Identifikácia sekvenčných variantov v kontexte neistoty

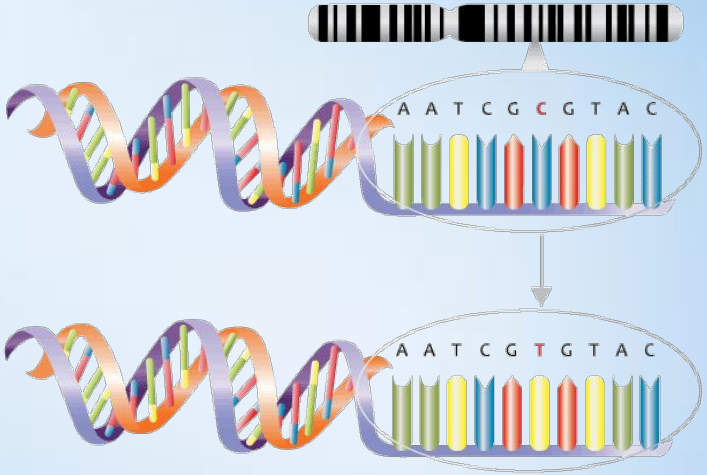
Školiteľ: doc. Mgr. Tomáš Vinař
Tomáš Janeta

Úvod

- V práci sa zaoberáme problémom hľadania sekvenčných variantov
- Popisujeme dva základné prístupy k tomuto problému
- Tieto prístupy implementujeme a testujeme na rôznych typoch dát
- Skúmame vplyv použitia reťazcov vygenerovaných pomocou CTC vrstvy namiesto klasických čítaní na výsledky oboch algoritmov

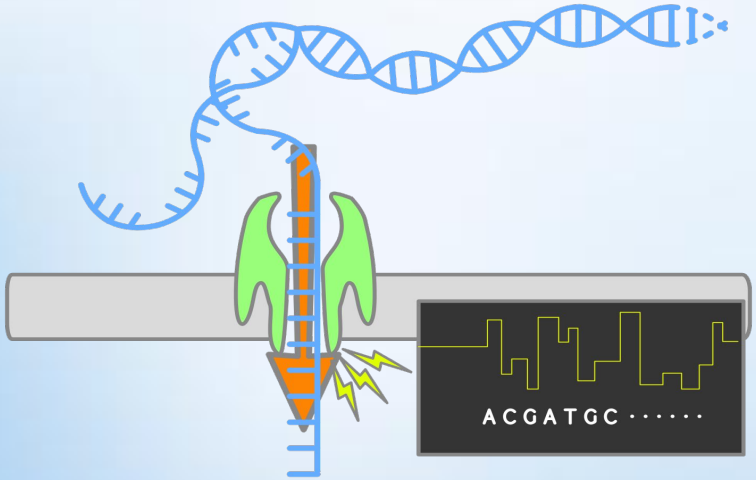
Genetický variant

- Nositeľom genetickej informácie v prírode je DNA
- Variantom nazývame zmenu v DNA postupnosti



Nanopórové sekvenovanie

- Metóda sekvenovania DNA
- Výhody - dlhé čítania, rýchlosť, dostupnosť prvých dát už počas čítania
- Nevýhody - veľké množstvo nepresností vzniknutých počas sekvenovania



CTC vrstva

- CTC je metóda z oblasti neurónových sietí, používaná pri úlohách kde treba dlhšej postupnosti hodnôt X priradiť kratšiu postupnosť znakov Y z konečnej množiny

Vzorkovanie

- Zo surového signálu, ktorý je výsledkom nanopórového sekvenovania vygenerujeme pomocou CTC vrstvy pravdepodobnostné rozdelenia nad množinou znakov A,C,T,G,N
- Pomocou týchto rozdelení vygenerujeme reťazce ktoré použijeme namiesto čítaní

Cieľ práce

- Opísať a porovnať niekoľko základných prístupov pri hľadaní sekvenčných variantov
- Implementovať tieto prístupy
- Vyhodnotiť oba prístupy na rôznych typoch dát
- Skúmať vplyv vzorkovania na výsledky implementovaných prístupov

Motivácia

- Hľadanie variantov (angl.variant calling) má široké využitie v medicíne a pri testovaní antibiotík
- V súčasnosti existuje mnoho metód na hľadanie variantov, v našej práci opisujeme niektoré z nich
- Najznámejšie nástroje na hľadanie variantov sú GATK, Bcftools,...

