EFFICIENT CONSTRUCTION OF A COMPRESSED INDEX FOR LARGE TEXT COLLECTIONS

Študent: Klára Sládečková
Školiteľ: doc. Mgr. Bronislava Brejová, PhD.
Konzultant: Andrej Baláž, MSc.

CIEĽ

- KOMPRESIA
- Biologické dáta sú repetívne
- Zameranie sa na pangenomické dáta – množina sekvencií rôznych jedincov toho istého druhu majúcich rovnaký význam

BWT MATICA

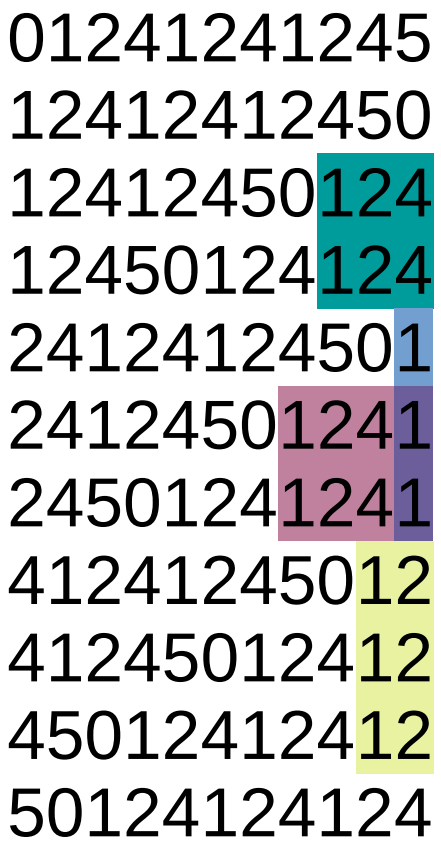
- Matica lexikograficky zoradených rotácií pôvodného reťazca

12450 →

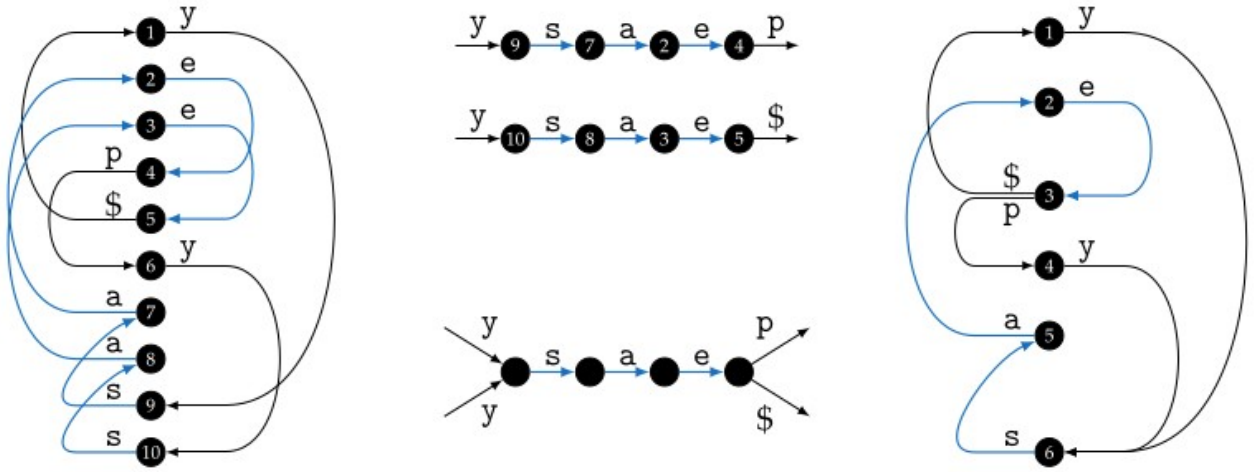
01245
12450
24501
45012
50124

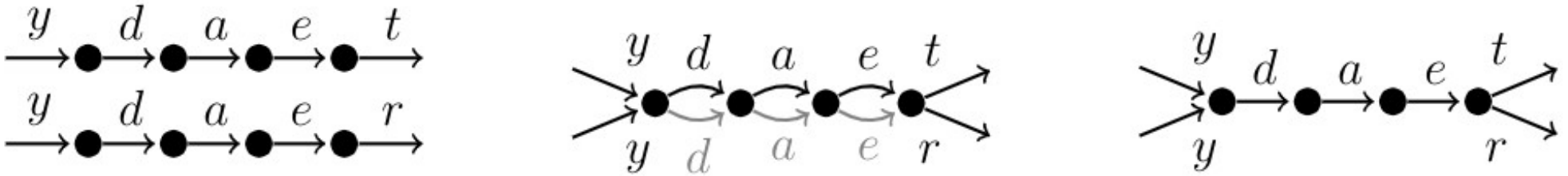
BLOK

- Obdĺžnik zarovnaný k pravému kraju BWT matice
- Riadky sú ekvivalentné
- (Uvažujeme len maximálne bloky na výšku aj šírku)



