PANEURÓPSKA VYSOKÁ ŠKOLA
FAKULTA INFORMATIKY

FI-123456-12345

Pomocne Algorithmy na tvorbu Dna Kanalov pre Sekvencovanie

Bakalárska práca

2026

Patrik Homola

PANEURÓPSKA VYSOKÁ ŠKOLA
FAKULTA INFORMATIKY

Pomocne Algorithmy na tvorbu Dna Kanalov pre Sekvencovanie

Bakalárska práca

Patrik Homola

Študijný program:	Aplikovaná informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Ústav aplikovanej informatiky
Školiteľ:	doc. RNDr. Peter Farkas, PhD.

Bratislava 2026

PANEURÓPSKA VYSOKÁ ŠKOLA

Fakulta informatiky

ZADANIE PRÁCE

Meno a priezvisko študenta:
Evidenčné číslo práce:
Študijný odbor:
Študijný program:
Forma a metóda štúdia:
Vedúci práce:
Ústav/katedra:
Dátum zadania práce:

Názov:

Anotácia: Cieľom práce je vytvorenie komplexného systému pre navigáciu v interiéroch s napojením na rozličné webové zdroje s možnosťou ich editovania a návrh shaderov pre ich zobrazenie. Súčasťou práce bude analýza použitia v reálnom nasadení vrátane implementácie a zobrazovania na rozličných dotykových zariadeniach.

.....

RNDr. Ján Lacko, PhD.
vedúci práce

.....

Ing. Juraj Štefanovič, PhD.
vedúci ústavu

Poďakovanie:

Pri vypracovaní tejto práce by som sa rád poďakoval za odbornú pomoc pánovi doc. RNDr. Peter Farkas, PhD., ktorý mi poskytol mnoho cenných rád a konzultácií.

Poďakovanie môžete napísať podľa seba.

Čestné prehlásenie:

Čestne vyhlasujem, že záverečnú prácu som vypracoval samostatne a že som uviedol všetku použitú literatúru.

. . . . . .

Patrik Homola

Abstrakt

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer eu lacus leo. Nulla egestas purus non dignissim tincidunt. In sit amet tellus bibendum, lobortis magna ut, sodales mi. Etiam vel eros efficitur purus ultrices vulputate et quis justo. Nunc ultrices tellus a dui mattis, eget laoreet arcu tempor. Donec vestibulum, magna ac fringilla lacinia, libero risus fringilla arcu, nec consectetur nibh justo vel sem. Quisque gravida sit amet elit ut aliquam.

Abstrakt obsahuje informáciu o cieľoch práce, jej stručnom obsahu a v závere abstraktu sa charakterizuje splnenie cieľa, výsledky a význam celej práce. Súčasťou abstraktu je 3 - 5 kľúčových slov.

Kľúčové slová: Kľúčové slovo 1, Kľúčové slovo 2, Kľúčové slovo 3

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer eu lacus leo. Nulla egestas purus non dignissim tincidunt. In sit amet tellus bibendum, lobortis magna ut, sodales mi. Etiam vel eros efficitur purus ultrices vulputate et quis justo. Nunc ultrices tellus a dui mattis, eget laoreet arcu tempor. Donec vestibulum, magna ac fringilla lacinia, libero risus fringilla arcu, nec consectetur nibh justo vel sem. Quisque gravida sit amet elit ut aliquam.

Abstrakty v anglickom a slovenskom jazyku sa musia zhodovať.

Keywords: Keyword 1, Keyword 2, Keyword 3

Obsah

Zoznam obrázkov

Zoznam skratiek a značiek

BPMN Business Process Model and Notation
Notácia pre modelovanie business procesov

DNA Deoxyribonucleic Acid
Deoxyribonucleic Acid - DNA je molekula nesúca genetické informácie používané v rastlinách, zvieratách a väčšine organizmov. Je tvorená dvoma reťazcami nukleotidov, ktoré tvoria dvojitú špirálu.

FASTQ Fast Quality
Fast Quality - Formát súboru používaný na ukladanie sekvenčných dát spolu s kvalitatívnymi skóre pre každú bázu.

OCL Object Constraint Language
Jazyk pre špecifikáciu obmedzení, dotazov a vyhľadávacích operácií v UML

RNA Ribonucleic Acid
Ribonucleic Acid - RNA je molekula podobná DNA, ktorá hrá kľúčovú úlohu v procese prepisu a prekladu genetickej informácie. Na rozdiel od DNA je zvyčajne jednovláknová a obsahuje uracil namiesto thyminu.

