

IDENTIFIKÁCIA A ANALÝZA REPETITÍVNYCH SEKVENCIÍ V NANOPÓROVÝCH DÁTACH

SAMUEL MOLČAN

ŠKOLITEĽ: DOC. MGR. TOMÁŠ VINAŘ, PHD.

26.6.2019

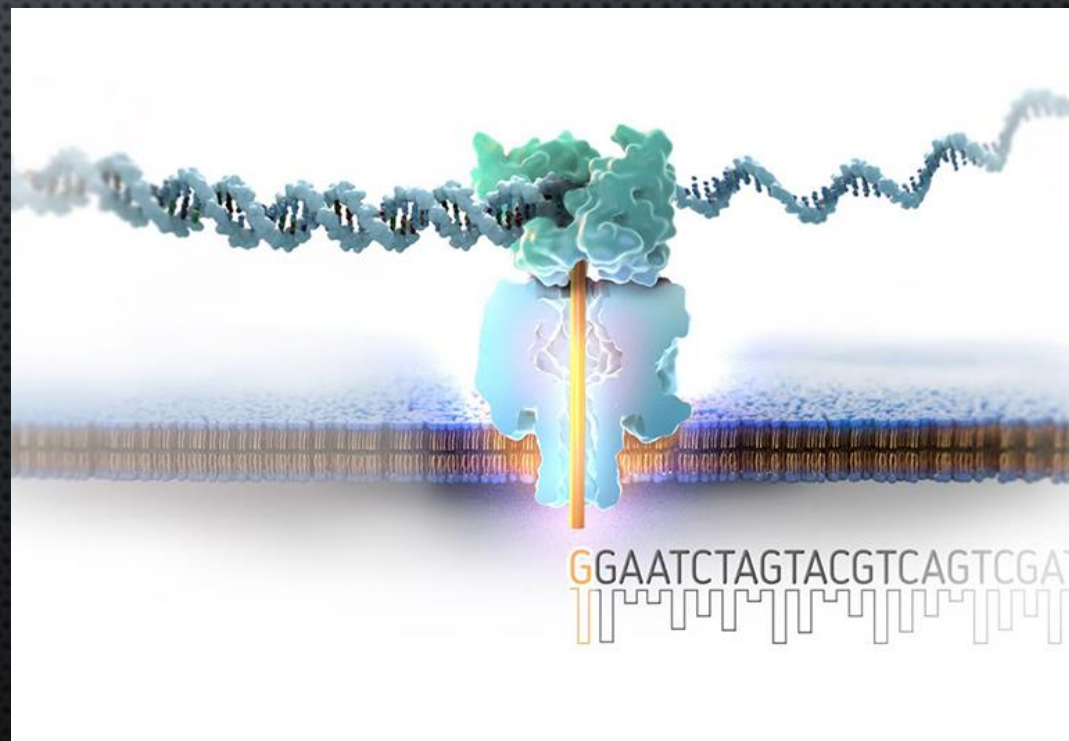
GENETICKÁ INFORMÁCIA - DNA