TUNELOVANIE





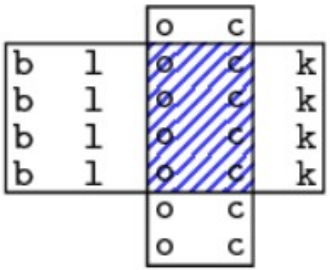
CENA BLOKU

- |bwt| - |t-bwt|
- w – šírka bloku b
- h – výška bloku b
- (cena(b) = (w-1)*(h-1))

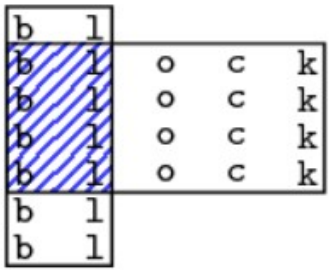
01241241245
12412412450
12412450124
12450124124
24124124501
24124501241
24501241241
41241245012
41245012412
45012412412
50124124124

3

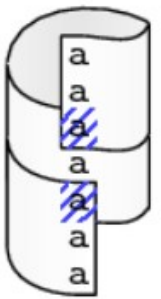
KOLÍZIA BLOKOV



a) compensable collision



(b) critical collision



(c) critical self-collision

PROBLÉM VÝBERU BLOKOV

- Zostroj množinu blokov tak, aby
  - Žiaden blok nebol samo-kolidujúci
  - Žiadna dvojica blokov v množine nebola kriticky kolidujúca
  - Veľkosť ztunelovanej bwt bola minimálna
- NP-úplný problém

U. Baier and K. Dede, "BWT Tunnel Planning is Hard But Manageable," 2019 Data Compression Conference (DCC), Snowbird, UT, USA, 2019, pp. 142-151, doi: 10.1109/DCC.2019.00022. keywords: {Tunneling;Data compression;Transforms;Complexity theory;Planning;Compressors;Tools;Burrows Wheeler transform;data compression;tunneling},

DE BRUIJN GRAPH EDGE MINIMIZATION

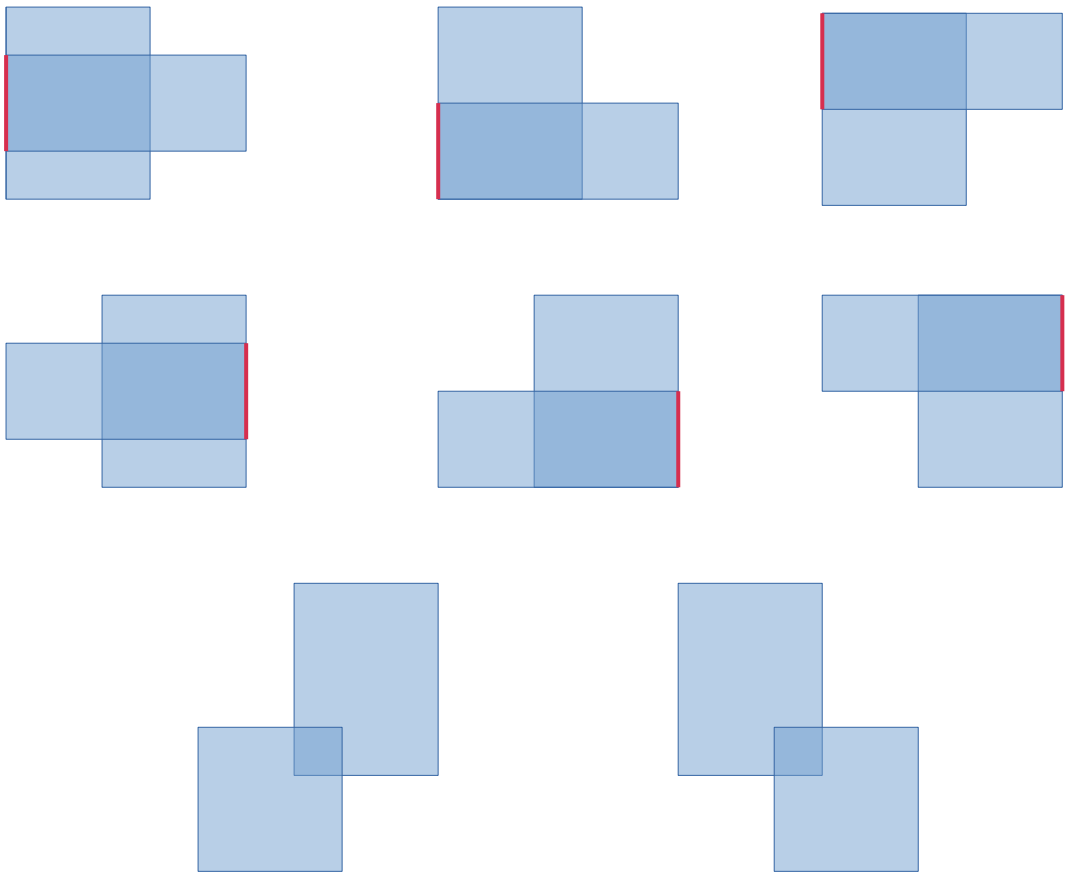
- State-of-the-art algoritmus
- 2020
- Priemerný čas – *O(n*log s)*
- Neuvažuje kompenzovateľné kolízie