Pileup prístup

- Priamočiary prístup
- Nájde pozície ktoré sa nezhodujú s referenciou
- Prístup má dva vstupné parametre:
  - a - minimálny počet čítaní ktoré musia podporovať variant, aby ho algoritmus uznal ako validný
  - r - minimálne percento čítaní, obsahujúcich pozíciu s variantom ktoré ho musia podporovať aby ho algoritmus uznal ako validný

FB prístup

- Sofistikovanejší prístup
- Inšpirovaný softvérovým nástrojom Freebayes
- Berie do úvahy aj zarovnanú sekvenciu
- Prístup má dva vstupné parametre:
  - a - minimálny počet čítaní ktoré musia podporovať variant, aby ho algoritmus uznal ako validný
  - r - minimálne percento čítaní, obsahujúcich pozíciu s variantom ktoré ho musia podporovať aby ho algoritmus uznal ako validný
  - p - minimálny počet pozícií, o ktoré budeme každý podozrivý interval rozširovať doprava aj doľava

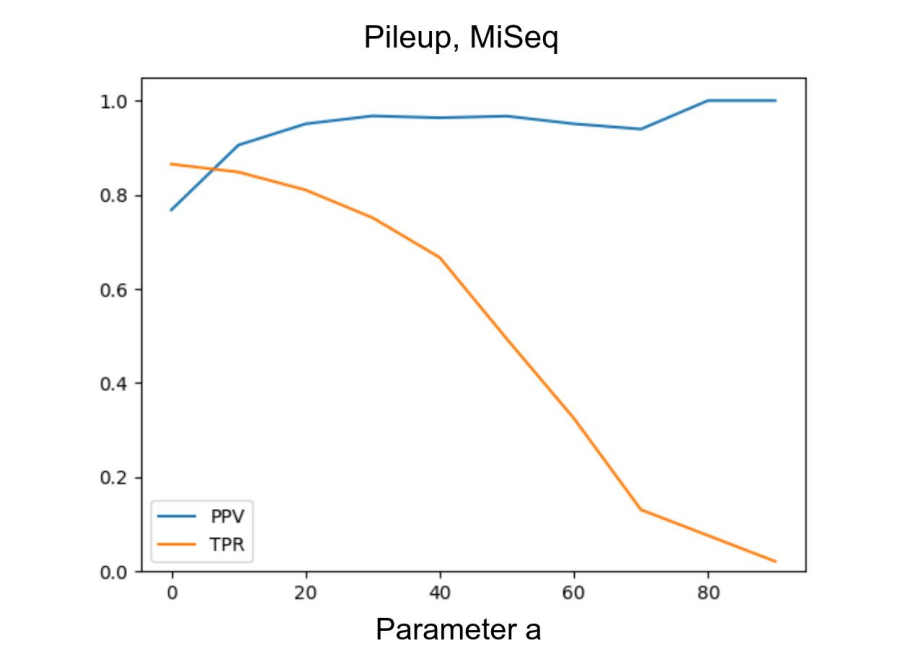
Vyhodnotenie

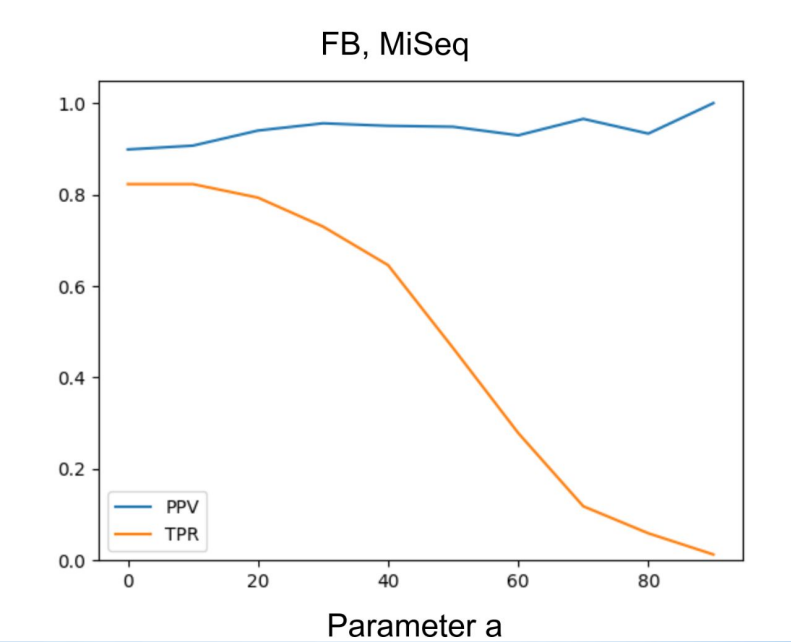
- Vstupom pre vyhodnotenie je zoznam variantov ktoré považujeme za správne
- Tieto varianty sme získali pomocou nástroja Bcftools
- Pri vyhodnocovaní sme sa riadili dvoma indikátormi:
  - positive predictive value (PPV) - pomer všetkých správnych variantov ktoré boli nájdené ku všetkým variantom ktoré boli nájdené
  - true positive rate (TPR) - pomer všetkých správnych variantov ktoré boli nájdené ku všetkým správnym variantom

Prvá vzorka

- Dáta pochádzajú z genómu baktérie Eschericia coli
- Dáta boli získané pomocou technológie MiSeq
- Čítania sa vyznačujú vysokou presnosťou a obsahujú predovšetkým SNPy.