- REPETITÍVNE SEKVENCIE - TANDEMOVÉ OPAKOVANIA
- SEKVENOVANIE DNA – MINION
- PREKLADAČ BÁZ

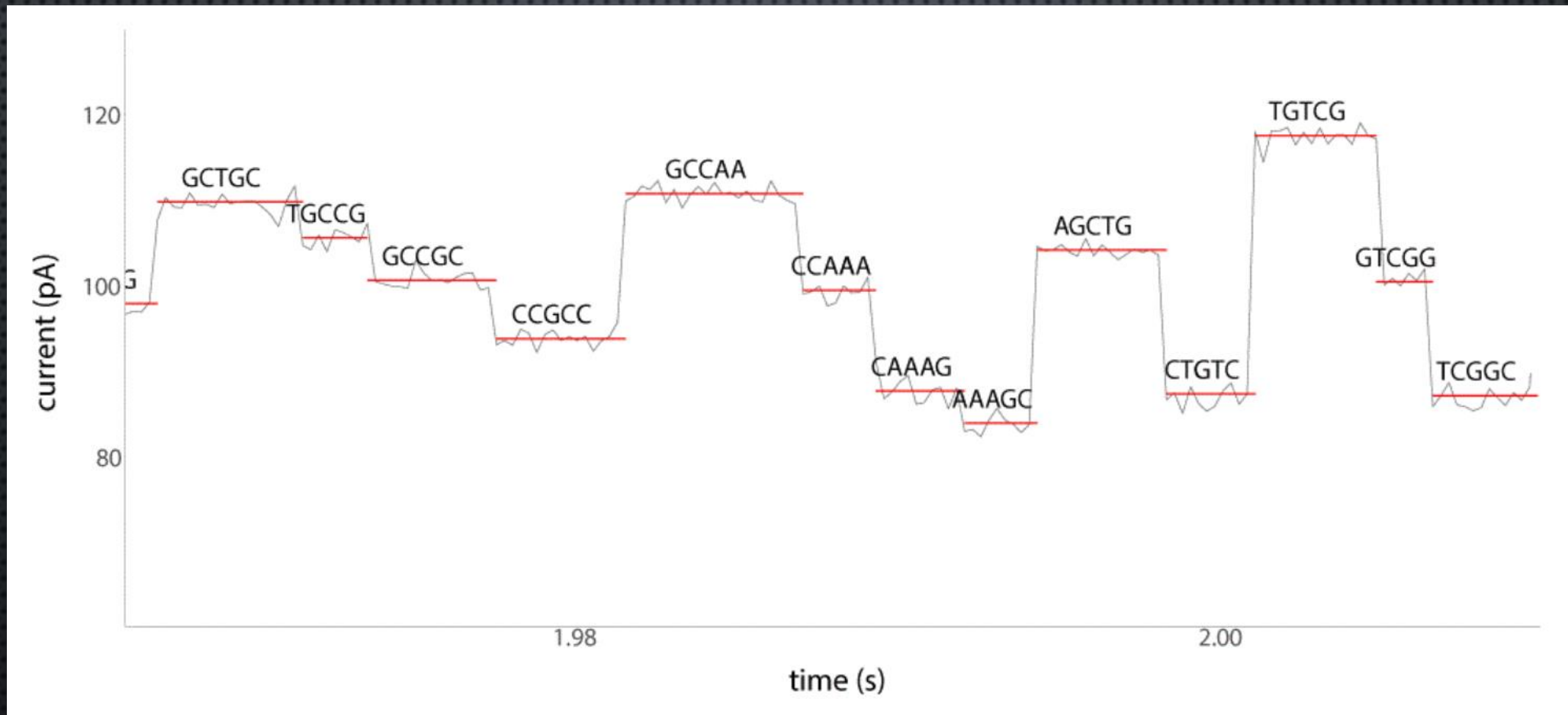
TANDEMOVÉ OPAKOVANIA



MINION SEKVENÁTOR



VÝSTUPNÝ SIGNÁL SEKVENÁTORA MINION