Uwe Baier, Thomas Büchler, Enno Ohlebusch, Pascal Weber, Edge minimization in de Bruijn graphs, Information and Computation, Volume 285, Part B, 2022, 104795, ISSN 0890-5401,

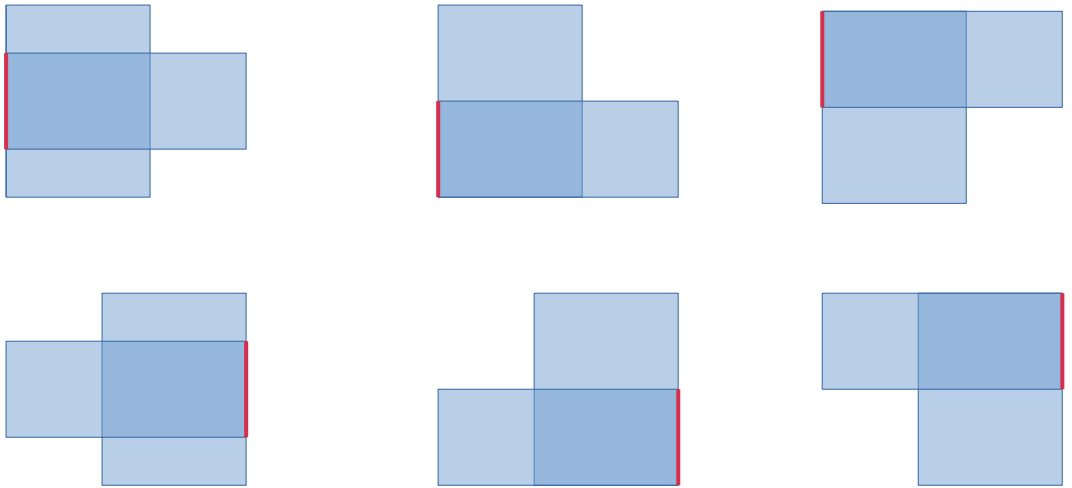
HEURISTIKA

- Vypočítaj maximálne bloky ($O(n)$)
- Zbav sa samo-kolidujúcich blokov ($O(n*w\_max*log(w\_max))$)
- Zbav sa kritických kolízií

