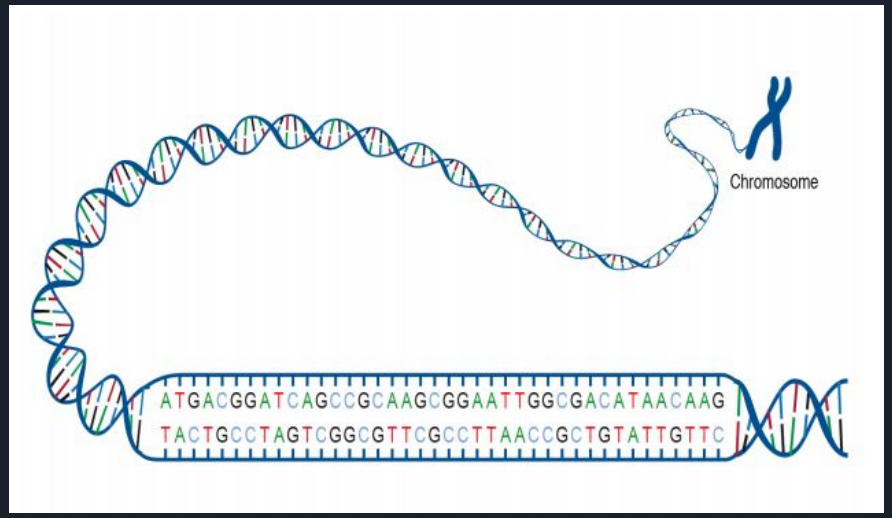
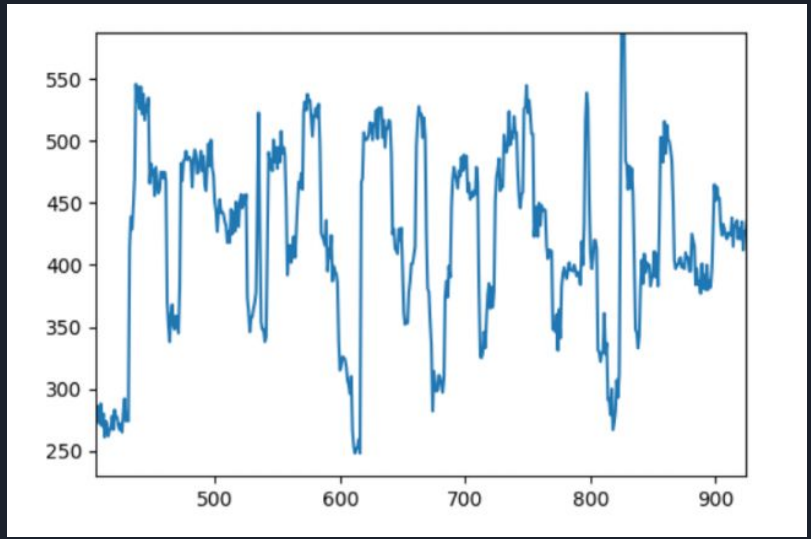
Dátové štruktúry pre Selektívne sekvenovanie

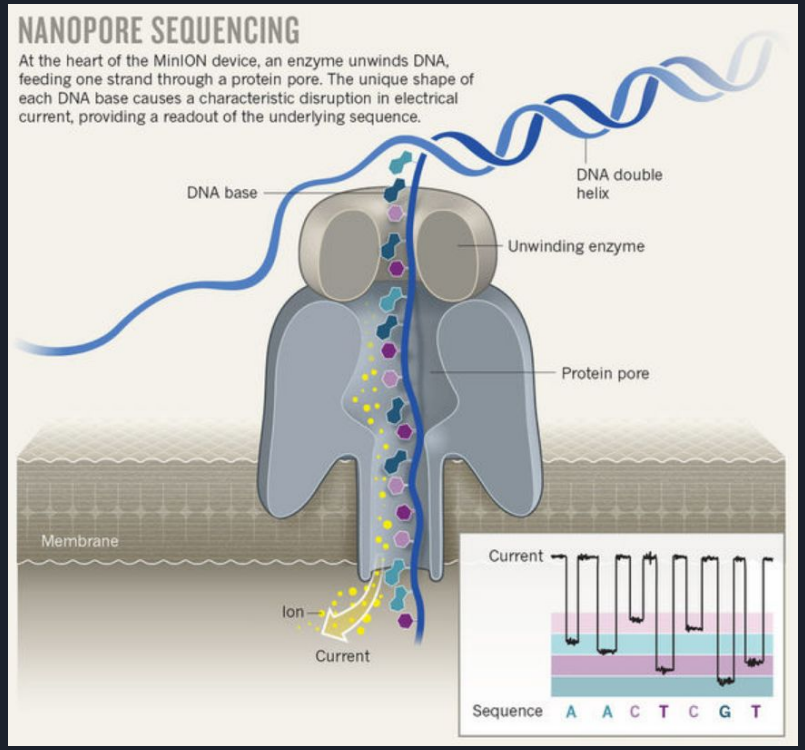
Meno a priezvisko: Andrej Korman
Školiteľ: doc. Mgr. Tomáš Vinař, PhD.