Porovnanie oboch prístupov na prvej vzorke

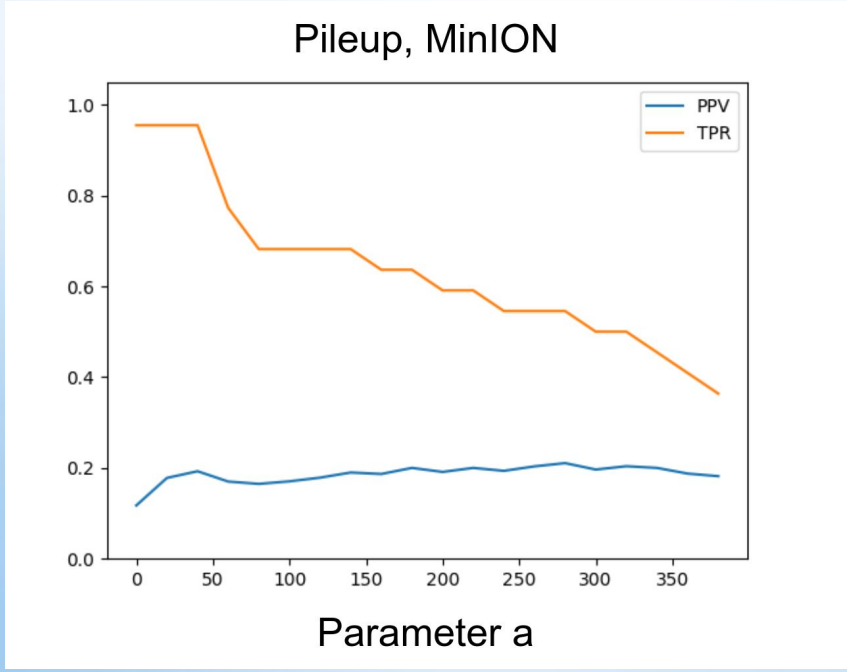


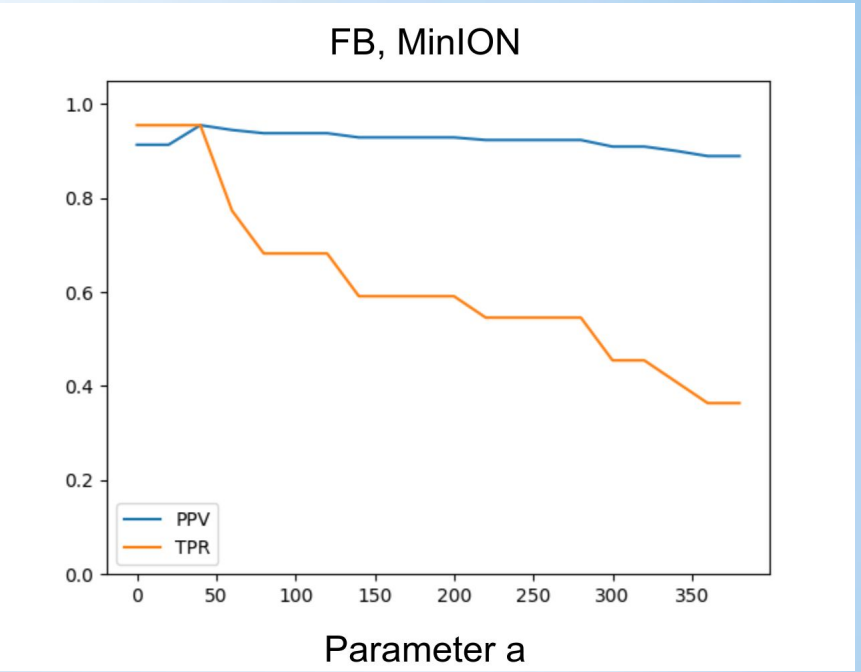


Druhá vzorka

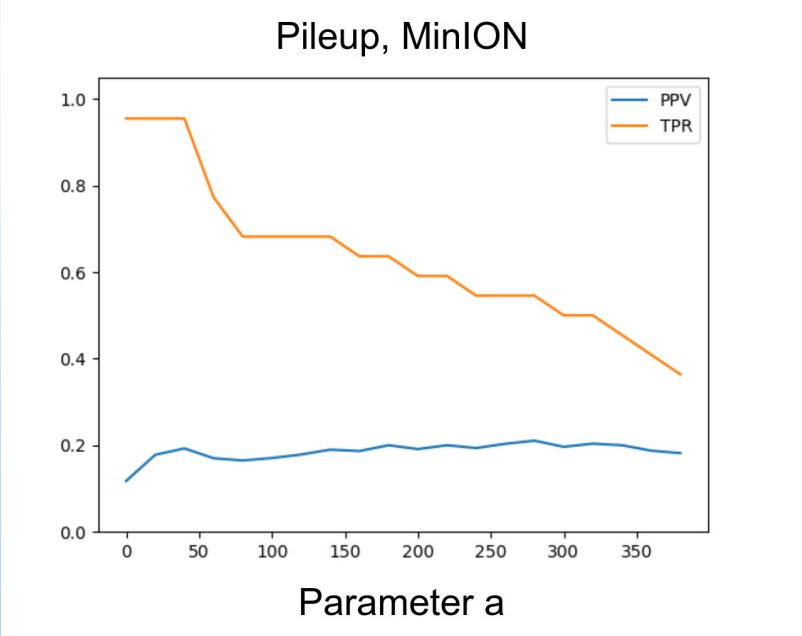
- Dáta pochádzajú z genómu vírusu SARS-COV-2
- Dáta boli získané pomocou technológie nanopórového sekvenovania
- Čítania sa vyznačujú nízkou presnosťou a obsahujú veľké množstvo nesprávnych variantov, predovšetkým delécií