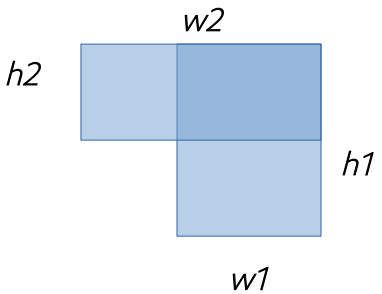
KRITICKÉ KOLÍZIE

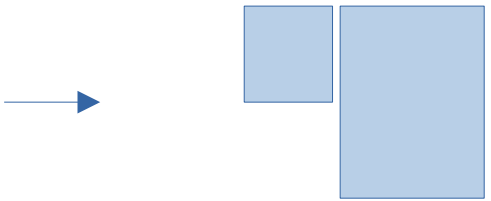


KRITICKÉ KOLÍZIE

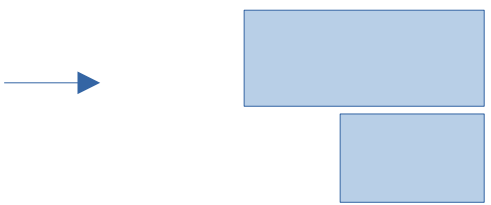


KRITICKÉ KOLÍZIE

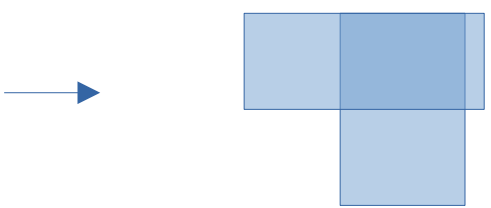




VERTIKÁLNE ROZDELENIE

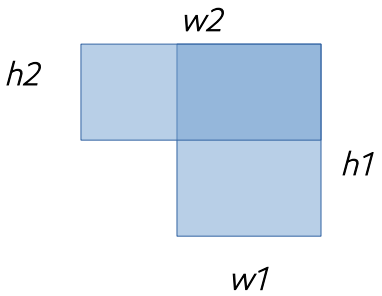


HORIZONTÁLNE ROZDELENIE

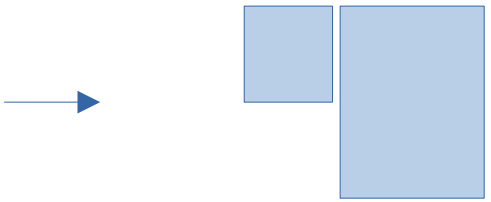


ZÚŽENIE

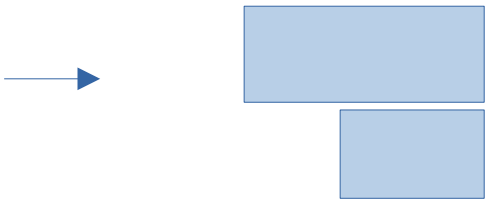
KOLÍZIE V PRAVOM STĹPCI



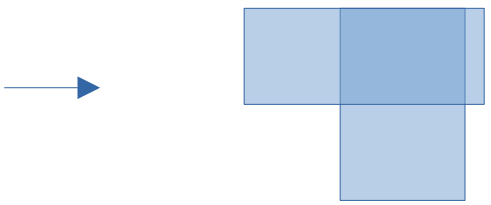
(h2-1) * (w2-1) -
(w1-1)*(h1-h2)



- h2

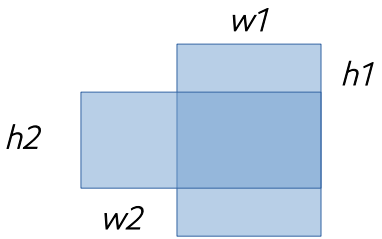


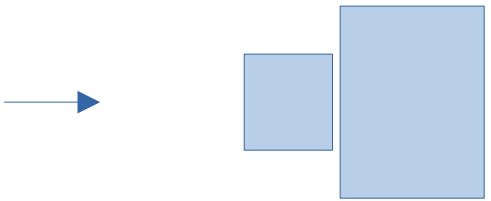
- w1



- (h1-h2)

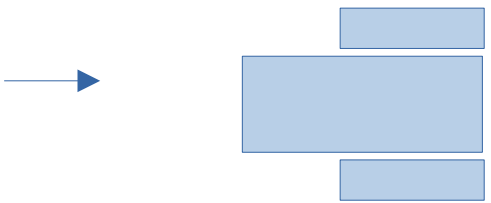
COLLISIONS ON THE RIGHT



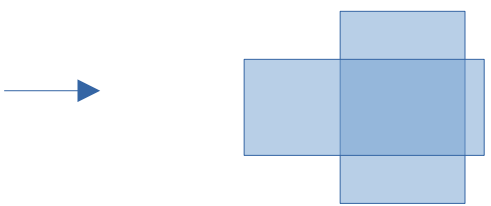


- h2

(h2-1) * (w2-1) -
(w1-1)*(h1-h2)

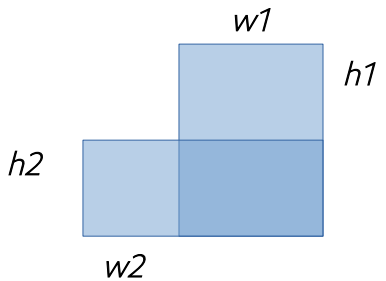


- 2*w1

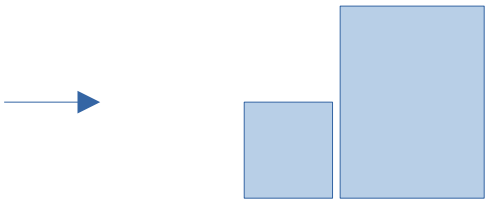


- (h1-h2)

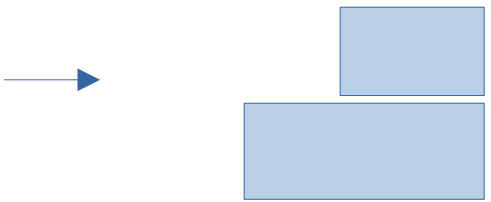
COLLISIONS ON THE RIGHT



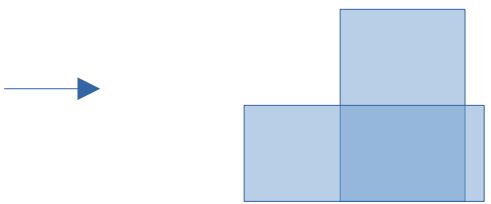
(h2-1) * (w2-1) -
(w1-1)*(h1-h2)