SNR Signal-to-Noise Ratio
Signal-to-Noise Ratio - Poměr medzi úrovňou požadovaného signálu a úrovňou šumu, ktorý môže ovplyvniť kvalitu dát získaných z nanopórového sekvenovania.

UML Unified Modeling Language
Univerzálny jazyk pre vizuálne modelovanie systémov

1 Úvod

V tejto bakalárskej práci sa venujem problematike ukladania dát do DNA a spracovaniu sekvenačných dát získaných pomocou nanopórového sekvenovania. Práca je rozdelená do niekoľkých kapitol, ktoré postupne predstavujú motiváciu, teoretické východiská, návrh riešenia, implementáciu a vyhodnotenie výsledkov.

2 Motivácia

V dnešnej digitálnej dobe sa množstvo generovaných dát exponenciálne zvyšuje, čo vytvára potrebu efektívnych a trvalo udržateľných metód ich ukladania. Tradičné úložné médiá, ako sú pevné disky, SSD disky a magnetické pásky, majú obmedzenú životnosť a kapacitu, čo vedie k častým migráciám dát a zvýšeným nákladom na údržbu. Naopak, DNA ako médium pre ukladanie dát ponúka niekoľko významných výhod:

- **Vysoká hustota ukladania:** DNA má neuveriteľne vysokú hustotu ukladania dát, kde jeden gram DNA môže teoreticky uložiť až 215 petabajtov (215 miliónov gigabajtov) dát. To znamená, že obrovské množstvo informácií môže byť uložené v mimoriadne malom objeme.
- **Dlhodobá stabilita:** DNA je chemicky stabilná molekula, ktorá môže prežiť tisíce rokov za správnych podmienok. Na rozdiel od tradičných médií, ktoré sa môžu degradovať počas niekoľkých dekád, DNA môže uchovávať dáta po veľmi dlhú dobu bez straty integrity.
- **Energetická efektívnosť:** Ukladanie dát do DNA nevyžaduje neustálu energiu na udržiavanie dát, na rozdiel od elektronických úložísk, ktoré potrebujú napájanie na zachovanie informácií. To vedie k výrazným úsporám energie a znižuje ekologickú stopu dátových centier.
- **Ľahká replikácia a prenosnosť:** DNA môže byť ľahko kopírovaná pomocou biologických procesov, čo umožňuje jednoduché zálohovanie a prenos dát medzi rôznymi miestami bez potreby špecializovaného hardvéru.
- **Odolnosť voči technologickému zastaraniu:** Na rozdiel od tradičných úložných médií, ktoré môžu rýchlo zastarať v dôsledku technologického pokroku, DNA ako médium zostáva nezmenené a čitateľné pomocou základných biologických techník.

Napriek týmto výhodám, proces čítania DNA pomocou nanopórov prináša špecifické výzvy. Pri sekvenovaní DNA cez nanopór sa DNA molekula prechádza cez nanopór po jednotlivých nukleotidoch, pričom elektrický signál identifikuje bázy. Tento proces je však náchylný na chyby, keďže DNA sa môže pohybovať rýchlo alebo nepravidelne, čo vedie k nesprávnemu určeniu sekvencie nukleotidov. Tieto chyby v čítaní majú priamy vplyv na integritu uložených dát a vyžadujú sofistikované metódy kódovania a korekcie chýb. Avšak s pokračujúcim vývojom technológií nanopórového sekvenovania sa presnosť postupne zvyšuje, čo robí z DNA perspektívnu alternatívu k tradičným metódam ukladania dát.

Cieľom tejto práce je prispiet k zefektívneniu procesu ukladania dát do DNA vytvorením DNA kanála s lepšími vlastnosťami pre čítanie a zápis dát. Optimalizáciou týchto procesov môžeme zvýšiť rýchlosť a presnosť ukladania a načítavania informácií, čo je kľúčové pre praktické využitie DNA ako úložného média v budúcnosti.

DNA kanál predstavuje špecifický spôsob, akým sú dáta kódované a ukladané do DNA molekúl. Tento kanál zahŕňa výber vhodných sekvencií nukleotidov, ktoré minimalizujú chyby pri čítaní a zápise dát, ako aj optimalizáciu procesov syntézy a sekvenovania DNA.