Úvod do problematiky





Úvod do problematiky



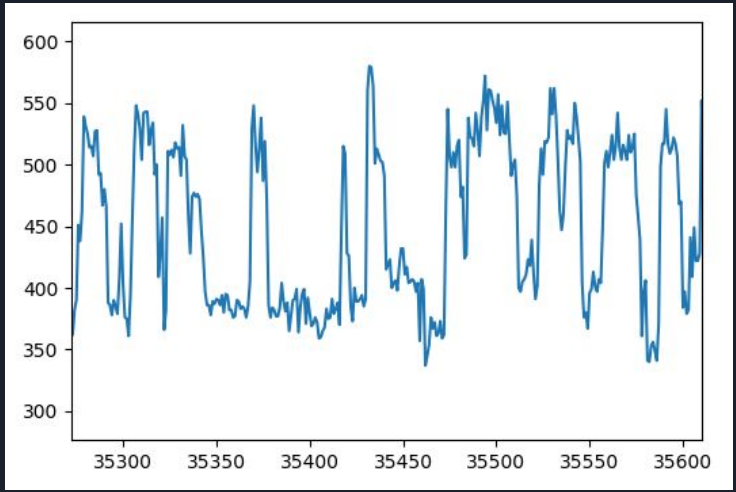
kmer	mean	sd
AAAAAA	-0.53125	0.33288004
AAAAAC	-0.8561053	0.33288004
AAAAAG	-0.6227140	0.33288004
AAAAAT	-0.8553921	0.33288004
AAAACA	-1.3231160	0.33288004
AAAACC	-1.4454362	0.33288004
AAAACG	-1.3438284	0.33288004
AAAACT	-1.3678160	0.33288004
AAAAGA	-0.7254464	0.33288004
AAAAGC	-0.8928571	0.33288004
AAAAGG	-0.8106306	0.33288004
AAAAGT	-1.0230346	0.33288004
AAAATA	-1.3443526	0.33288004
AAAATC	-1.6689948	0.33288004