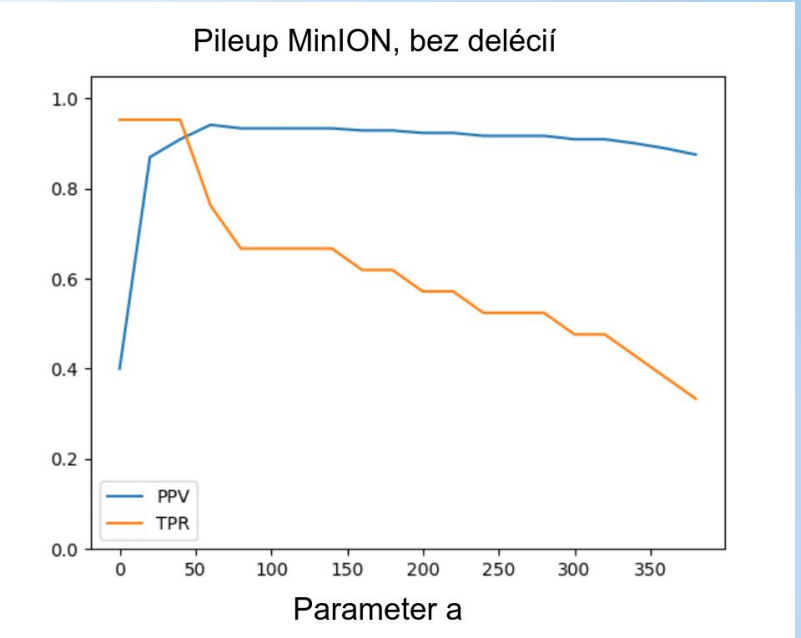
Porovnanie oboch prístupov na druhej vzorke