- h2

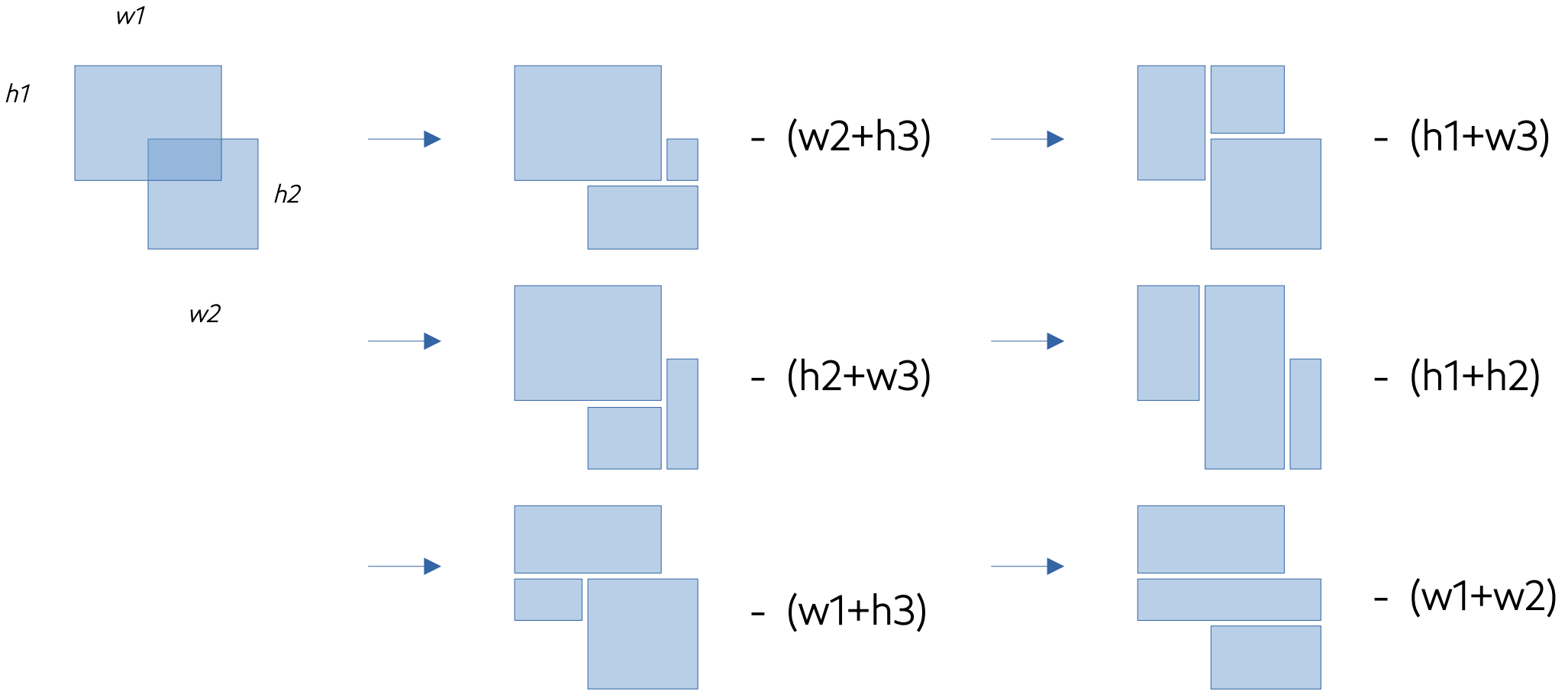


- w1

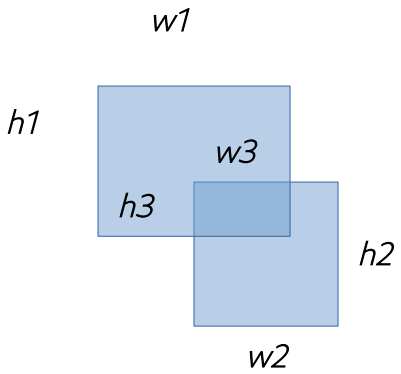


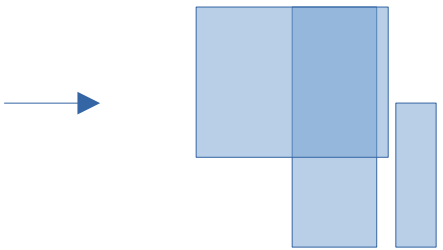
- (h1-h2)

ROHOVÉ KOLÍZIE

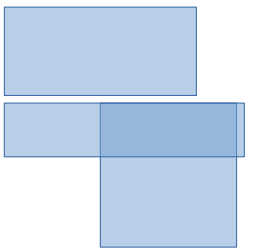


ROHOVÉ KOLÍZIE

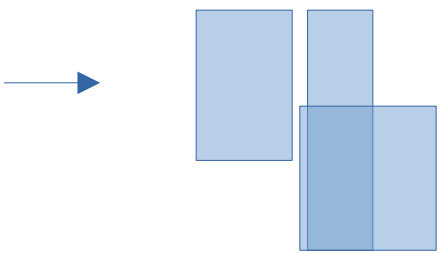




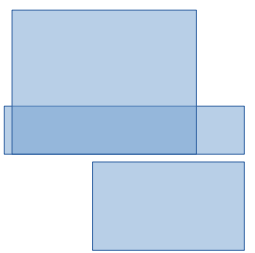
- (2*h2-h3)