ŠTANDARDNÝ SPÔSOB HĽADANIA TANDEMOVÝCH OPAKOVANÍ

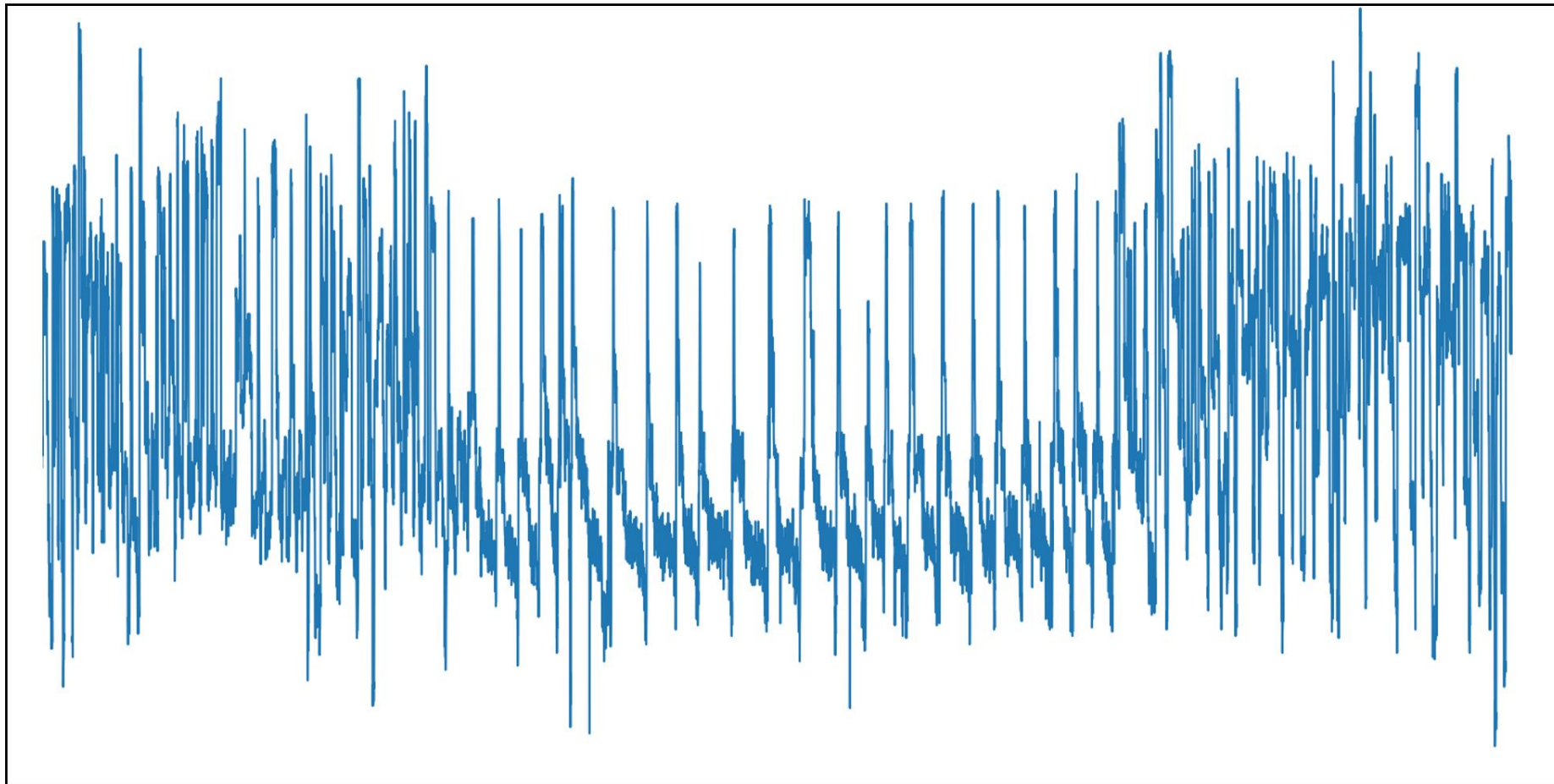
- TANDEM REPEAT FINDER, TANTAN
- PREKLAD BÁZ, VIACNÁSOBNÉ ZAROVNANIE, KONSENZUS - VÝPOČTOVO NÁROČNÉ
- HĽADANIE LEN V REŤAZCI PRELOŽENÝCH BÁZ - NEPRESNÉ

