Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Identifikácia variantov v dátach nanopórového sekvenovania

Bakalárska práca

2018
Eduard Batmendijn

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Identifikácia variantov v dátach nanopórového sekvenovania

Bakalárska práca

Študijný program:	Informatika
Študijný odbor:	2508 Informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	doc. Mgr. Tomáš Vinař, PhD.

Bratislava, 2018
Eduard Batmendijn

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Eduard Batmendijn
Študijný program: informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Identifikácia variantov v dátach nanopórového sekvenovania
Variant Identification in Nanopore Sequencing Data

Anotácia: Nanopórové sekvenovanie produkuje sekvencie signálov, ktoré je možné porovnať so známou referenčnou DNA sekvenciou. Cieľom práce je navrhnúť a naimplementovať metódy, ktoré by umožnili identifikovať miesta, kde sa prečítané sekvencie signálov odlišujú od referencie. Efektívne metódy zohľadňujúce špecifické vlastnosti dát umožnia spoľahlivú identifikáciu takýchto variantov aj z dát s nízkym pokrytím.

Vedúci: doc. Mgr. Tomáš Vinař, PhD.
Katedra: FMFI.KAI - Katedra aplikovanej informatiky
Vedúci katedry: prof. Ing. Igor Farkaš, Dr.

Dátum zadania: 31.10.2017

Dátum schválenia: 31.10.2017 doc. RNDr. Daniel Olejár, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Abstrakt

Tu bude abstrakt.

Kľúčové slová:

Abstract

Slovak translation of abstract coming here.

Keywords:

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Keď už budeme vedieť, o čom nakoniec práca bude, na tomto mieste stručne zoznámime čitateľa s problematikou a štruktúrou našej práce.

Kapitola 1

Nanopórové sekvenovanie DNA

Genetická informácia je v prírode často kódovaná deoxyribonukleovou kyselinou (DNA[1]). DNA je tvorená dvoma vláknami spletenými do tvaru dvojzávitnice. Každé vlákno obsahuje postupnosť dusíkatých báz, ktorá kóduje informáciu. V DNA sa vyskytujú štyri dusíkaté bázy: adenín (A), cytozín (C), guanín (G) a tymín (T). Postupnosti báz v jednotlivých vláknach sú komplementárne: na pozíciách, kde má prvé vlákno adenín (resp. cytozín, guanín, tymín) má druhé vlákno tymín (resp. guanín, cytozín, adenín). Na zrekonštruovanie celej informácie nám teda stačí poznať poradie báz v jednom z vlákien.

Proces zisťovania poradia báz v DNA sa nazýva *sekvenovanie* DNA. Techniky sekvenovania DNA boli známe už v sedemdesiatych rokoch minulého storočia a od vtedy sa stále vyvíjajú. Pri sekvenovaní sa určí poradie dusíkatých báz vo fragmentoch DNA, nazývaných *čítania*. Z dostatočného počtu prekrývajúcich sa čítaní sa potom dá zrekonštruovať celá postupnosť báz v DNA. Pre rôzne sekvenačné technológie sú typické rôzne dĺžky čítaní, ktoré produkujú.

Jednou z najnovších sekvenačných technológií je nanopórové sekvenovanie. Vyznačuje sa dlhými čítaniami, nízkou cenou a dostupnosťou prvých dát už počas sekvenovania, ale aj veľkým množstvom chýb v jednotlivých čítaniach. Pri nanopórovom sekvenovaní sa vo vhodne zvolenej membráne vytvorí *nanopór*, t. j. otvor s priemerom rádovo 1nm. Membránou sa oddelia dve komory s elektrolytom, pričom v jednej z komôr sa nachádza aj predpripravená vzorka DNA. Po zavedení elektrického napätia medzi komorami začne nanopórom tiecť iónový prúd. Vďaka elektroforéze a za pomoci enzýmov sa jedno vlákno DNA postupne oddeľuje od druhého a prechádza nanopórom. Časť vlákna, ktorá sa práve nachádza v najužšej časti nanopóru, má vplyv na elektrický prúd tečúci cez nanopór. Rôzne bázy ovplyvňujú elektrický prúd rôznym spôsobom. Pri sekvenovaní sa meria priebeh elektrického prúdu v čase a na základe jeho zmien sa potom určuje, aké bázy prešli cez nanopór.