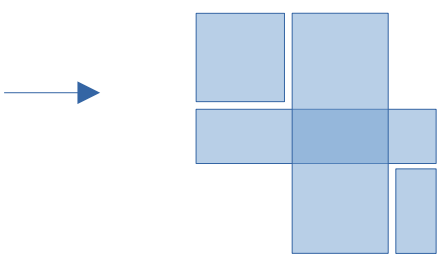
- (w1+h2-h3)



- (2*h1-h3)



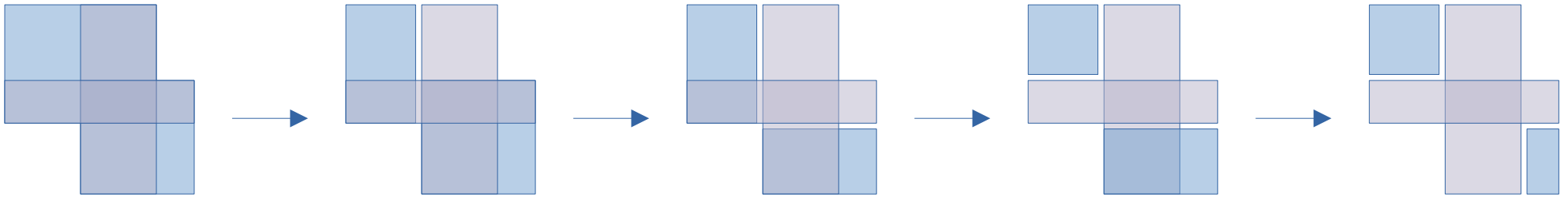
- (w2+h1-h3)

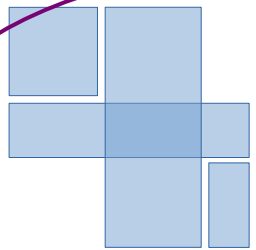


- (h1+h2-2*h3 + w1+w2-2*w3)