3 Ciele a zadanie práce

3.1 Uloha

Cieľom tejto úlohy je simulovať chyby, ktoré vznikajú pri čítaní DNA pomocou nanopórového sekvenovania. Nanopórové sekvenovanie je inovatívna technológia, ktorá umožňuje rýchle a efektívne čítanie genetického materiálu. Úlohou je vytvoriť generátor, ktorý na základe pravdepodobnosti chyby náhodne pridáva chyby do reálneho DNA reťazca. Tento generátor by mal byť schopný simulovať rôzne typy chýb, ako sú záměny báz, vloženia alebo vynechania báz, ktoré sa môžu vyskytnúť počas procesu sekvenovania.

Táto simulácia bude slúžiť na testovanie a vyhodnocovanie algoritmov pre spracovanie sekvenačných dát, ktoré musia byť robustné voči týmto chybám. Cieľom je zlepšiť presnosť a spoľahlivosť analýzy genetických informácií získaných pomocou nanopórového sekvenovania.

4 Návrh riešenia

4.1 Navrh riesenia

Navrhnutie riešenia úlohy simulácie chýb pri čítaní DNA pomocou nanopórového sekvenovania zahŕňa niekoľko kľúčových krokov. Cieľom je vytvoriť systém, ktorý dokáže generovať realistické chyby v DNA sekvenciách na základe pravdepodobností chýb pozorovaných v skutočných nanopórových dátach.

Aby sme mohli dosiahnuť tento cieľ, najprv je potrebné získať reálne DNA sekvencie a ich zodpovedajúce výstupy z nanopórového sekvenovania. Tieto dáta budú slúžiť ako základ pre analýzu a modelovanie chýb. Pre každý reálny DNA reťazec je potrebné získať dostatočné množstvo výstupných sekvencií, ktoré boli generované počas sekvenovania. Čím väčšie množstvo výstupov máme pre jeden vstupný reťazec, tým lepšie môžeme pochopiť variabilitu a pravdepodobnosti chýb, ktoré sa vyskytujú počas procesu sekvenovania.

Problem je ze nevieme Ake je realne dna. Riesenie tohto problému spočíva v použití techník zhlukovania (clustering) na identifikáciu skupín výstupných sekvencií, ktoré pravdepodobne pochádzajú z rovnakého pôvodného DNA reťazca. Týmto spôsobom môžeme vytvoriť konsenzuálny reťazec pre každý zhluk, ktorý bude slúžiť ako náš "reálny"vstupný reťazec. Cim vacsi zhluk najdeme, tym vacsia sanca ze konzesus bude reprezentovat realny vstup a tym lepsie data pre nasu analýzu generované dna. Týmto spôsobom získame nielen reálny reťazec, ale aj informácie o tom, akými chybami sa tento reťazec líšil od jednotlivých výstupov. Pre každú bázu v reálnom reťazci tak môžeme vytvoriť mapu, ktorá ukazuje, na aký reťazec bola táto báza prečítaná v rôznych výstupoch.

5 Technologické pozadie

Táto časť predstavuje základné technológie a algoritmy použité v práci.

6 Nanopore

Nanopore je technológia používaná na sekvenovanie DNA a RNA molekúl. Táto metóda umožňuje čítanie genetického materiálu tým, že molekuly prechádzajú cez nanometrové póry, čo vedie k zmene elektrického prúdu, ktorý je následne analyzovaný na určenie sekvencie nukleotidov. Je to revolučná technológia v oblasti genomiky, ktorá umožňuje rýchle a presné sekvenovanie s minimálnymi nákladmi a zariadeniami prenosných veľkostí. Aplikácie nanopórov zahŕňajú diagnostiku chorôb, výskum genetických porúch a monitorovanie environmentálnych vzoriek.

Nanopore sekvenovanie je obzvlášť užitočné pre jeho schopnosť čítať dlhé sekvencie DNA, čo zjednodušuje zostavovanie genómov a identifikáciu štruktúrnych variácií. Funguje na princípe detekcie zmien v elektrickom prúde, keď molekula prechádza cez nanopór, čo umožňuje priame čítanie sekvencií bez potreby amplifikácie alebo značenia. DNA alebo RNA molekuly sú vedené cez nanopór pomocou elektrického poľa, pričom každá báza spôsobuje charakteristickú zmenu prúdu, ktorá je zaznamenaná a analyzovaná softvérom na určenie sekvencie.