NOVÝ PRÍSTUP HĽADANIA TANDEMOVÝCH OPAKOVANÍ

- OPAKUJÚCE SA POSTUPNOSTI V SUROVOM SIGNÁLE
KOREŠPONDUJÚ S TANDEMOVÝMI OPAKOVANIAMÍ

REPETITÍVNY SIGNÁL

napätie

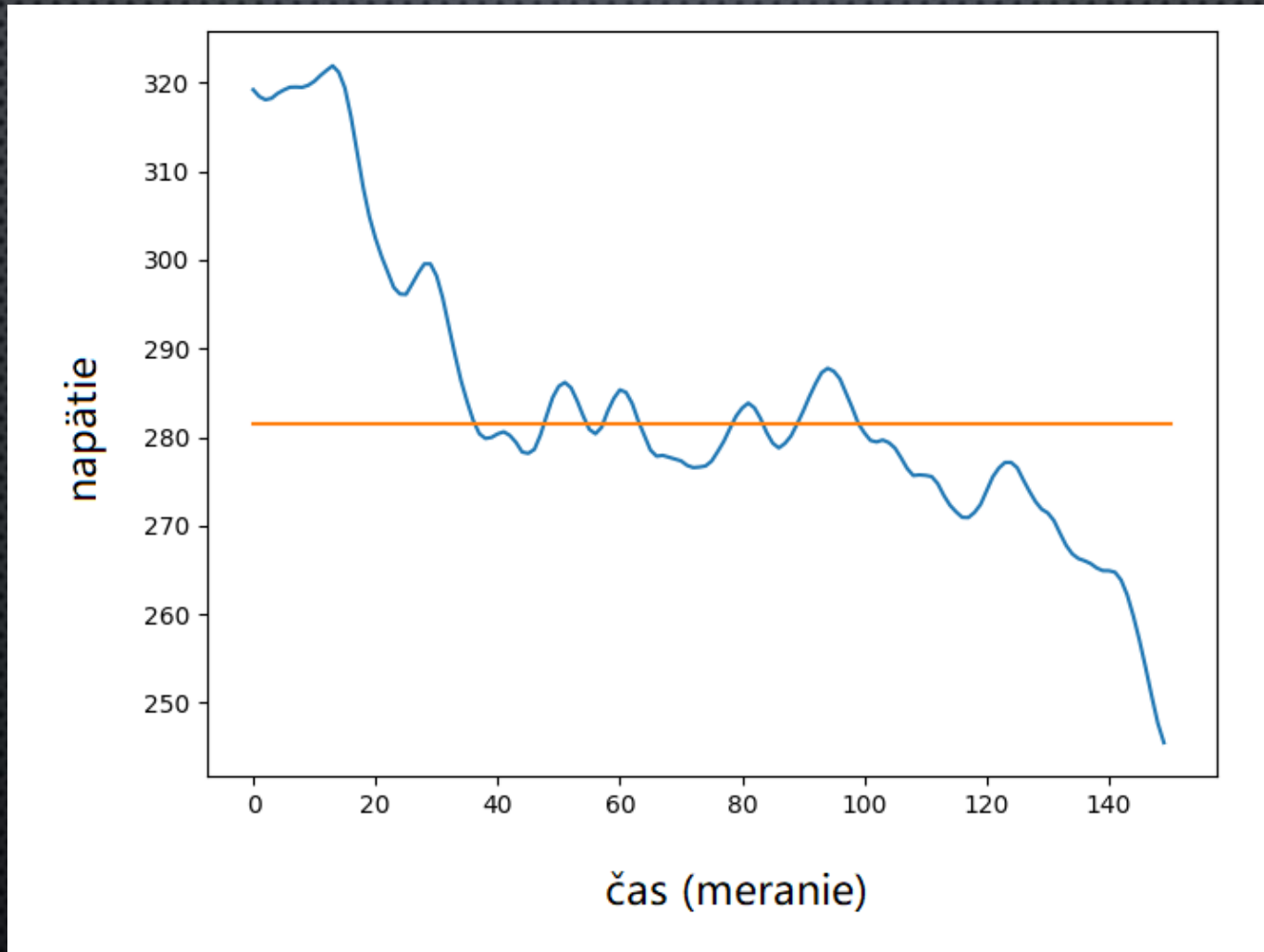


čas (meranie)

