EFEKTÍVNA KONŠTRUKCIA KOMPRIMOVANÉHO INDEXU PRE VEĽKÉ ZBIERKY TEXTOV

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

MOTIVÁCIA

- Rastúci počet dát a aj potreba ich efektívneho spravovania
- Biologické dáta sú repetitívne
- Pangenomické dáta – množina sekvencií (genómov) z rôznych jedincov toho istého druhu
- Komprimovaná dátová štrúktúra podporujúca indexovanie

BWT MATICA

- Matica lexikograficky zoradených rotácií pôvodného reťazca

$readysteadygo\$ \longrightarrow$

F	L
$\$ r e a d y s t e a d y g o$	
$a d y g o \$ r e a d y s t e$	
$a d y s t e a d y g o \$ r e$	
$d y g o \$ r e a d y s t e a$	
$d y s t e a d y g o \$ r e a$	
$e a d y g o \$ r e a d y s t$	
$e a d y s t e a d y g o \$ r$	
$g o \$ r e a d y s t e a d y$	
$o \$ r e a d y s t e a d y g$	
$r e a d y s t e a d y g o \$$	
$s t e a d y g o \$ r e a d y$	
$t e a d y g o \$ r e a d y s$	
$y g o \$ r e a d y s t e a d$	
$y s t e a d y g o \$ r e a d$	

BWT

- Burrows-Wheelerová transformácia (stĺpec L)
- Lineárna konštrukcia
- Lineárna rekonštrukcia pôvodného reťazca
- Možnosť kompresie – vďaka zoskupovaniu rovnakých kontextov do súvislých sekcií
- Rýchle vyhľadávanie $O(m + occ)$

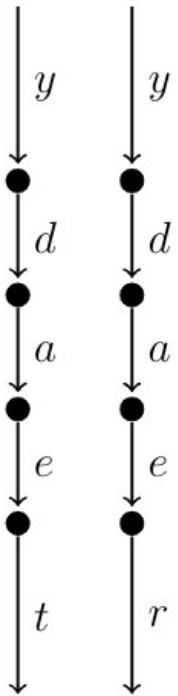
BLOK

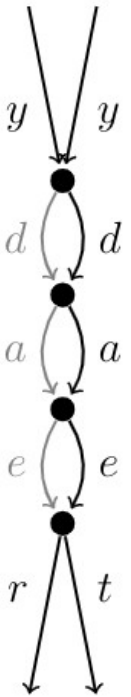
- Hlavná jednotka kompresie
- Skupina rovnakých podreťazcov pôvodného reťazca

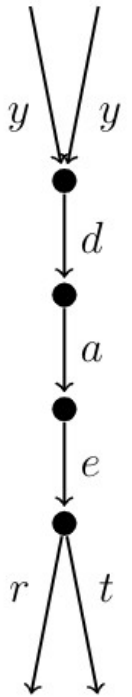
$readysteadygo\$ \longrightarrow$

- Uvažujeme iba maximálne bloky (na výšku aj šírku)

TUNELOVANIE



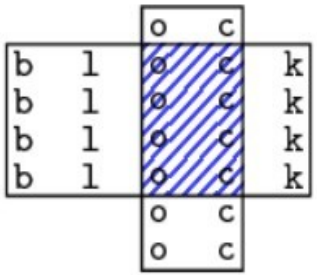




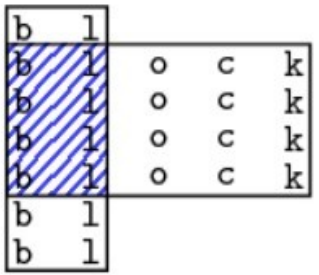
PROFIT BLOKU

- Výhoda tunelovania daného bloku (alebo množiny blokov)
- | BWT | - | TBWT |
- $w \cdot (h-1)$

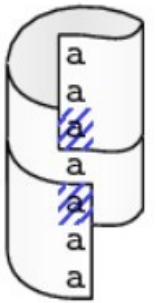
KOLÍZIA BLOKOV



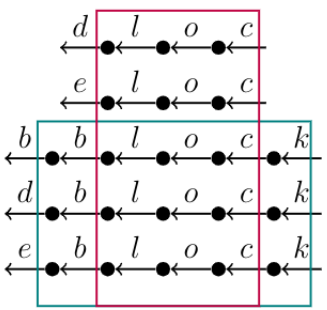
a) kompenzovateľná kolízia

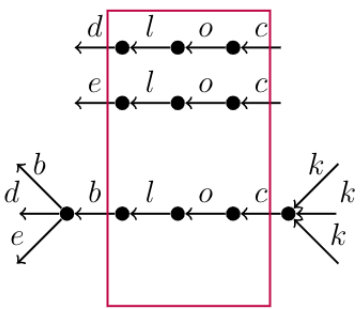


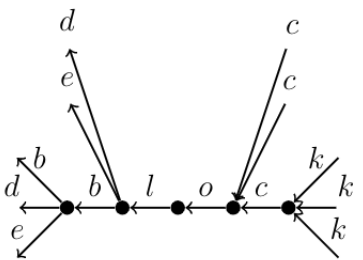
b) kritická kolízia



c) kritická samokolízia







PROBLÉM VÝBERU BLOKOV

- Množina blokov, ktorá neobsahuje kritickú kolíziu a maximalizuje profit tunelovania
- 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},