Úvod do problematiky

referencia

ACGCTGACTG...ACTCACTCTCCG

surový signál
hľadaný v referencii

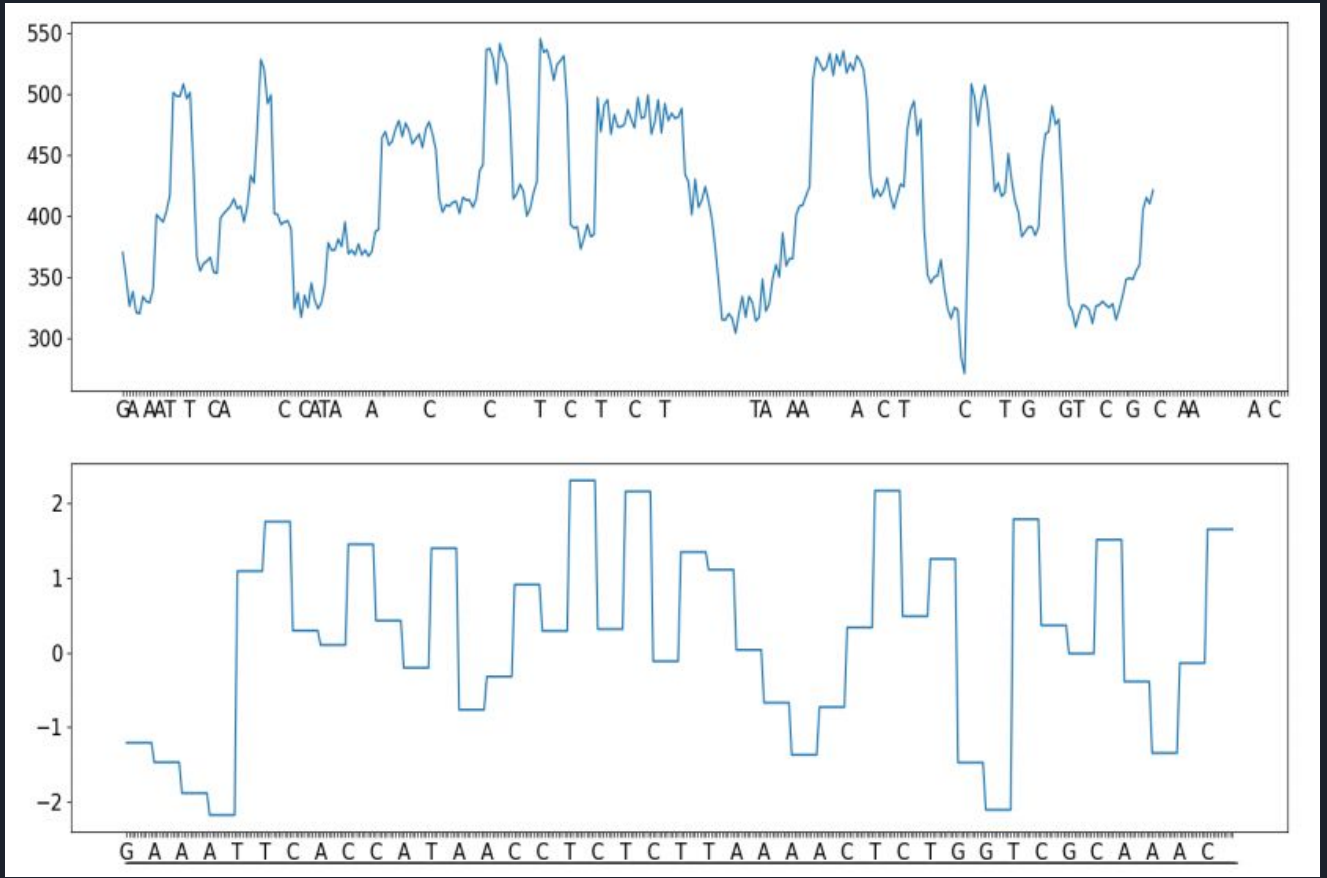


Navrhované riešenie

1. Nasimulovať umelý signál z referencie
2. Vytvoriť alternatívnu reprezentáciu tohto signálu
3. Vytvoriť index, ktorý nám umožní rýchlo odpovedať na otázku, či sa signál nachádza v referencii

- alternatíva fungujúceho pomalšieho riešenia [Loose et al., 2016]