[1] z anglického ***d**eoxyribo**n**ucleic **a**cid*

TODO: obrázok

1.1 Sekvenátor MinION

Prístroj MinION je nanopórový sekvenátor vyrábaný firmou Oxford Nanopore Technologies. V našej práci budeme používať dáta získané týmto sekvenátorom. V MinIONe sa používa polymérová membrána, do ktorej sú zasadené proteínové nanopóry. Sekvenátor obsahuje stovky nanopórov, dokáže teda sekvenovať niekoľko DNA vlákien súčasne. Vo verzii, s ktorou pracujeme, prechádza vlákno DNA cez nanopór rýchlosťou približne 400 báz za sekundu. Hodnota elektrického prúdu sa zaznamenáva 4000-krát za sekundu, teda v priemere zhruba 10- krát na bázu. Namerané hodnoty prúdu sa pre každé čítanie ukladajú do zvlášť súboru vo formáte `.fast5`. Tieto nespracované dáta budeme nazývať *surový signál*, alebo, ak nebude hroziť nedorozumenie, skrátene iba *signál*.

TODO: obrázok

1.1.1 Normalizácia signálu

Surový signál závisí nielen od úseku DNA nachádzajúceho sa v nanopóre, ale aj od ďalších faktorov, ktoré sa pre rôzne čítania môžu líšiť. Pred ďalším spracovaním je preto potrebné surový signál znormalizovať.

Jednou z metód normalizácie je *mediánová normalizácia*, ktorú navrhujú Stoiber et. al. v [1].

Definícia 1. Nech $a_1, a_2, \ldots, a_n \in \mathbb{R}$. Symbolom

$$\text{MEDIAN}_{i=1}^{n}(a_i)$$

budeme značiť medián hodnôt $a_1, a_2, \ldots, a_n$.

Definícia 2. Nech $r_1, r_2, \ldots, r_n$ sú namerané hodnoty surového signálu. Nech

$$M = \text{MEDIAN}_{i=1}^{n}(r_i)$$

a nech

$$D = \text{MEDIAN}_{i=1}^{n}(|r_i - M|).$$

Mediánovo znormalizovaný signál je postupnosť $s_1, s_2, \ldots, s_n$ určená predpisom

$$s_i = \frac{r_i - M}{D}.$$

1.1.2 Určovanie báz

Pri určovaní báz sa využíva fakt, že rôzne bázy pri svojom prechode nanopórom ovplyvňujú signál rôznym charakteristickým spôsobom. V praxi však signál nie je ovplyvnený iba jednou bázou. Pracuje sa preto s predpokladom, že signál je ovplyvnený k po sebe idúcimi bázami, ktoré sú práve najbližšie k nanopóru.

Kapitola 2

Zarovnanie signálu k referencii

V tejto kapitole popíšeme algoritmus na zarovnanie surového signálu k referenčnej sekvencii.

Kapitola 3

Identifikácia jednonukleotidových polymorfizmov

V tejto časti navrhneme techniku na odhaľovanie jednonukleotidových polymorfizmov (SNP[1]) v sekvenovaných dátach.

[1] z anglického ***s**ingle **n**ucleotide **p**olymorphism*

Kapitola 4

Identifikácia inzercií a delécií

Našu techniku identifikácie SNP v sekvenovaných dátach sa pokúsime rozšíriť aj na inzercie a delécie.

Kapitola 5

Testovanie

V tejto kapitole uvedieme metodiku a výsledky testovania nášho prístupu na reálnych dátach.

Kapitola 6

Implementácia

V tejto kapitole odkážeme čitateľa na zverejnenú implementáciu nami navrhovaných techník.

Záver

Na záver zhrnieme výsledky, ktoré sa nám podarilo dosiahnuť.

Literatúra

[1] Marcus H Stoiber, Joshua Quick, Rob Egan, Ji Eun Lee, Susan E Celniker, Robert Neely, Nicholas Loman, Len Pennacchio, and James B Brown. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *bioRxiv*, 2017.