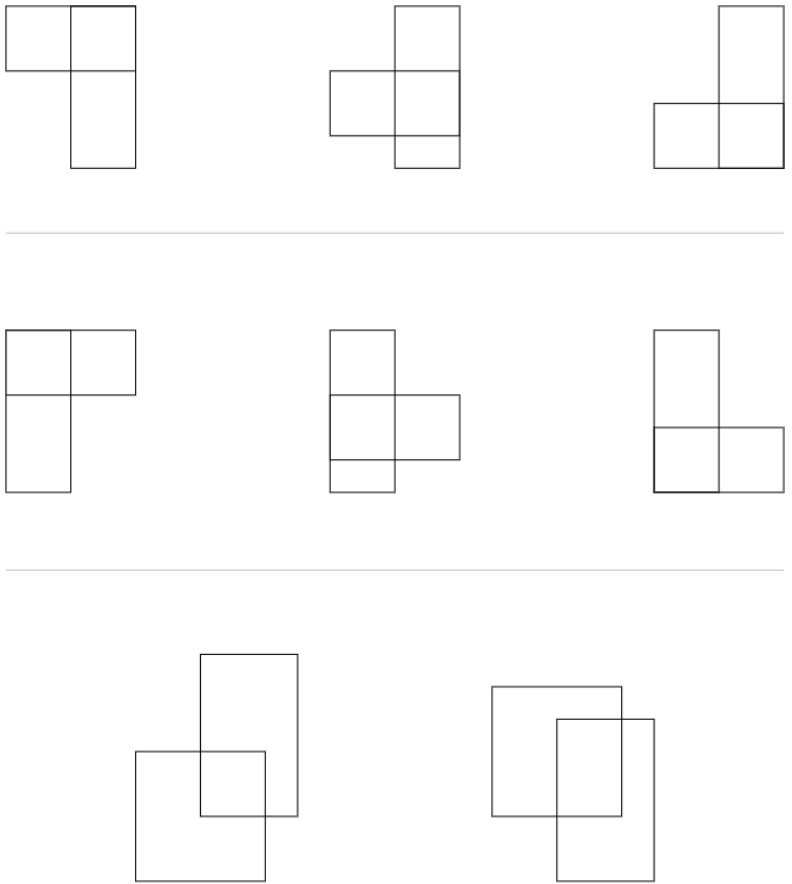
DE BRUIJN GRAPH EDGE MINIMIZATION

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

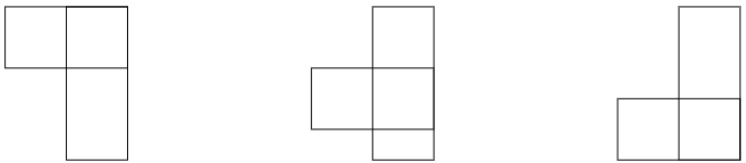
Uwe Baier, Thomas Büchler, Enno Ohlebusch, Pascal Weber, Edge minimization in de Bruijn graphs, Information and Computation, Volume 285, Part B, 2022, 104795, ISSN 0890-5401,

NAŠA HEURISTIKA

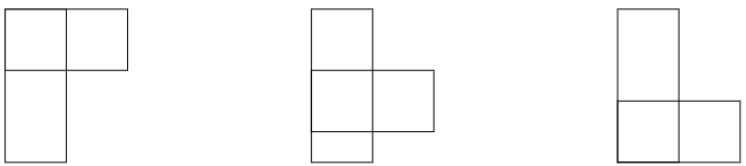
KRITICKÉ KOLÍZIE



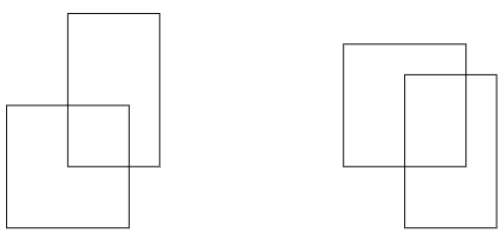
KRITICKÉ KOLÍZIE



→ vpravo zarovnané (kritické) kolízie

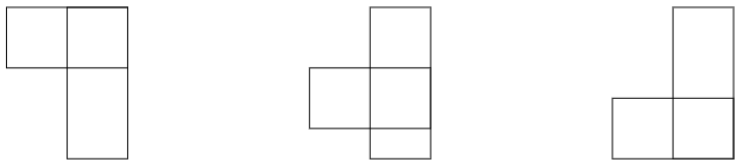


→ vľavo zarovnané (kritické) kolízie

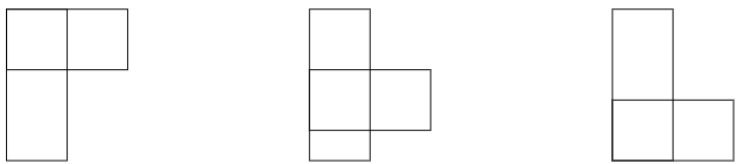


→ rohové (kritické) kolízie

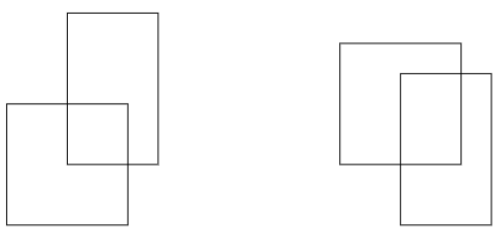
KRITICKÉ KOLÍZIE



$$\mathcal{O}(N \log N + m)$$

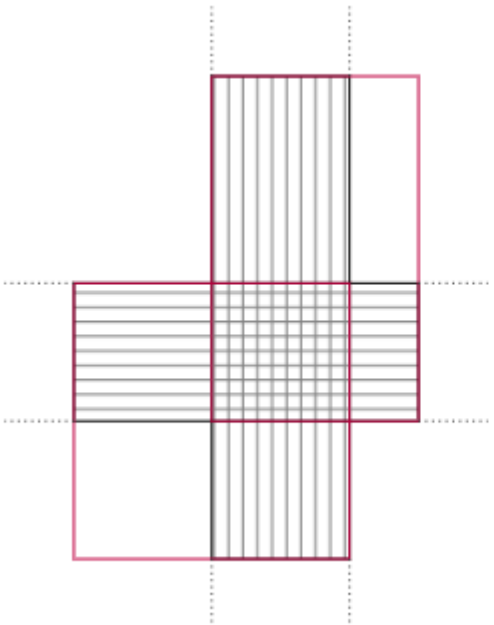


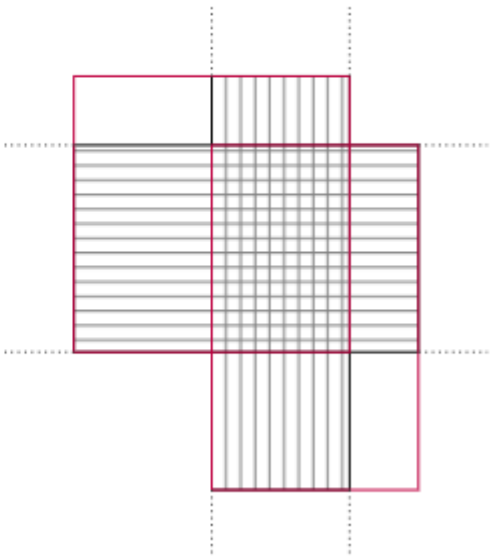
$$\mathcal{O}(N \log N + m + N w_{max})$$