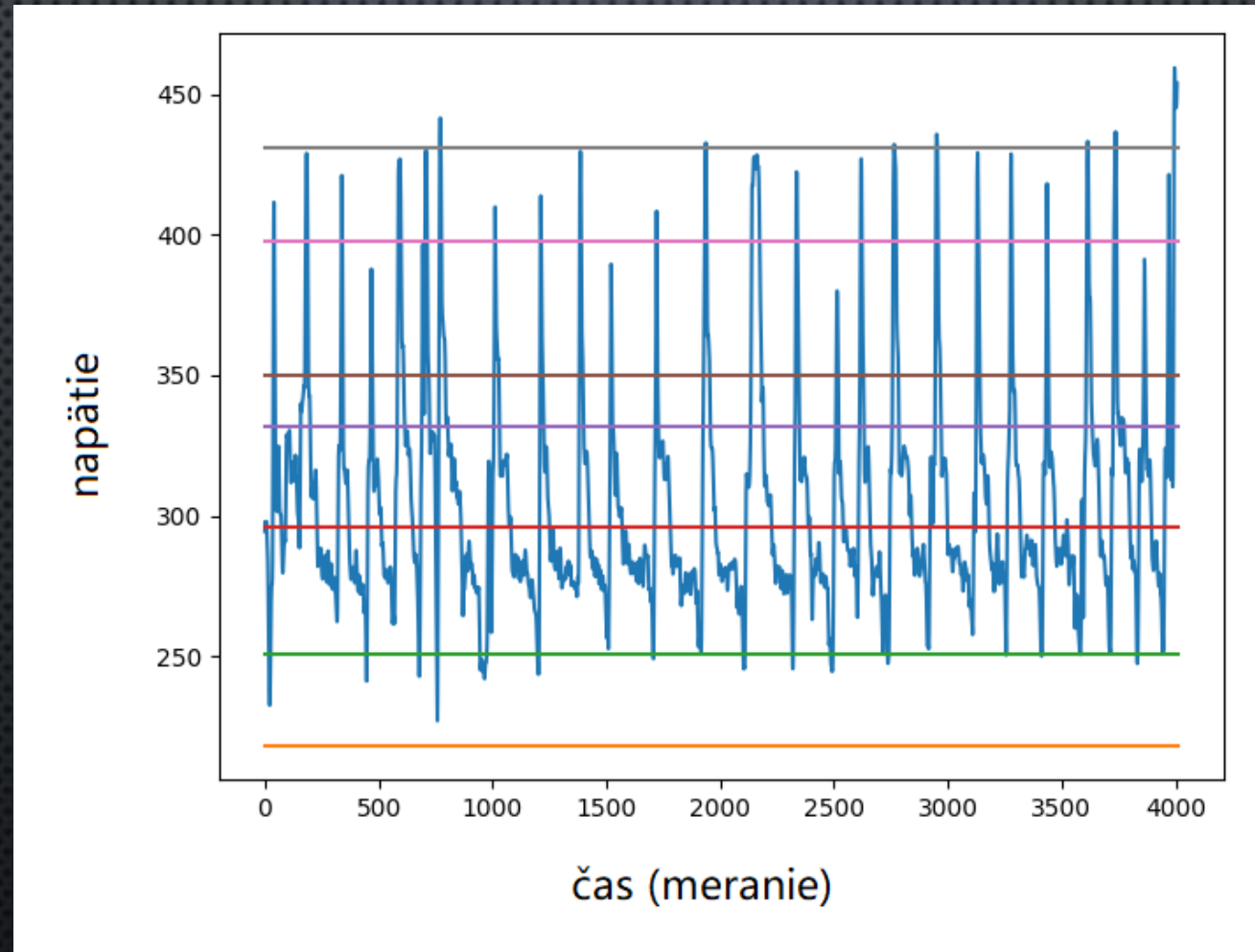
HĽADANIE TANDEMOVÝCH OPAKOVANÍ V SIGNÁLE - DISKRETIZÁCIA SIGNÁLU

- ÚPRAVA SIGNÁLU (VYROVNANIE, ODSTRÁNENIE EXTRÉMNYCH POZOROVANÍ)
- VYTVORENIE FUNKČNÉHO POČTU KATEGÓRIÍ
- DYNAMICKÉ HĽADANIE HRANÍC - PROBLÉM OSCILÁCIE
- KOMPRESIA REŤAZCA
- HODNOTIACE KOMPONENTY – SKÓROVACIA MATICA, INZERCIE, DELÉCIE