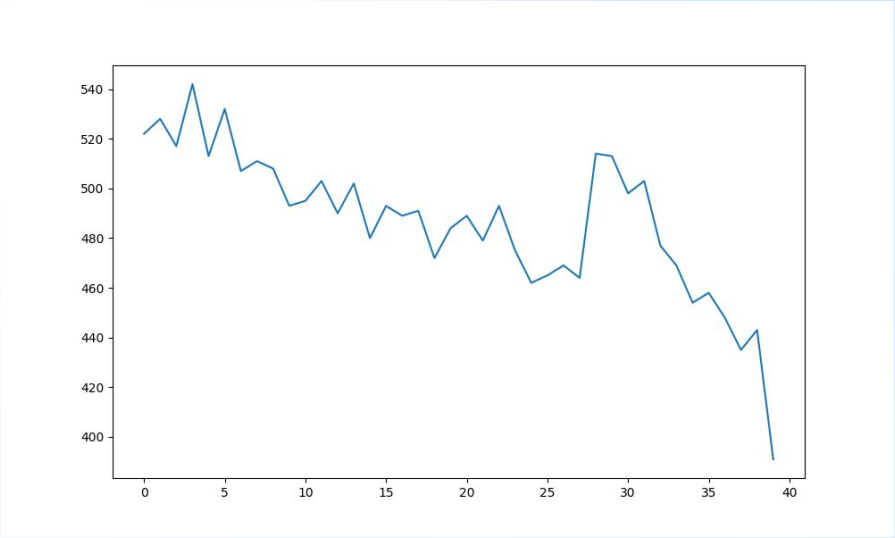


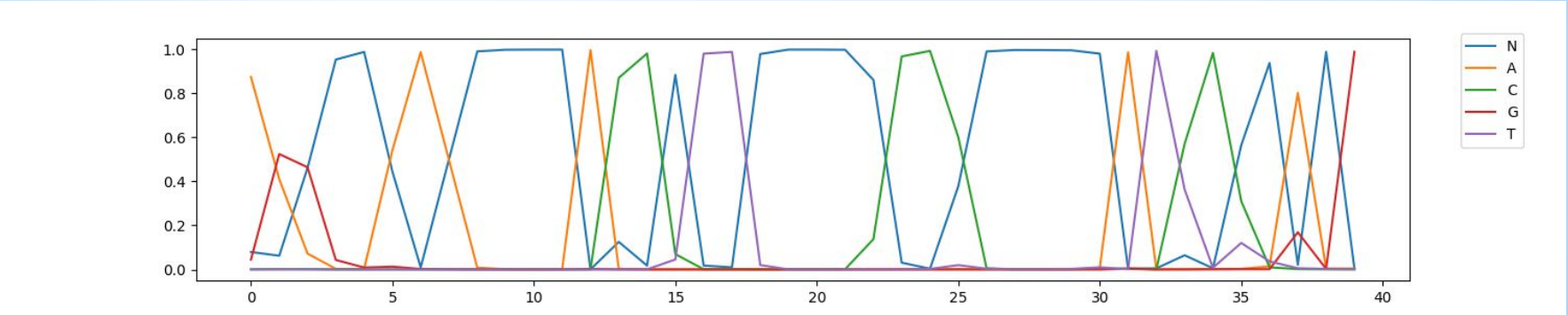
Porovnanie pileup prístupu s deléciami a bez