$$\Omega(N w_{max} \log(N w_{max}))$$

ROHOVÉ KOLÍZIE



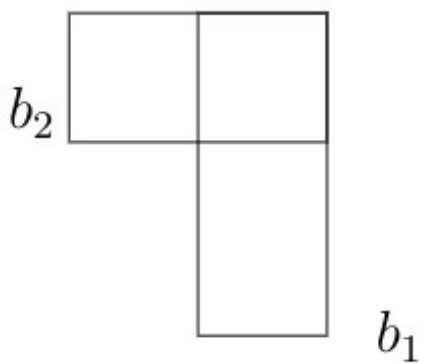


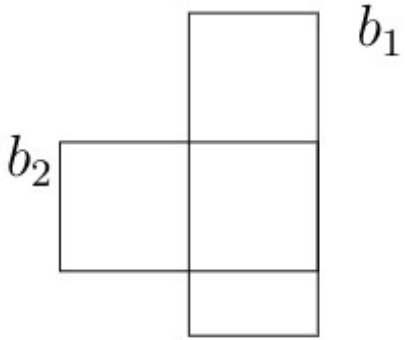
ZAROVNANÉ KOLÍZIE

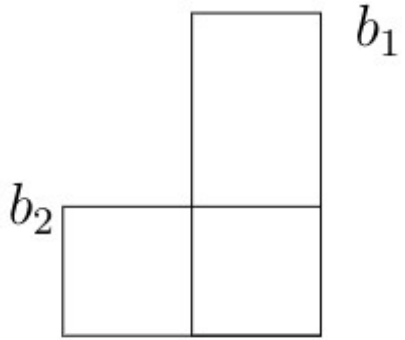
$$w_1 \cdot (h_1 - 1) \; + \; w_2 \cdot (h_2 - 1) \; - \; w_1 \cdot (h_2 - 1)$$

$$= \; w_1 \cdot (h_1 - 1) + (h_2 - 1) \cdot (w_2 - w_1)$$

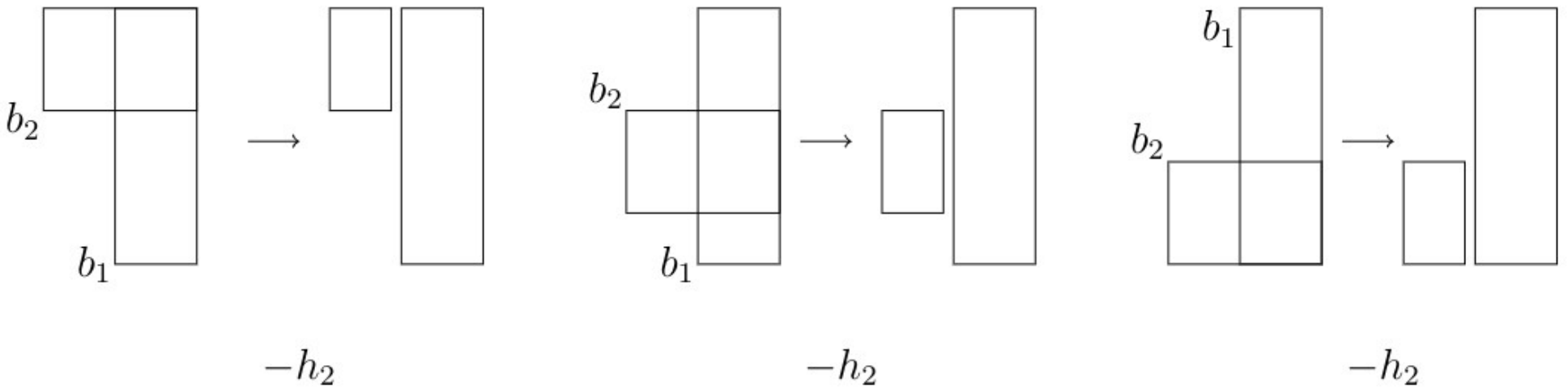
$$= \; w_1 \cdot (h_1 - h_2) + w_2 \cdot (h_2 - 1).$$