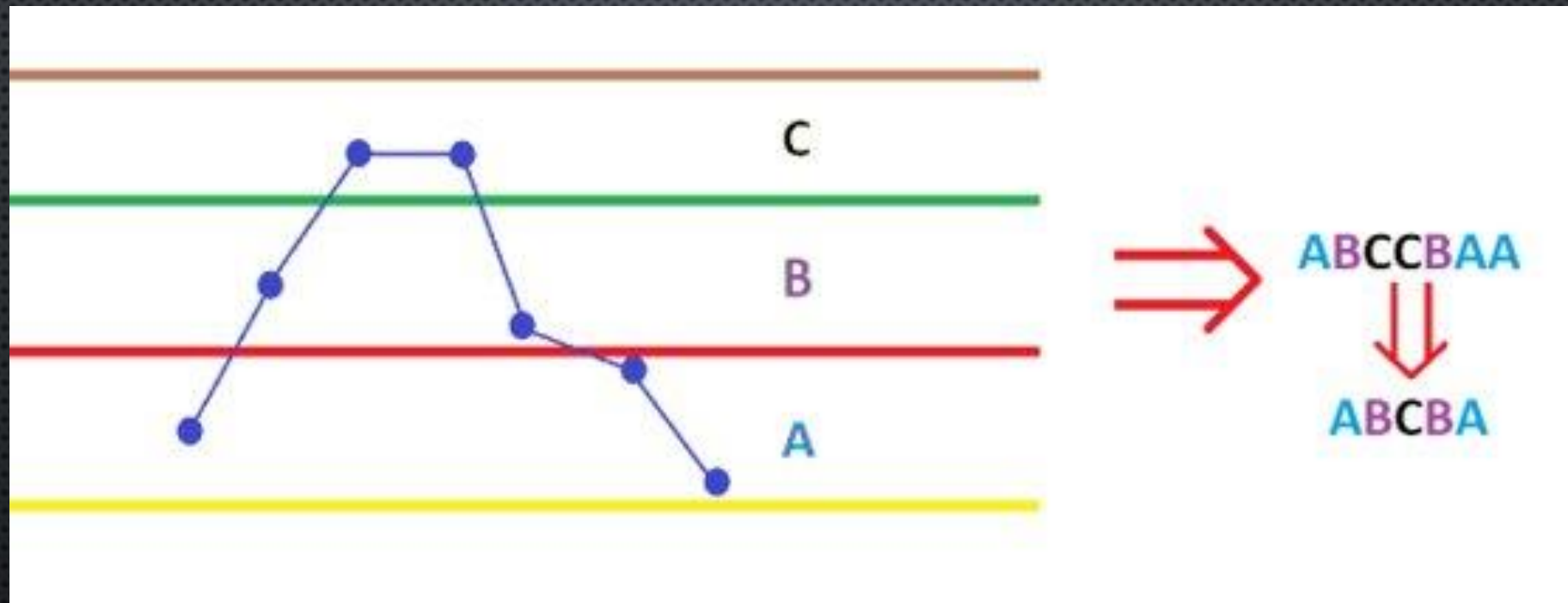
OSCILÁCIA



DYNAMICKÉ ROZDELENIE HRANÍC



KOMPRESIA REŤAZCA



SKÓROVACIA MATICA

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	J	Z	X	*	
A	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	
R	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
N	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
D	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
C	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
Q	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
E	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
G	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
H	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
I	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
L	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
K	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
M	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
F	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
P	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5	-5
S	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5	-5
T	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5	-5
W	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5	-5
Y	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5	-5
V	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5	-5
B	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5	-5
J	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5	-5
Z	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1	-5
X	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3	1
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	3	5	3

TESTOVANIE A VÝSLEDKY

- SENZITIVITA A SPECIFICITA, ZHODA POČTU OPAKOVANÍ
- TESTOVANIE 31 104 KOMBINÁCIÍ PARAMETROV NA TRÉNOVACEJ MNOŽINE
- TESTOVACIA MNOŽINA