ROHOVÉ KOLÍZIE

ROHOVÉ KOLÍZIE

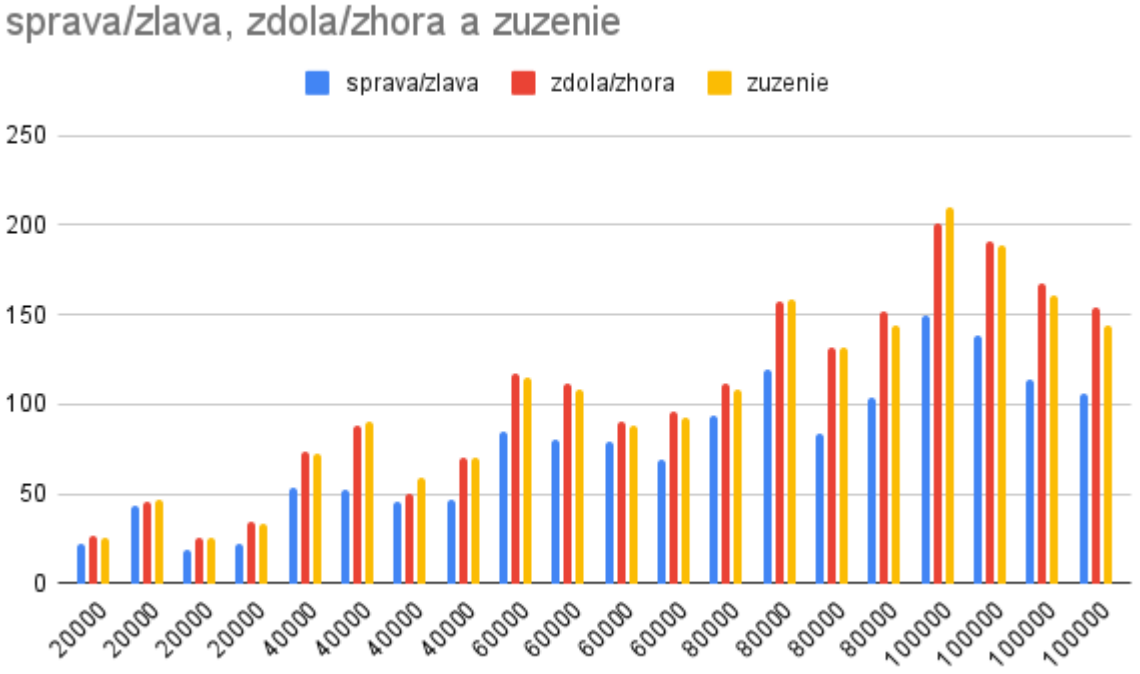


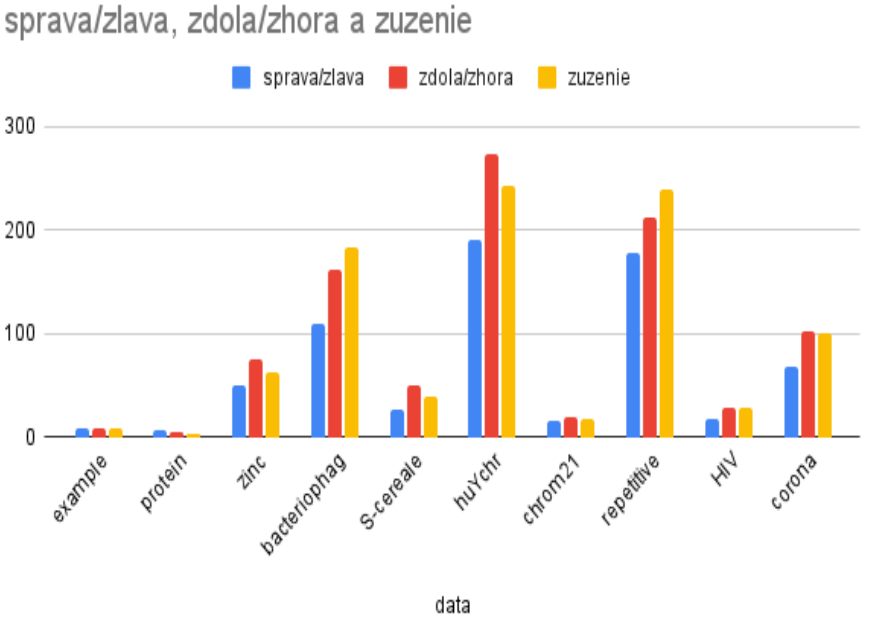


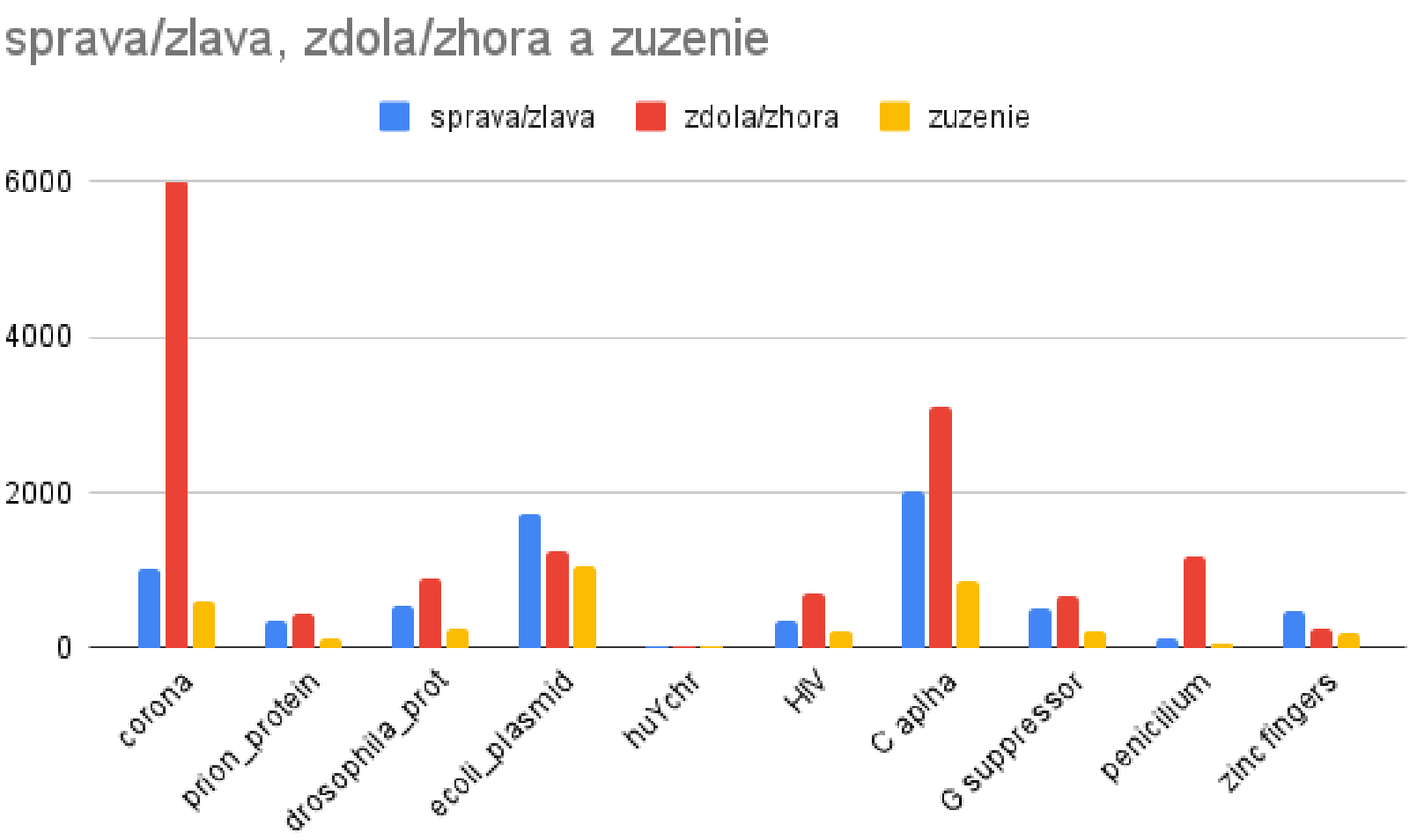
- (h1+h2-2*h3 + w1+w2-2*w3)