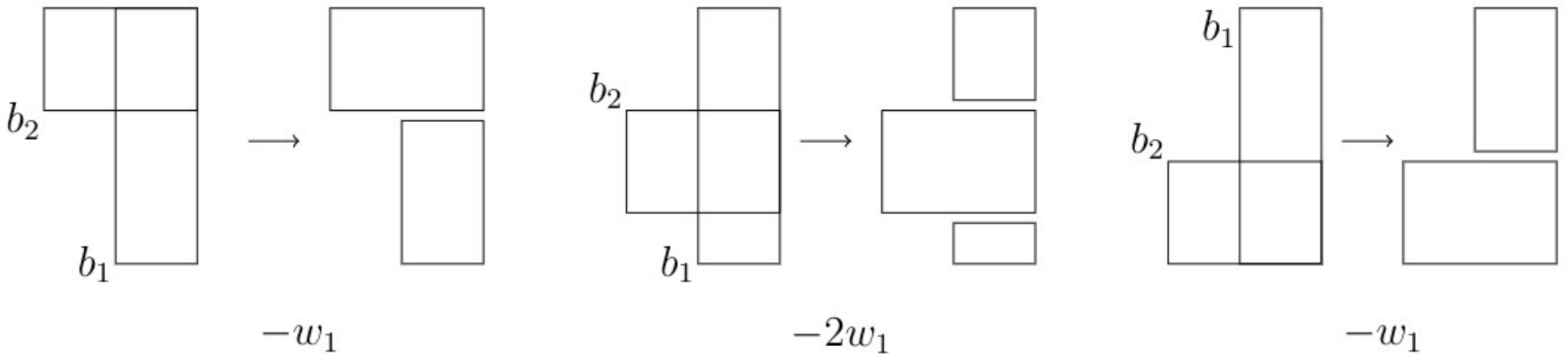




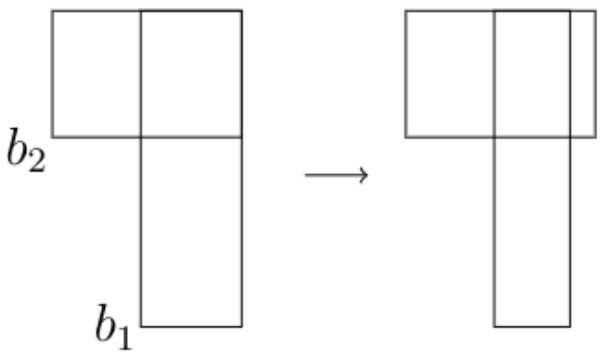
VERTIKÁLNA DIVÍZIA



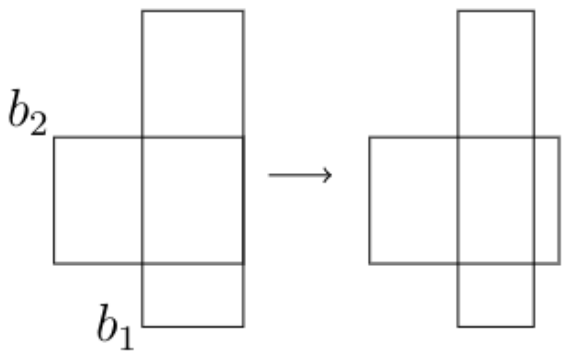
HORIZONTÁLNA DIVÍZIA



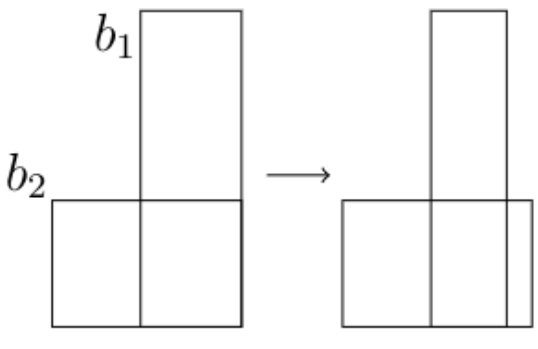
ZÚŽENIE



$-(h_1 - h_2)$



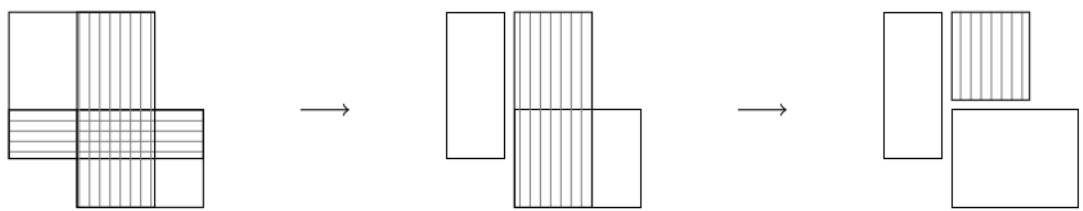
$-(h_1 - h_2)$