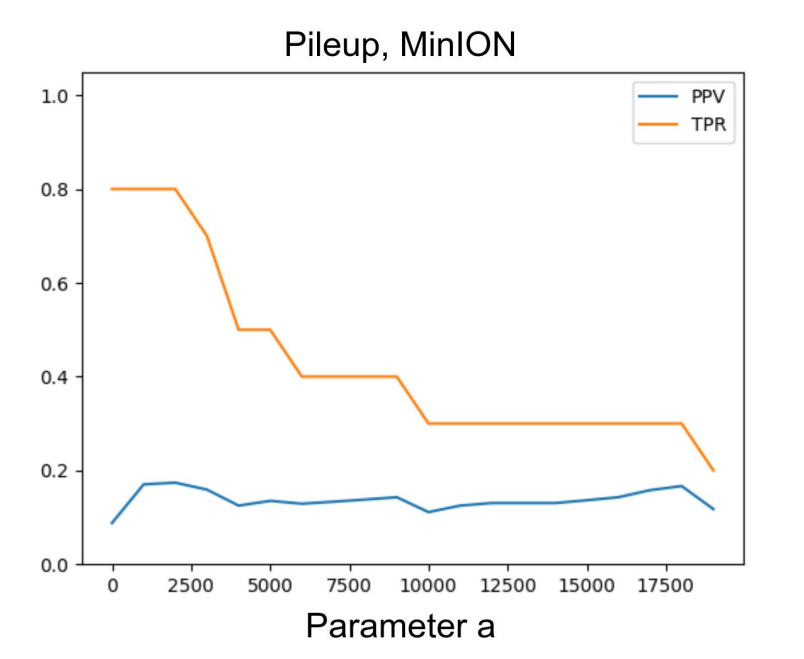


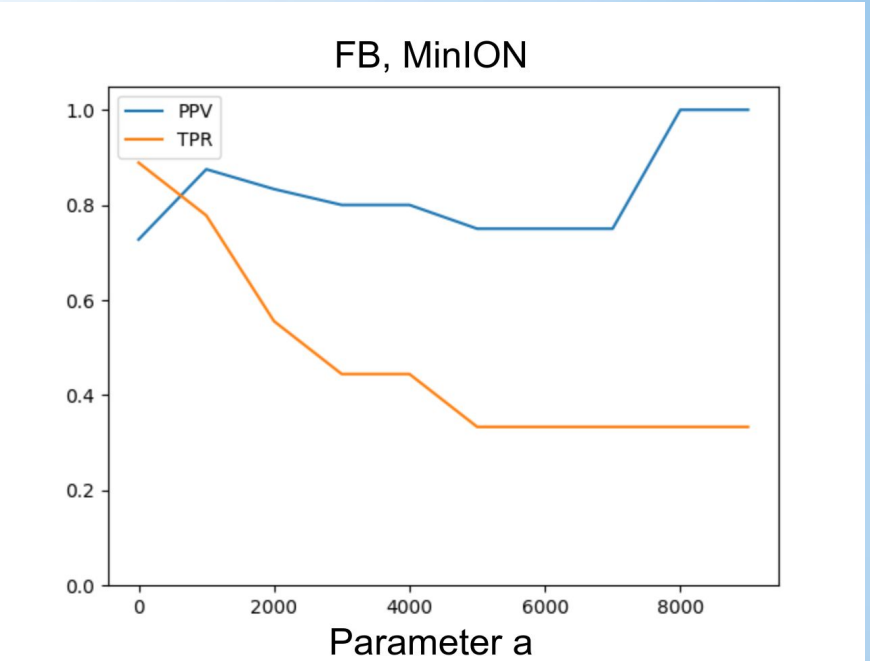
Vzorkovanie





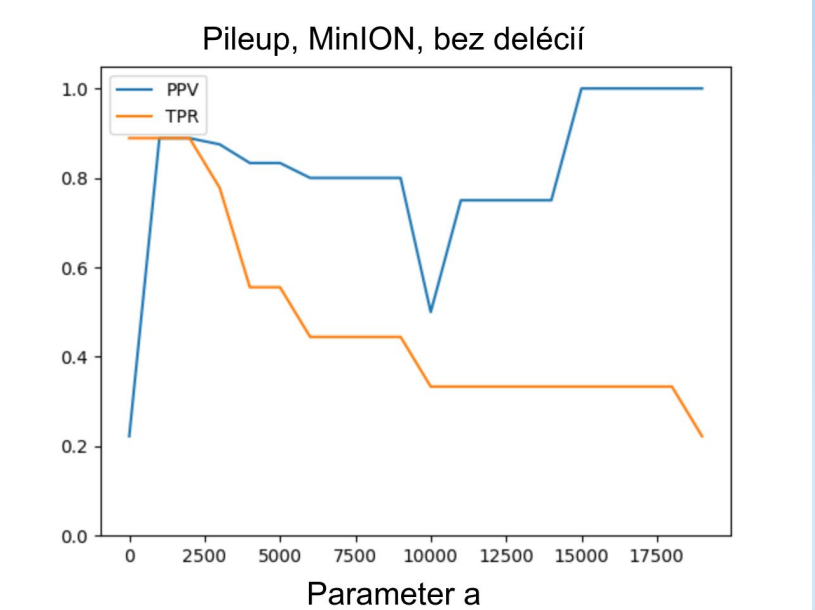
Vzorkovanie





Pripomienky oponenta

1. Obr. 4.2: skúmali ste výraznejší prepad hodnoty PPV pri “a=10000”? Ako si ho vysvetľujete?



Pripomienky oponenta

2. Dokázalo by podľa Vás vzorkovanie pomocou CTC vrstvy pomôcť spoľahlivo určiť varianty v úsekoch s veľmi nízkym pokrytím?

3. Sú vzorky v experimentoch v časti 4.1 pre každú hodnotu n generované nanovo, alebo je po zvýšení n k existujúcej množine pridaný rozdiel? O tomto v texte nie je zmienka.

Pripomienky oponenta

4. Malo by podľa Vás zmysel brať do úvahy pôvodné čítania s väčšou váhou ako z neho vygenerované vzorky?

5. Malo by podľa Vás zmysel namiesto náhodného vzorkovania len deterministicky vygenerovať k najpravdepodobnejsich vzoriek z každého čítania?

Zdroje obrázkov

- https://cdn.openpr.com/T/2/T212642922_g.jpg
- https://miro.medium.com/max/3840/0*glhxCHhF3SBOsMYa.png
- https://nanoporetech.com/sites/default/files/s3/minion-product3.jpg

Ďakujem za pozornosť