MOŽNÉ KOMBINÁCIE

- Horizontálne rozdelenie + zúženie
- Vertikálne rozdelenie + zúženie
- Horizontálne rozdelenie + vertikálne rozdelenie
- Vertikálne rozdelenie + vertikálne rozdelenie
- (horizontálne rozdelenie + horizontálne rozdelenie)


sprava/zlava, zdola/zhora a zuzenie
sprava/zlava
zdola/zhora
zuzenie
250
200
150
100
50
0
20000
20000
20000
20000
40000
40000
40000
40000
60000
60000
60000
60000
80000
80000
80000
80000
100000
100000
100000
100000


sprava/zlava, zdola/zhora a zuzenie
sprava/zlava
zdola/zhora
zuzenie
300
200
100
0
example
protein
zinc
bacteriophag
S-cereale
huYchr
chrom21
repetitive
HIV
corona
data


sprava/zlava, zdola/zhora a zuzenie
sprava/zlava
zdola/zhora
zuzenie
6000
4000
2000
0
corona
prion_protein
drosophila_prot
ecoli_plasmid
huYchr
HIV
C alpha
G suppressor
penicillium
zinc fingers

Input file	Initial size	Tunneled size	Size ratio	Time	Strategy
		12	0.67	0.1	dBGEM
example.txt	18	10	0.56	0.001	heuristic3
		5076	0.99	0.06	dBGEM
protein.fasta	5109	5043	0.99	0.2	heuristic
		10029	0.97	0.1	dBGEM
zinc_fingers.fa	10345	9697	0.94	3.3	heuristic
		33343	0.98	0.1	dBGEM
bacteriophage.fasta	34041	32155	0.94	38	heuristic
		6288	0.92	0.06	dBGEM
S-cereale.fasta	6837	6089	0.89	1.3	heuristic
		3518	0.95	0.06	dBGEM
huYchr.fasta	3693	3022	0.82	0.19	heuristic
		5022	0.25	0.06	dBGEM
chrom21_rep.fasta	20001	4699	0.23	0.87	heuristic
		1881	0.62	0.06	dBGEM
repetitive.txt	3019	796	0.26	0.02	heuristic
		9400	0.33	0.08	dBGEM
HIV.txt	28060	8731	0.31	2.8	heuristic

Input file	Initial size	Tunneled size	Size ratio	Time	Strategy
example.txt	18	12	0.67	0.1	dBGEM
		10	0.56	0.001	heuristic3
corona_virus.fasta	275303	58294	0.21	0.21	dBGEM
		43458	0.16	75	heuristic
prion_protein.fasta	15614	5881	0.38	0.1	dBGEM
		4854	0.31	0.8	heuristic
drosophila_protein.fasta	23134	9600	0.41	0.9	dBGEM
		8494	0.37	2.1	heuristic
ecoli_plasmid.fasta	190975	114653	0.60	0.3	dBGEM
		104845	0.55	447	heuristic
huYchr.fasta	4470	3837	0.86	0.06	dBGEM
		3652	0.82	0.5	heuristic
HIV.fasta	19706	9027	0.46	0.08	dBGEM
		8297	0.42	2	heuristic
Calpha.fasta	94119	42034	0.45	0.15	dBGEM
		40039	0.43	59.1	heuristic
G_suppressor.fasta	14048	5379	0.38	0.07	dBGEM
		5050	0.36	0.6	heuristic
penicilium.fasta	8640	3292	0.38	0.06	dBGEM
		2368	0.27	0.21	heuristic
zinc_fingers.fasta	22981	9305	0.40	0.08	dBGEM
		9197	0.40	2.4	heuristic

ĎAKUJEM ZA POZORNOSŤ