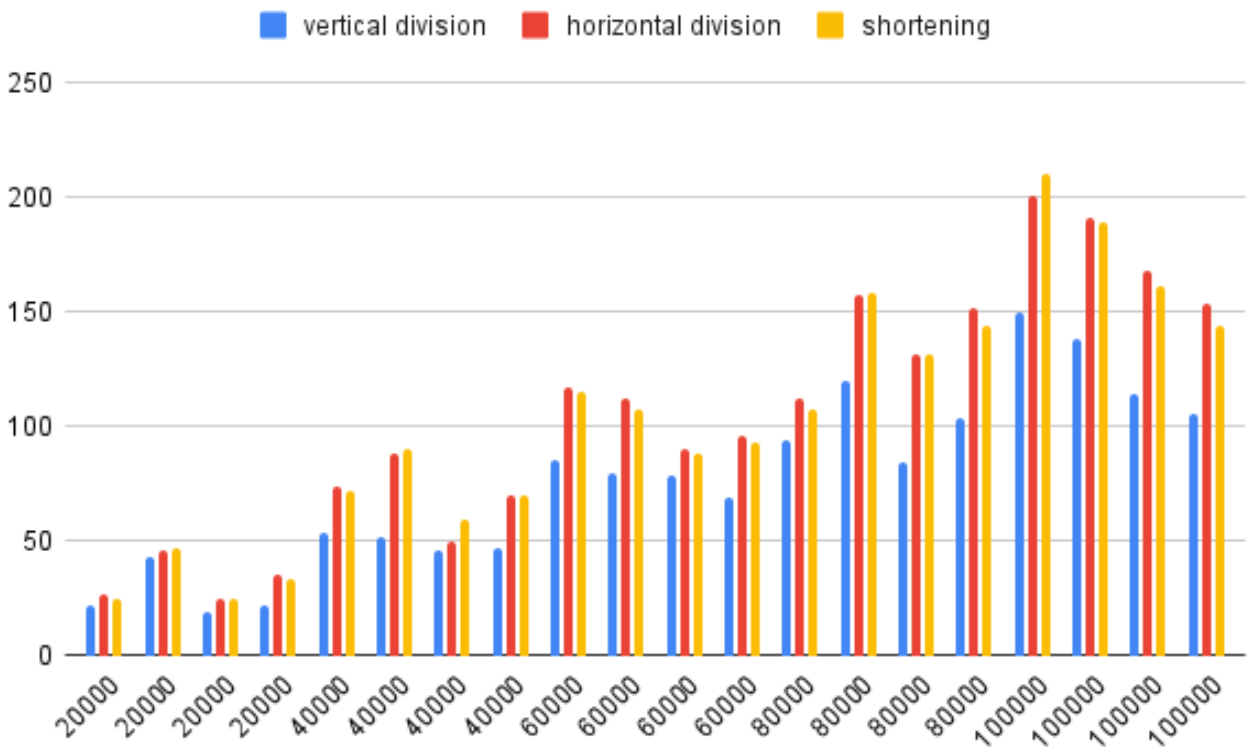
$-(h_1 - h_2)$

ROHOVÉ KOLÍZIE

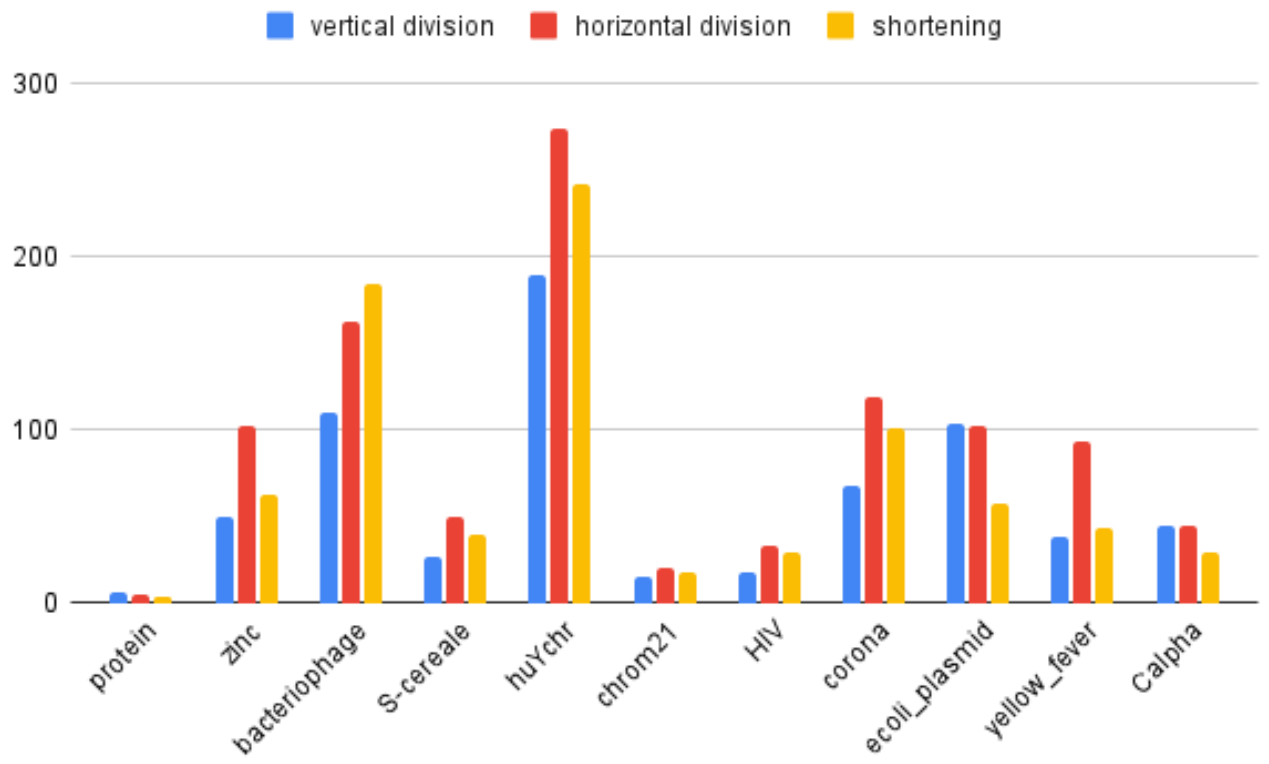
- vertikálna divízia + vertikálna divízia
- vertikálna divízia + horizontálna divízia
- horizontálna divízia + horizontálna divízia
- vertikálna divízia + zúženie
- horizontálna divízia + zúženie