Simulácia signálu

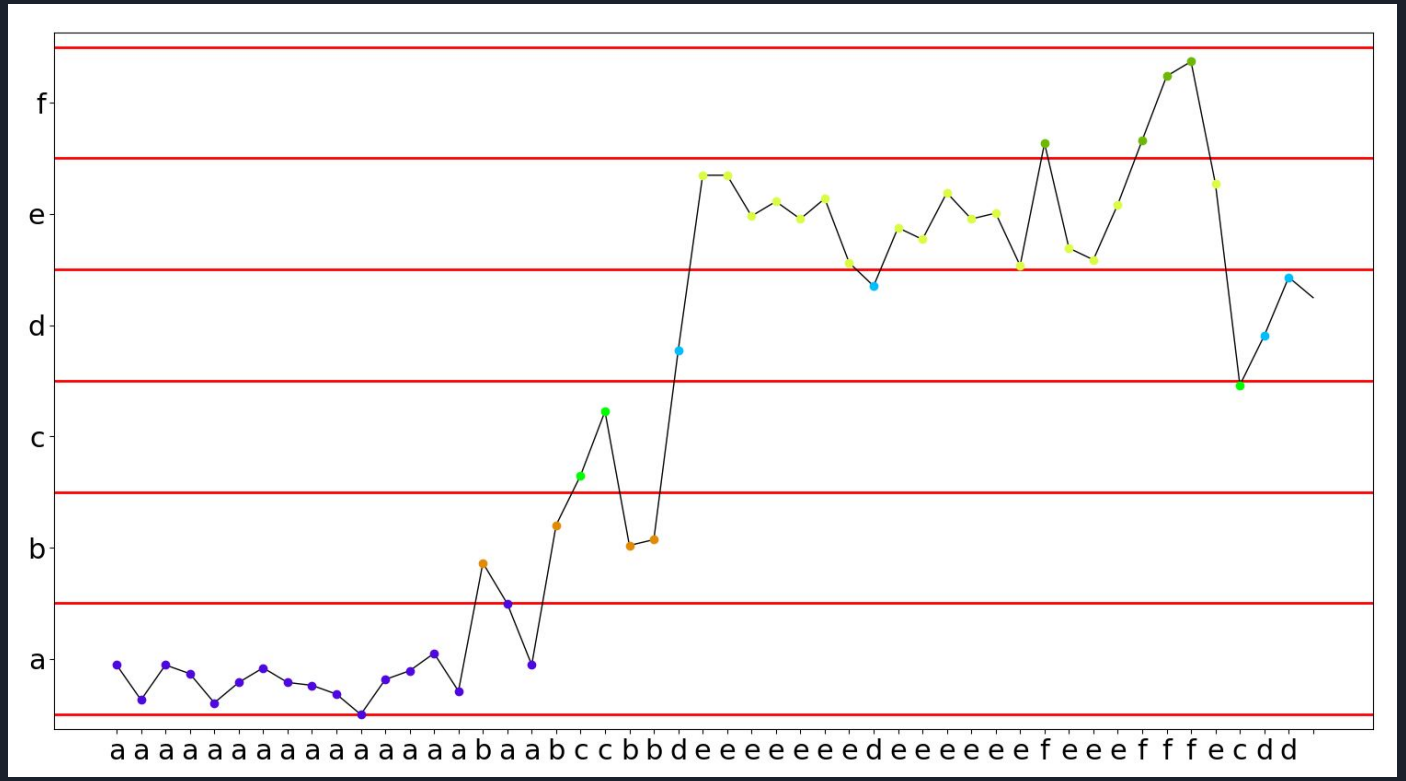


Diskretizácia signálu

- snaha o lepšiu reprezentáciu signálu

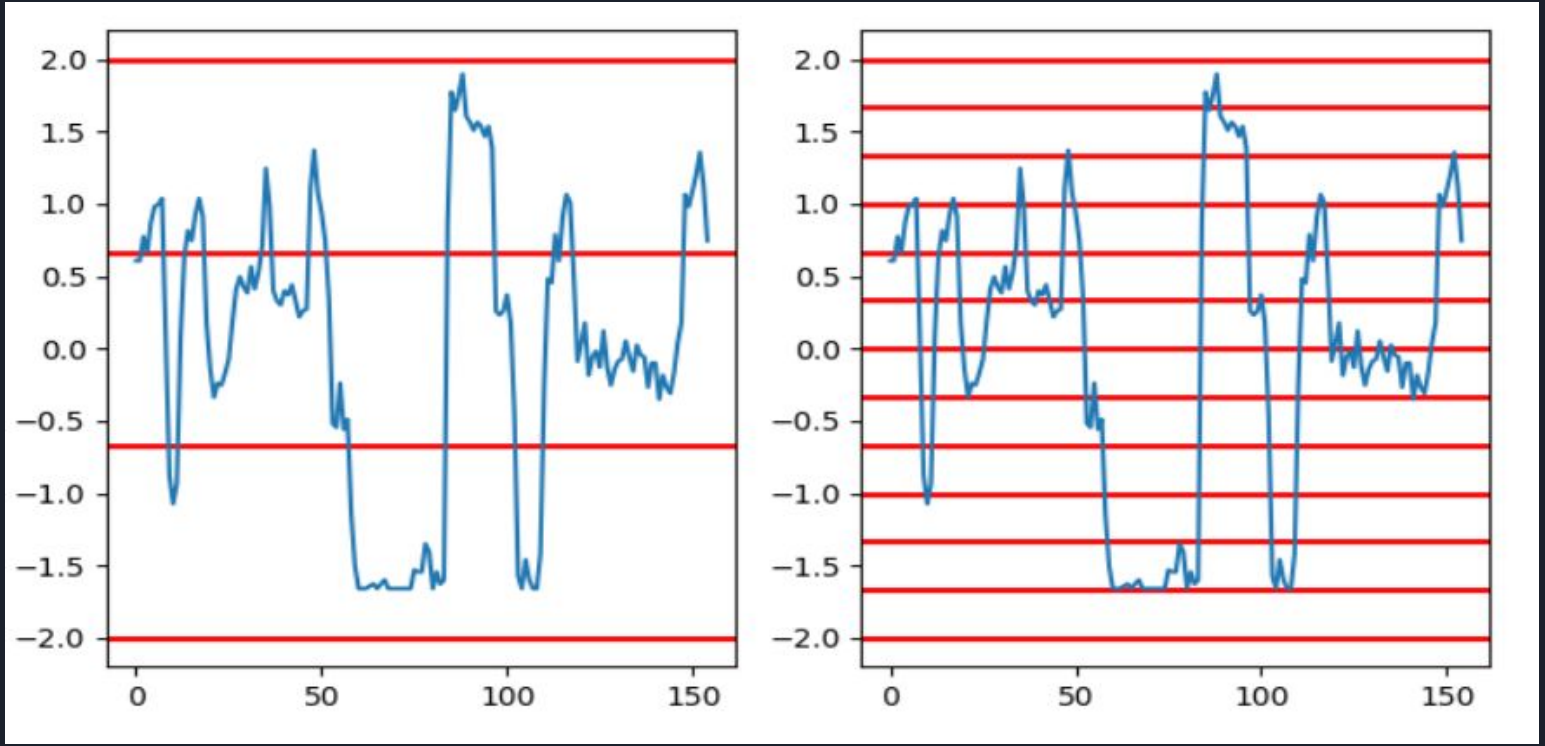
- umožnenie vyhľadávania v signále

Diskretizácia signálu