6.1 Prekladanie DNa/Mra do signálu

Prekladanie DNA alebo RNA sekvencií do elektrického signálu v nanopórovom sekvenovaní zahŕňa niekoľko kľúčových krokov:

- **Príprava vzorky:** DNA alebo RNA molekuly sú pripravené na sekvenovanie, často zahŕňajúce fragmentáciu a pridanie adaptérnych sekvencií.
- **Vedenie cez nanopór:** Molekuly sú vedené cez nanopór pomocou elektrického poľa, ktoré spôsobuje ich pohyb.
- **Detekcia prúdu:** Keď molekula prechádza cez nanopór, každá báza spôsobuje charakteristickú zmenu v elektrickom prúde, ktorý je zaznamenaný ako časová séria dát.
- **Analýza signálu:** Zaznamenaný elektrický signál je analyzovaný pomocou algoritmov na identifikáciu sekvencie nukleotidov na základe zmien prúdu.
- **Preklad do sekvencie:** Softvér prekladá analyzovaný signál späť do sekvencie DNA alebo RNA, čo umožňuje výskumníkom získať genetické informácie z pôvodnej molekuly.

Výstupom z procesu nanopórového sekvenovania je sekvencia nukleotidov (A, T, C, G pre DNA alebo A, U, C, G pre RNA) reprezentujúca genetickú informáciu pôvodnej molekuly. Tento výstup je často vo forme textového súboru (napríklad FASTQ formát), ktorý obsahuje sekvencie spolu s kvalitatívnymi skóre pre každú bázu. Vzhľadom na technické obmedzenia a variabilitu v procese sekvenovania je bežné, že jednotlivé DNA alebo RNA

retiazce sú prečítané niekoľkokrát (tzv. "coverageälebo "depth of coverage"). Toto opakované čítanie zvyšuje spoľahlivosť a presnosť výslednej sekvencie, pretože umožňuje korekciu chýb a identifikáciu variácií v genetickom materiáli. No napriek tomu, že sa retiazce čítajú viackrát, stále môže dôjsť k nepresnostiam v sekvencii kvôli šumu v signáli.

Retazce DNA alebo RNA môžu byť čítané rôznou rýchlosťou počas nanopórového sekvenovania, čo môže ovplyvniť kvalitu a presnosť zaznamenaného signálu. Rýchlosť prechodu molekuly cez nanopór je ovplyvnená viacerými faktormi, vrátane veľkosti molekuly, jej konformácie, interakcií s nanopórom a podmienok prostredia (napríklad teplota a iónová sila).

Keďže DNA a RNA molekuly sú veľmi malé, signál zaznamenaný počas nanopórového sekvenovania odráža interakcie jednotlivých nukleotidov (báz) s nanopórom. Každá báza má jedinečný tvar a elektrické vlastnosti, ktoré ovplyvňujú spôsob, akým mení elektrický prúd pri prechode cez nanopór. Preto je možné získať signál pre každú jednotlivú bázu v molekule, čo umožňuje presné čítanie sekvencie.

7 Dorado Basecaller

Dorado je pokročilý basecaller vyvinutý spoločnosťou Oxford Nanopore Technologies (ONT) na prekladanie surového elektrického signálu získaného z nanopórového sekvenovania na sekvenciu nukleotidov (A, T, C, G pre DNA alebo A, U, C, G pre RNA). Je navrhnutý tak, aby poskytoval vysokú presnosť a rýchlosť pri spracovaní sekvenačných dát, čo je kľúčové pre rôzne aplikácie v oblasti genomiky a bioinformatiky.

7.1 Názov podkapitoly

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer eu lacus leo. Nulla egestas purus non dignissim tincidunt.

8 Metódy spracovania sekvencií

9 Cluster clover

Klastrovací algoritmus je metóda, ktorá rozdeľuje množinu dát do skupín (klastrov) tak, aby dáta v rámci jednej skupiny boli si navzájom podobné a dáta z rôznych skupín boli odlišné Qu et al., 2022. Tieto algoritmy sa často používajú v bioinformatike na analýzu sekvencií DNA alebo iných biologických dát.

Clover algoritmus je efektívny klastrovač navrhnutý špeciálne pre DNA sekvencie v oblasti DNA-based data storage. Využíva stromovú štruktúru na rýchle a presné zoskupovanie sekvencií podľa ich podobnosti, čím zvyšuje efektivitu a škálovateľnosť procesu Qu et al., 2022. Clover umožňuje efektívne spracovanie veľkých množín sekvencií a je vhodný pre moderné aplikácie v oblasti ukladania dát do DNA.