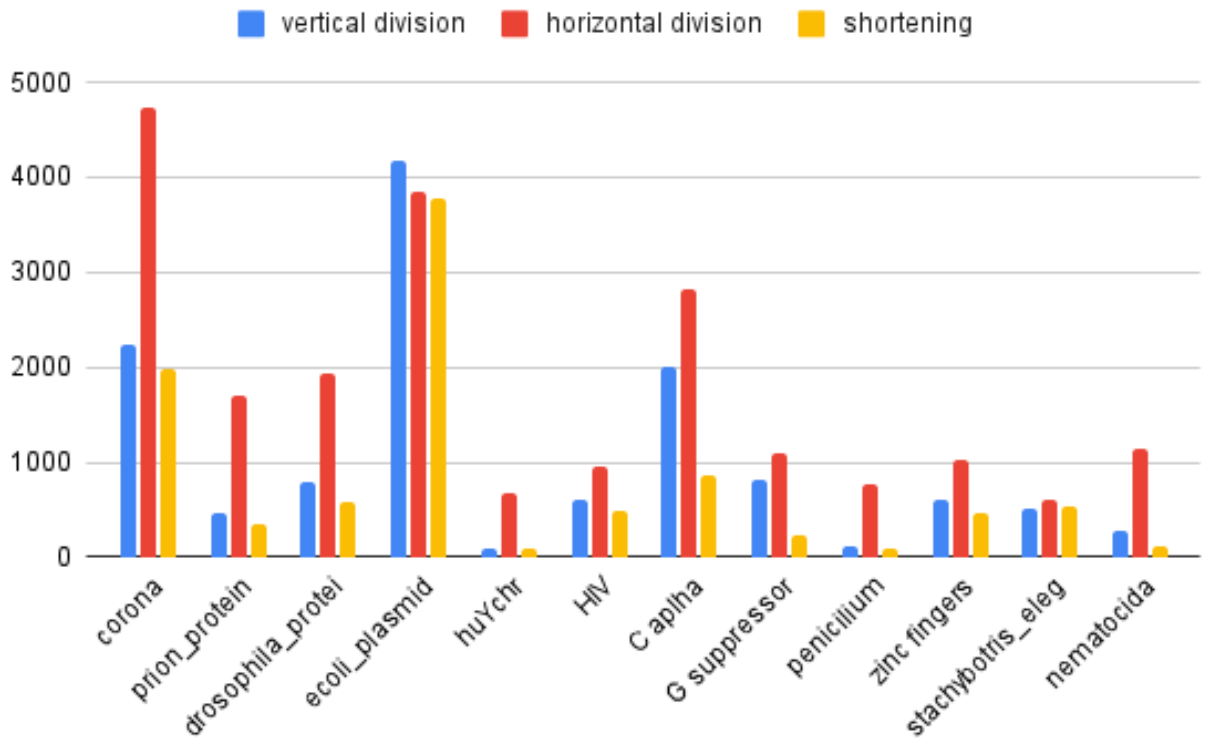
NÁHODNÉ REŤAZCE



BIOLOGICKÉ DÁTA



PANGENOMICKÉ DÁTA



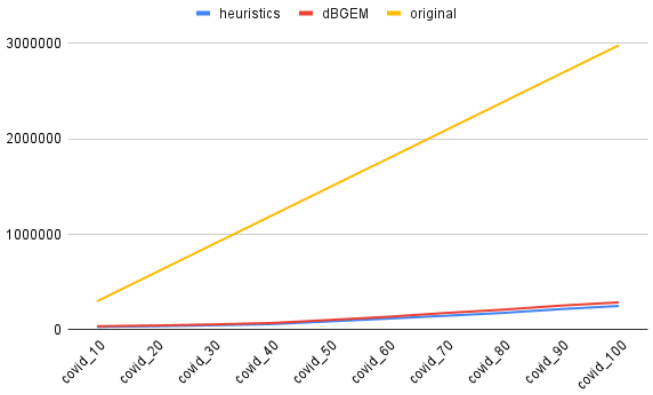
HEURISTIKA

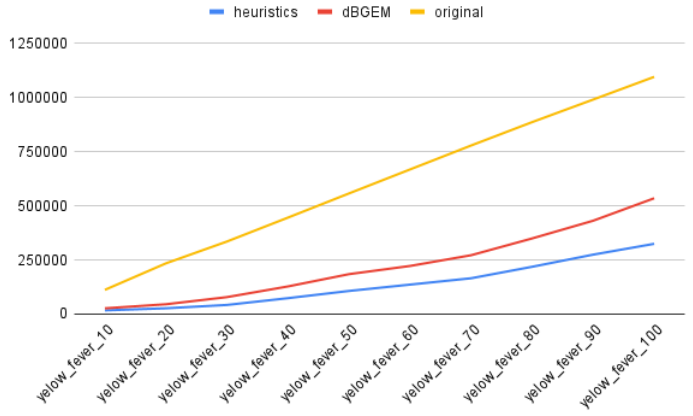
- Výpočet maximálnych blokov
- Vpravo zarované kolízie – vertikálna divízia
- Odstránenie samokolidujúcich a malých blokov
- Vľavo zarovnané kolízie – skrátenie

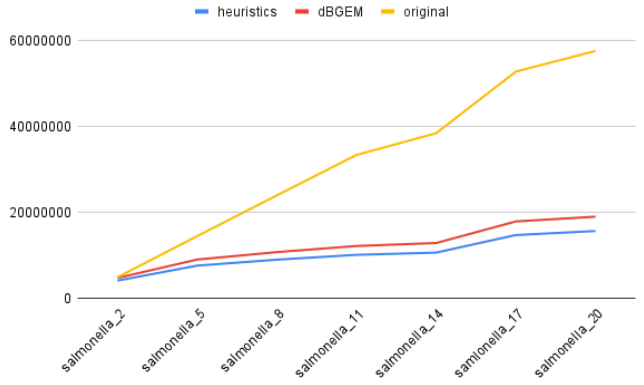
- $\mathcal{O}(N w_{max} \log N + n)$

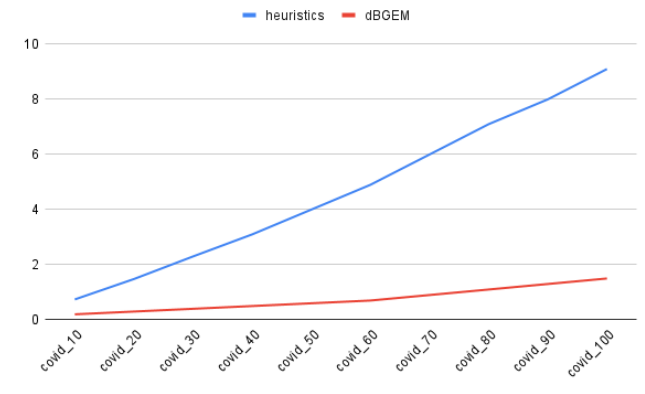
Input file	Initial size	Tunneled size	Size ratio	Time	Strategy
example.txt	18	12	0.67	0.1	dBGEM
		10	0.56	0.001	heuristics
repetitive.txt	3019	1881	0.62	0.06	dBGEM
		1306	0.43	0.005	heuristics
protein.fasta	5109	5076	0.99	0.06	dBGEM
		4753	0.99	0.02	heuristics
zinc_fingers.fa	10345	10029	0.97	0.10	dBGEM
		8724	0.84	0.03	heuristics
bacteriophage.fasta	34041	33343	0.98	0.10	dBGEM
		28965	0.85	0.10	heuristics
S-cereale.fasta	6837	6288	0.92	0.06	dBGEM
		5359	0.78	0.02	heuristics
huYchr.fasta	3693	3518	0.95	0.06	dBGEM
		2979	0.81	0.01	heuristics
yellow_fever1.fasta	11128	10833	0.97	0.08	dBGEM
		9438	0.85	0.03	heuristics
HIV1.txt	28060	9400	0.33	0.08	dBGEM
		8026	0.29	0.07	heuristics
C_alpha1.fasta	16841	16348	0.97	0.09	dBGEM
		14345	0.85	0.03	heuristics
corona_virus1.fasta	30592	29895	0.98	0.11	dBGEM
		26105	0.85	0.09	heuristics
ecoli_plasmid1.fasta	33854	32891	0.97	0.11	dBGEM
		28635	0.85	0.08	heuristics
salmonicida1.fasta	20578	19893	0.97	0.10	dBGEM
		17318	0.84	0.06	heuristics