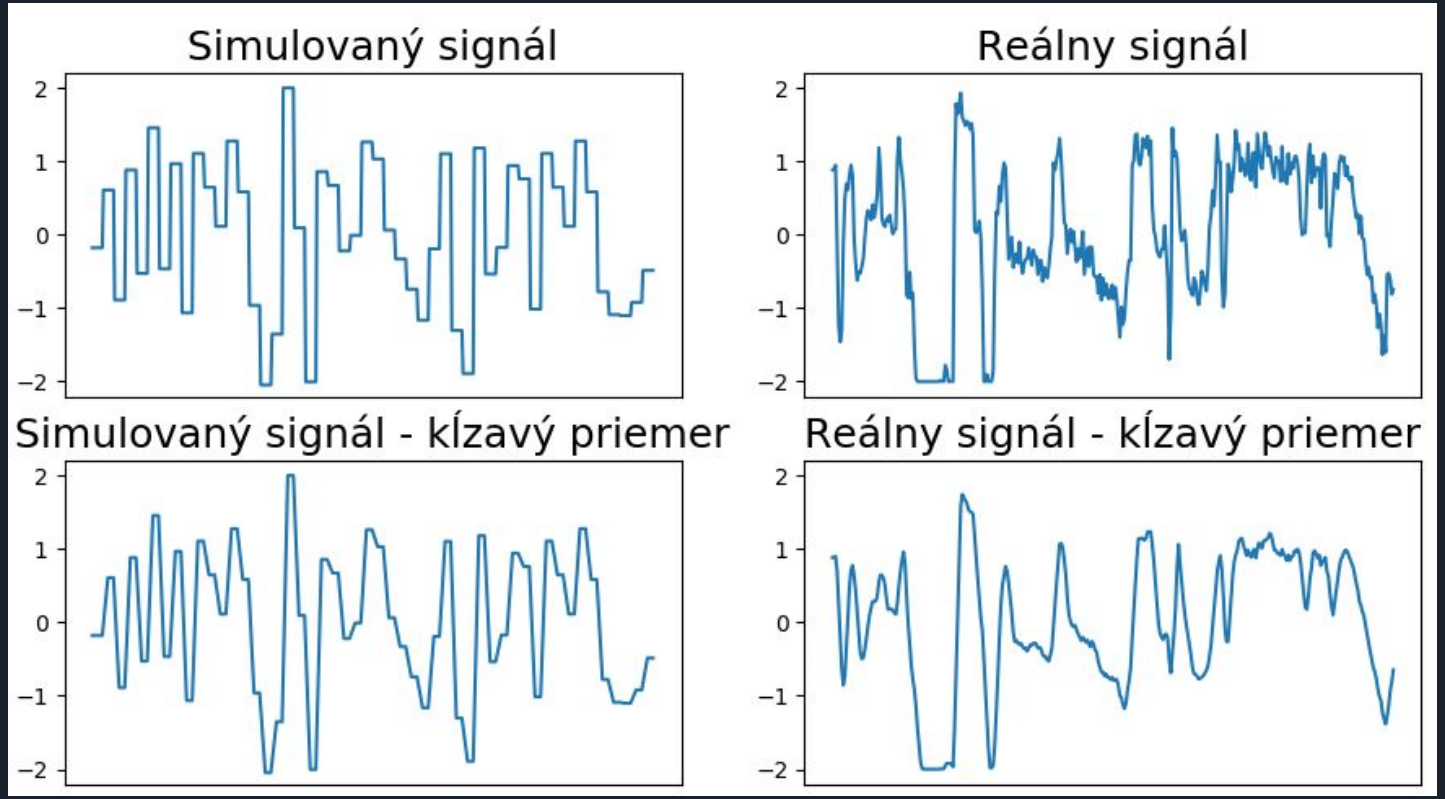


a b a b c b d e d e f e f e c d

Diskretizácia signálu



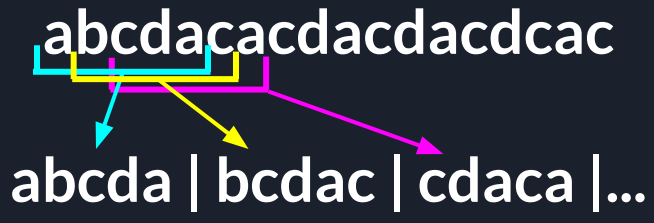
Simulovaný a reálny signál



Rozdelenie na k-tice

l = 4 (počet úrovní)

k = 5 (dĺžka k-tic)



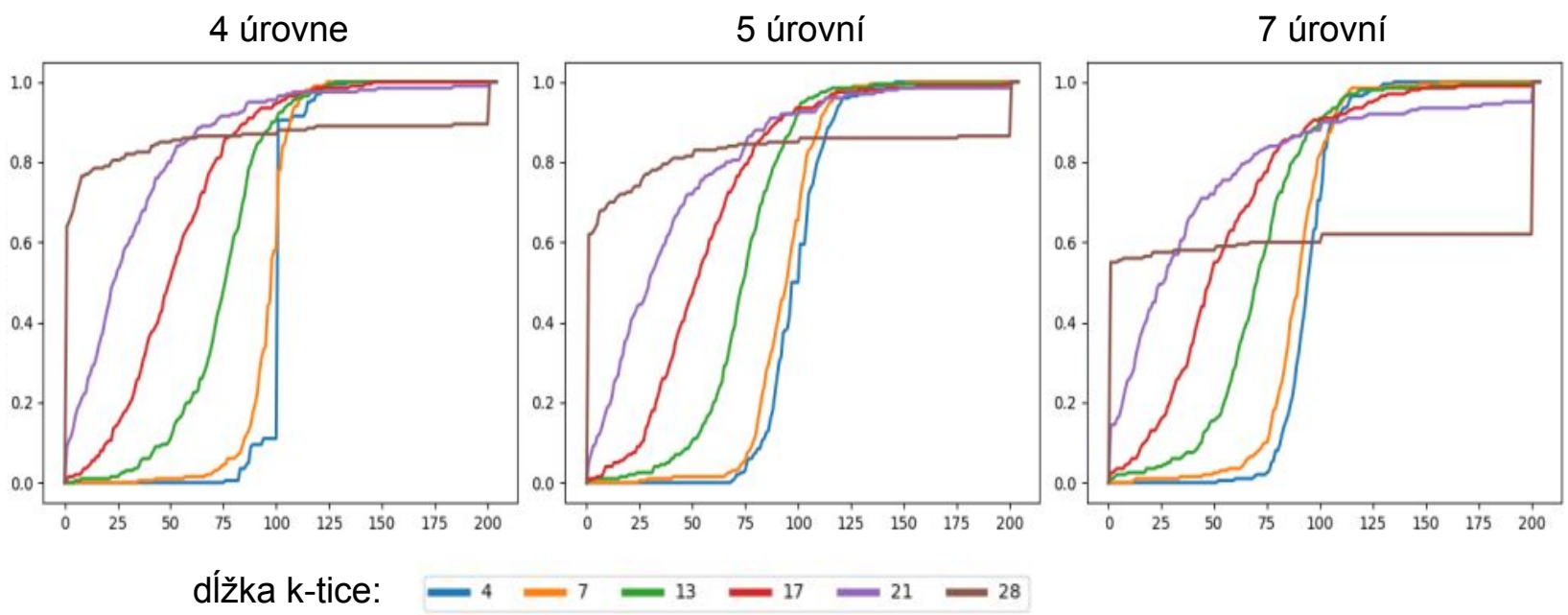
Zdieľané k-tice

Umelý signál	abcdacacda	{abcd, bcda, cdac, daca, acac, cacd, acda}
Reálny signál	cabcdabac	{cabc, abcd, bcda, cdab, daba, abac}
Náhodný signál	abcdbdad	{acbc, cbcb, bcbd, cbda, bdad}

dobré zdieľané k-tice (2)

