruijn graph edge minimization

~Opt = dolná hranica kompresie

Ďakujem za pozornosť