M



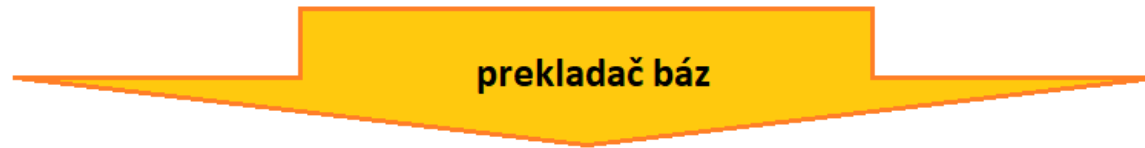
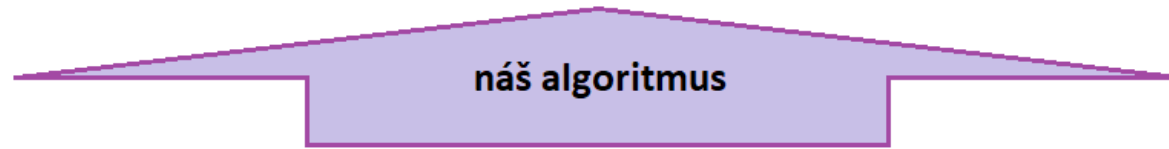
$M \cap R$



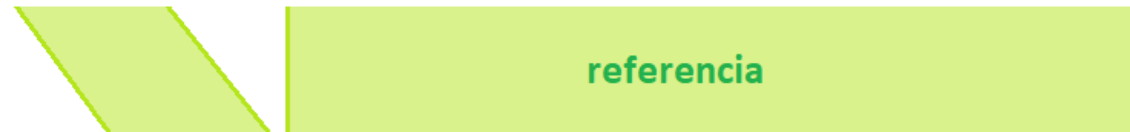
R



GGATCAGATCTAAGTTAAGTTAAGTTCCTAGTAA



AATCGTATCGATTACCTTACCTTACCGGCTACTGG



AATCTCGATTACCTTACCTTACCGGCTA

Úplná zhoda	1. sused	2. sused	Nezhoda	Začiatok medzery	Rozšírenie medzery	Počet kategórií	Maximálna dĺžka vzoru	Senzitivita (%)	Specificita (%)	Zhoda iterácií (%)
5	4	1	-7	6	5	7	14	3.33	4.69	50.73
7	4	1	-8	6	5	8	8	7.97	2.14	39.19
7	3	2	-9	6	4	10	9	7.71	3.17	36.54
6	3	2	-8	6	5	7	10	21.05	3.34	41.76
7	3	2	-7	6	5	10	18	25.39	2.61	50.26
6	4	3	-9	5	4	7	20	31.38	4.41	33.22
5	3	2	-7	5	4	9	24	33.74	2.45	57.84
6	2	1	-6	6	4	7	22	48.78	3.40	38.04
7	4	3	-7	5	3	7	15	26.05	14.11	50.52
7	4	3	-6	5	3	7	11	26.28	14,73	45.30
priemerné hodnoty								18.6	3.7	39.8
maximálne hodnoty								59.0	14.7	60.9
detekcia v preloženom reťazci								10.55	15.94	54.88

metóda	senzitivita	specificita	zhoda počtu opakovaní
hľadanie naším algoritmom v signále	12.16%	34.23%	24.38%
štandardný algoritmus v preloženom reťazci	6,25%	19,75%	44.08%

Tabuľka 4.3: Výsledky finálneho testovania

ROZŠÍRENIE

- POMOCOU STROJOVÉHO UČENIA
- NÁJŠŤ TANDEMOVÉ OPAKOVANIA AJ V NEREPETITÍVNYM SIGNÁLE
- ZVÝŠIŤ SPECIFICITU – NA OHLÁSENÉ ČASTI SPUSTIŤ DODATOČNÚ KONTROLU

ĎAKUJEM ZA POZORNOST