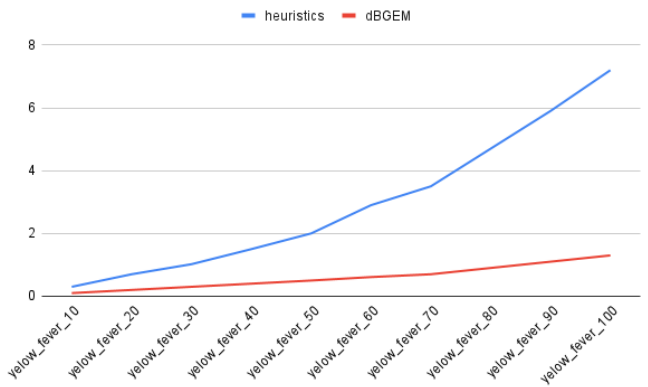
Input file	Initial size	# seq	Tunneled size	Size ratio	Time	Strategy
corona_virus.fasta	275303	9	58294	0.21	0.21	dBGEM
			40668	0.15	0.67	heuristic
prion_protein.fasta	15614	6	5881	0.38	0.10	dBGEM
			4622	0.30	0.04	heuristic
drosophila_protein.fasta	23134	9	9600	0.41	0.90	dBGEM
			7504	0.32	0.06	heuristic
ecoli_plasmid.fasta	190975	5	114653	0.60	0.30	dBGEM
			94965	0.50	0.55	heuristic
HIV.fasta	19706	7	9027	0.46	0.08	dBGEM
			6675	0.34	0.05	heuristic
Calpha.fasta	94119	7	42034	0.45	0.15	dBGEM
			35263	0.37	0.24	heuristic
G_suppressor.fasta	14048	7	5379	0.38	0.07	dBGEM
			4587	0.33	0.04	heuristic
penicilium.fasta	8640	11	3292	0.38	0.06	dBGEM
			2208	0.26	0.02	heuristic
zinc_fingers.fasta	22981	5	9305	0.40	0.08	dBGEM
			7551	0.33	0.06	heuristic
nematocida.fasta	39134	23	35735	0.91	0.13	dBGEM
			30249	0.77	0.10	heuristic
stachybotris.fasta	23348	30	10353	0.44	0.08	dBGEM
			7669	0.33	0.05	heuristic

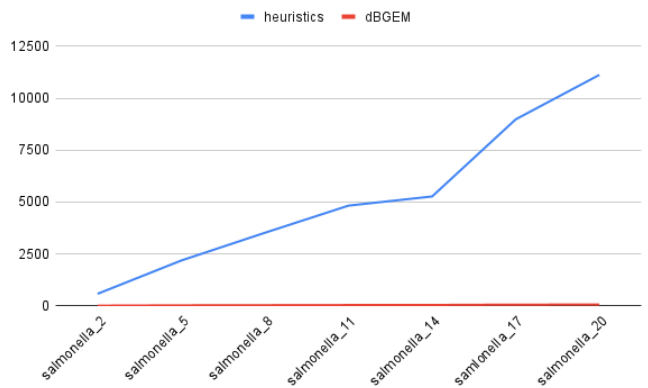

heuristics
dBGEM
original
3000000
2000000
1000000
0
covid_10
covid_20
covid_30
covid_40
covid_50
covid_60
covid_70
covid_80
covid_90
covid_100


heuristics
dBGEM
original
1250000
1000000
750000
500000
250000
0
yelow_fever_10
yelow_fever_20
yelow_fever_30
yelow_fever_40
yelow_fever_50
yelow_fever_60
yelow_fever_70
yelow_fever_80
yelow_fever_90
yelow_fever_100


heuristics
dBGEM
original
60000000
40000000
20000000
0
salmonella_2
salmonella_5
salmonella_8
salmonella_11
salmonella_14
salmonella_17
salmonella_20


heuristics
dBGEM
10
8
6
4
2
0
covid_10
covid_20
covid_30
covid_40
covid_50
covid_60
covid_70
covid_80
covid_90
covid_100