Viacnásobné zarovnanie sekvencií (Multiple Sequence Alignment, MSA) je proces usporiadania troch alebo viacerých biologických sekvencií (DNA, RNA alebo proteínov) tak, aby boli identifikované podobné regióny medzi nimi. Tieto podobnosti môžu odhaliť štrukturálne, funkčné alebo evolučné vzťahy medzi sekvenciami.

10 Základný princíp

Pri viacnásobnom zarovnaní sekvencií hľadáme také usporiadanie, ktoré maximalizuje podobnosť medzi všetkými sekvenciami súčasne. Do sekvencií sú vkladané medzery (gap characters, značené pomlčkou '-'), aby sa zodpovedajúce pozície dostali do rovnakých stĺpcov.

10.1 Príklad zarovnania

Majme tri jednoduché sekvencie:

- Sekvencia 1: `ACGT`
- Sekvencia 2: `AGT`
- Sekvencia 3: `ACCT`

Ich zarovnanie môže vyzerať nasledovne:

```
Seq1: A C G T
Seq2: A - G T
Seq3: A C C T
```

11 Význam a využitie

Viacnásobné zarovnanie má niekoľko dôležitých aplikácií:

- **Fylogenéza** – konštrukcia evolučných stromov
- **Identifikácia konzervovaných motívov** – nálezenie funkčne dôležitých oblastí
- **Predikcia štruktúry** – odvodenie sekundárnej a terciárnej štruktúry proteínov
- **Homológia** – určenie príbuznosti medzi sekvenciami

12 Zložitosť problému

Viacnásobné zarovnanie je výpočtovo náročný problém. Presné riešenie pomocou dynamického programovania má časovú zložitosť $O(L^n)$, kde L je dĺžka sekvencií a n je ich počet. Preto sa v praxi používajú heuristické algoritmy.

13 Skórovacia funkcia

Kvalita zarovnania sa hodnotí pomocou skórovacej funkcie, ktorá typicky obsahuje:

- **Match score** – bodovanie zhody medzi znakmi
- **Mismatch penalty** – penalizácia za nezhodu
- **Gap penalty** – penalizácia za medzery (môže byť lineárna alebo afinná)

14 Bežné algoritmy

Medzi najpoužívanejšie algoritmy pre MSA patria:

- **ClustalW/ClustalOmega** – progresívne zarovnanie pomocou sprievodného stromu
- **MUSCLE** – rýchly progresívny algoritmus s iteratívnym spresňovaním
- **T-Coffee** – kombinuje párové zarovnania do finálneho MSA
- **MAFFT** – využíva FFT pre rýchle nálezenie homológnych oblastí

15 Vlastná implementácia a experimenty

V tejto kapitole popisujem konkrétne kroky, ktoré som vykonal počas riešenia práce, vrátane spracovania dát, aplikácie algoritmov a vyhodnotenia výsledkov.

Najskôr som stiahol potrebný dataset a pripravil prostredie Dorado, ktoré je určené na spracovanie sekvenačných dát.

Následne som spustil basecaller Dorado, ktorým som preložil súbory vo formáte pod5 do formátu BAM, čo umožňuje ďalšie spracovanie dát.

Potom som použil algoritmus Clover, ktorý zoskupil sekvencie do jednotlivých klastrov na základe ich podobnosti.

Zo všetkých vytvorených klastrov som vybral tie, ktoré boli dostatočne veľké na ďalšiu analýzu.

Na každý vybraný cluster som aplikoval nástroj Medaka, aby som vytvoril konsenzuálnu sekvenciu pre daný cluster.

Pre každú katicu (k-mer) z konsenzuálnej sekvencie som vybral reprezentatívnu katicu z alignovaných sekvencií v danom clustri.

Nakoniec som všetky získané dáta uložil pre ďalšie spracovanie alebo analýzu.

16 Záver

V práci som predstavil možnosti a výzvy spojené s ukladaním dát do DNA a spracovaním sekvenačných dát z nanopórového sekvenovania. Navrhnuté a implementované metódy prispievajú k efektívnejšiemu spracovaniu a analýze týchto dát. Cieľ bakalárskej práce bol splnený.

Použitá literatúra

QU, Guanjin; YAN, Zihui; WU, Huaming, 2022. Clover: tree structure-based efficient DNA clustering for DNA-based data storage. *Briefings in Bioinformatics*. Roč. 23, č. 5, s. 1–16. Dostupné z DOI: 10.1093/bib/bbac336. Corresponding author: Huaming Wu, Center for Applied Mathematics, Tianjin University, Tianjin, 300072, China. E-mail: whming@tju.edu.cn.