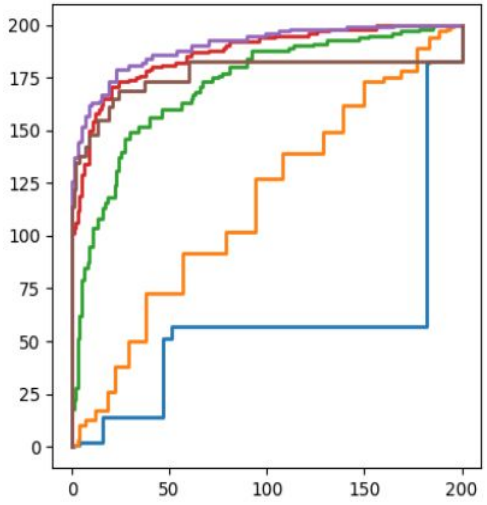
zlé zdieľané k-tice (1)

Výsledky - Graf 1

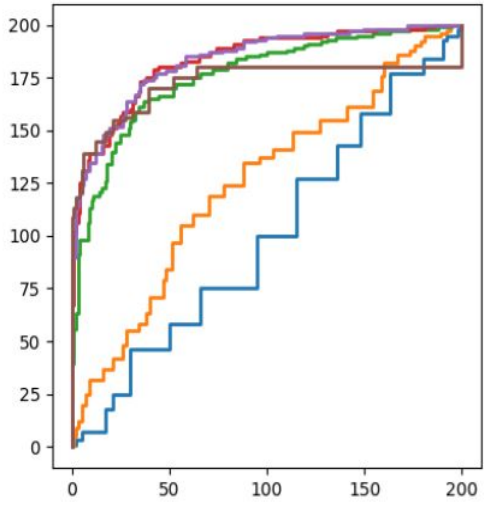


Výsledky - Graf 2

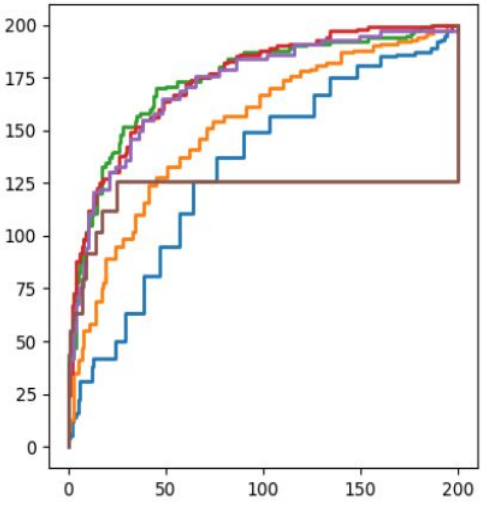
4 úrovne



5 úrovní



7 úrovní



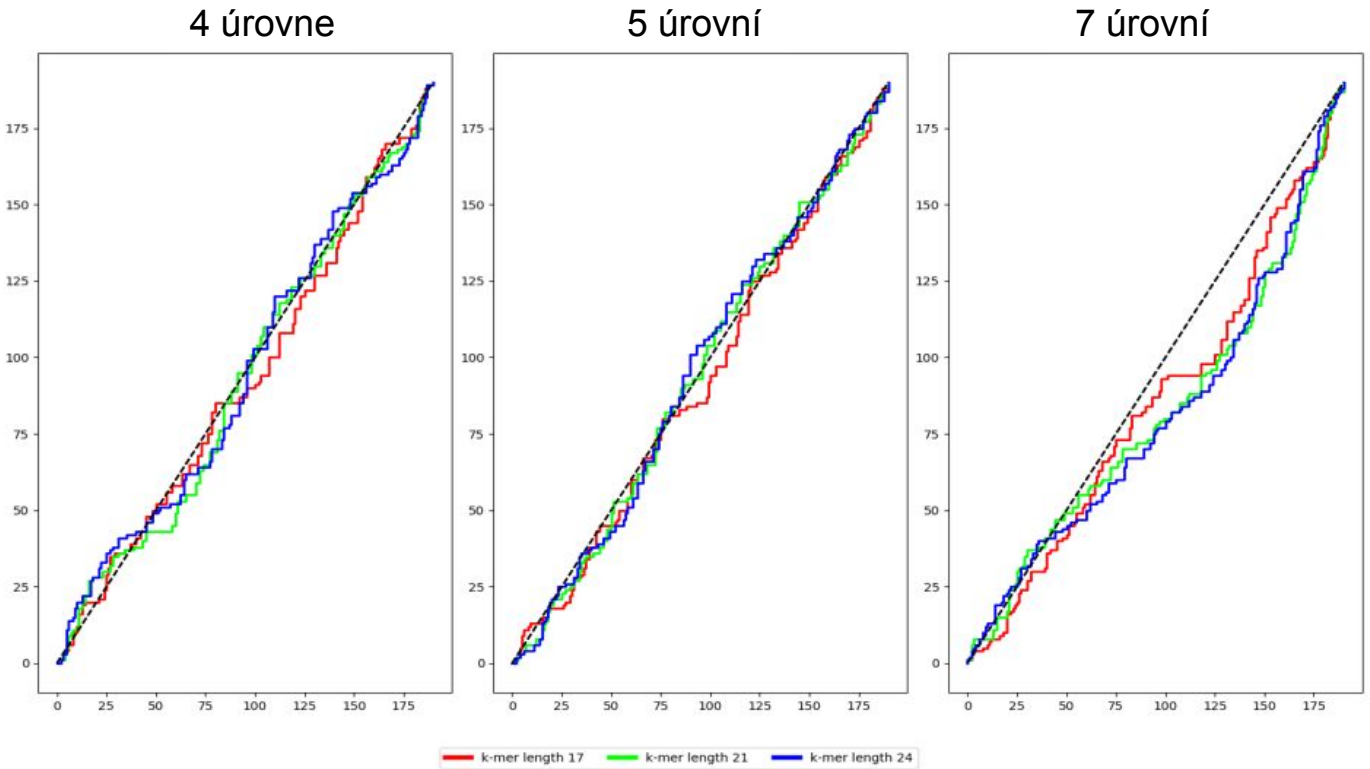
dĺžka k-tice:

Budovanie indexu 1

Postup:

1. Tvorba umelého signálu z DNA referencie
2. Vytvorenie úrovňového reťazca
3. Rozdelenie reťazca na prekrývajúce sa k-tice
4. Vloženie k-tic do hešovacej tabuľky

Výsledky - budovanie indexu 1



Budovanie indexu 2

Postup:

1. Použitie klasického zarovnávacieho algoritmu minimap2 [Li. H, 2018]
2. Nájdenie úrovňového reťazca reálneho signálu v referenčnom úrovňovom reťazci
3. Použitie substitúcie pre počet levelov počet úrovní4:

   ‘a’ -> ‘A’ ‘b’ -> ‘C’ ‘c’ -> ‘G’ ‘d’ -> ‘T’

4. To umožní použiť tradičné algoritmy na hľadanie DNA sekvencie vo veľkej referencii

Zhrnutie

- V práci sa podarilo preskúmať diskretizačnú metódu
    - sama o sebe má isté limitácie
    - viaceré metódy viedli k jej zlepšeniu
- Priamočiare budovanie indexu neúspešné
    - prevedených viacero experimentov identifikujúcich problémy
- Preformulovanie problému na iný bioinformatický problém
    - získanie informácií o kvalite diskretizačnej metódy
    - získanie intuície o ďalšom smerovaní
- Inšpirácia pre budúcu prácu v tejto oblasti

Moje vyjadrenie k posudku oponenta

- Vyjadrenie k rýchlosti metódy používajúcej algoritmus minimap2
  - Payne, Alexander, et al. "Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels." (2020)

- Iné typy filtrácie falošných hitov v indexe

- Vyjadrenie k celkovému záveru posudku

Zdroje obrázkov

https://mcic.osu.edu/genomics/nanopore-minion-sequencing

https://ucscgenomics.soe.ucsc.edu/ucsc-researchers-awarded-a-record-number-of-patents-last-year/

Výzvy bakalárskej práce

1. Prienik do problematiky a zorientovanie sa v nej
    a. Zapísanie si predmetu Metódy v bioinformatike
    b. Ročníkový projekt na bioinformatickú tému
2. Veľká časť práce tvorená experimentmi
    a. Časté stretnutia so školiteľom
    b. Štúdium iných prác
3. Problematika s neexistujúcim priamočiarym riešením
    a. Množstvo implementácii s veľkým percentom omylov
4. Prezentácia neúspešných experimentov