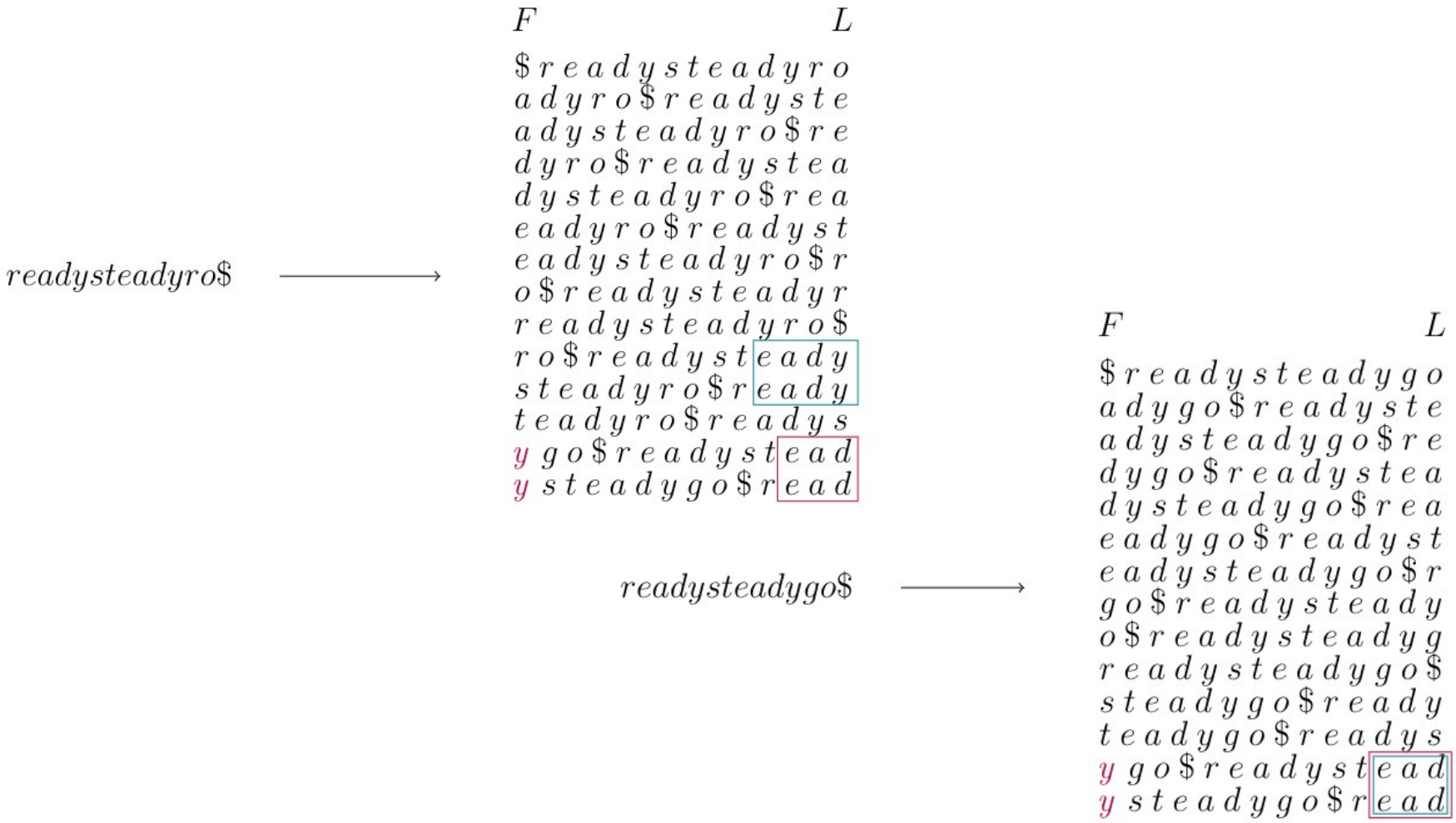

heuristics
dBGEM
8
6
4
2
0
yelow_fever_10
yelow_fever_20
yelow_fever_30
yelow_fever_40
yelow_fever_50
yelow_fever_60
yelow_fever_70
yelow_fever_80
yelow_fever_90
yelow_fever_100


heuristics
dBGEM
12500
10000
7500
5000
2500
0
salmonella_2
salmonella_5
salmonella_8
salmonella_11
salmonella_14
salmonella_17
salmonella_20

ZÁVER

- Heuristika tuneluje aj kompenzovateľné kolízie
- Veľmi vhodná pre dáta do veľkosti ~1 000 000bp
- Pre väčšie dáta je čas už v minútach až hodinách
- Čas heuristiky je blízky lineárnemu (v praxi)

ĎAKUJEM ZA POZORNOSŤ


readysteadyro$
F
L
$ r e a d y s t e a d y r o
a d y r o $ r e a d y s t e
a d y s t e a d y r o $ r e
d y r o $ r e a d y s t e a
d y s t e a d y r o $ r e a
e a d y r o $ r e a d y s t
e a d y s t e a d y r o $ r
o $ r e a d y s t e a d y r
r e a d y s t e a d y r o $
r o $ r e a d y s t e a d y
s t e a d y r o $ r e a d y
t e a d y r o $ r e a d y s
y g o $ r e a d y s t e a d
y s t e a d y g o $ r e a d
readysteadygo$
F
L
$ r e a d y s t e a d y g o
a d y g o $ r e a d y s t e
a d y s t e a d y g o $ r e
d y g o $ r e a d y s t e a
d y s t e a d y g o $ r e a
e a d y g o $ r e a d y s t
e a d y s t e a d y g o $ r
g o $ r e a d y s t e a d y
o $ r e a d y s t e a d y g
r e a d y s t e a d y g o $
s t e a d y g o $ r e a d y
t e a d y g o $ r e a d y s
y g o $ r e a d y s t e a d
y s t e a d y g o $ r e a d

Input file	Initial size	# seq	Tunneled size	Size ratio	Time	Strategy
corona_virus.fasta	275303	9	58294	0.21	0.21	dBGEM
			40668	0.15	0.67	heuristic
prion_protein.fasta	15614	6	5881	0.38	0.10	dBGEM
			4622	0.30	0.04	heuristic
drosophila_protein.fasta	23134	9	9600	0.41	0.90	dBGEM
			7504	0.32	0.06	heuristic
ecoli_plasmid.fasta	190975	5	114653	0.60	0.30	dBGEM
			94965	0.50	0.55	heuristic
HIV.fasta	19706	7	9027	0.46	0.08	dBGEM
			6675	0.34	0.05	heuristic
Calpha.fasta	94119	7	42034	0.45	0.15	dBGEM
			35263	0.37	0.24	heuristic
G_suppressor.fasta	14048	7	5379	0.38	0.07	dBGEM
			4587	0.33	0.04	heuristic
penicilium.fasta	8640	11	3292	0.38	0.06	dBGEM
			2208	0.26	0.02	heuristic
zinc_fingers.fasta	22981	5	9305	0.40	0.08	dBGEM
			7551	0.33	0.06	heuristic
nematocida.fasta	39134	23	35735	0.91	0.13	dBGEM
			30249	0.77	0.10	heuristic
stachybotris.fasta	23348	30	10353	0.44	0.08	dBGEM
			7669	0.33	0.05	heuristic

Input file	N	w_{max}
corona_virus.fasta	12015	22296
prion_protein.fasta	1456	2498
drosophila_protein.fasta	2427	1858
ecoli_plasmid.fasta	27385	33837
HIV.fasta	2044	2739
Calpha.fasta	10877	3285
G_suppressor.fasta	1502	1075
penicilium.fasta	603	750
zinc_fingers.fasta	2247	4334
nematocida.fasta	8750	846
stachybotris.fasta	2160	896

Input file	Initial size	∼Opt	Tunneled size	Size ratio	Time	Strategy
example.txt	18	10	10	0.56	0.10	ILPR
			12	0.67	0.10	dBGEM
			10	0.56	0.001	heuristic
repetitive.txt	3019		622	0.21	880	ILPR
			1881	0.62	0.06	dBGEM
			1306	0.43	0.01	heuristic
protein.fasta	5109	5019	5019	0.98	0.36	ILPR
			5076	0.99	0.06	dBGEM
			4753	0.93	0.02	heuristic
zinc_fingers.fa	10345	9031	9050	0.87	1111	ILPR
			10029	0.97	0.10	dBGEM
			8724	0.84	0.03	heuristic
bacteriophage.fasta	34041	29796	29796	0.88	7.5 hours	ILPR
			33343	0.98	0.10	dBGEM
			28965	0.85	0.10	heuristic
S-cereale.fasta	6837	5567	5569	0.81	730	ILPR
			6288	0.92	0.06	dBGEM
			5359	0.78	0.02	heuristic
huYchr.fasta	3693	1678	2370	0.64	520	ILPR
			3518	0.95	0.06	dBGEM
			2979	0.81	0.01	heuristic
chrom21_rep.fasta	20001	2166	4431	0.22	15 hours	ILPR
			5022	0.25	0.06	dBGEM
			4294	0.21	0.04	heuristic