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

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.

1.1 Nanopórové sekvenovanie DNA

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

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

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.

1.1.1 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*.

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 [4].

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

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

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

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

a nech

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

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

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

Určovanie báz

Na základe signálu nameraného MinIONom sa určuje, aké dusíkaté bázy prechádzali nanopórom, keď bol tento signál zaznamenaný. Táto úloha je pomerne náročná a v súčasnosti sa stále vyvíjajú lepšie a lepšie riešenia. Programy, ktoré určujú bázy, nazývame *basecallery*[2].

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. Skupinám k po sebe idúcich báz sa hovorí *k-mery*.

Ďalším problémom je, že vlákno DNA cez nanopór neprechádza konštantnou rýchlosťou. Jednotlivým bázam vo výslednej postupnosti preto môžu zodpovedať rôzne dlhé úseky signálu. Prvým krokom pri určovaní báz preto často býva rozdelenie signálu na úseky, v rámci ktorých bola hodnota signálu približne konštantná. Týmto úsekom sa hovorí *udalosti*. Pri ďalšom spracovaní sa predpokladá, že medzi jednotlivými udalosťami sa vlákno DNA väčšinou posunie o jednu bázu. Keďže však rozdelenie signálu na udalosti nemusí presne zodpovedať posunom DNA vlákna v nanopóre, uvažuje sa aj možnosť, že sa vlákno medzi udalosťami neposunulo, prípadne posunulo o viac než jednu bázu.

Niektoré basecallery (napr. Nanocall [3]) modelujú prechod DNA vlákna nanopórom ako skrytý Markovovský model. Skrytým stavom je k-mer, ktorý sa práve nachádza v nanopóre a ovplyvňuje signál. Viditeľným výstupom modelu sú udalosti. Z každého stavu (k-meru) sú možné prechody do štyroch k-merov, ktoré po ňom môžu nasledovať. Napríklad pre $k = 6$ sa zo stavu `ACCGCT` dá prejsť do stavov `CCGCTA`, `CCGCTC`, `CCGCTG` a `CCGCTT`. Okrem toho je ešte, s menšou pravdepodobnosťou, možný prechod naspäť do toho istého stavu (modelujúci udalosti, pri ktorých sa DNA vlákno neposunulo) a prechody modelujúce posun o viac ako jednu bázu. Pravdepodobnostné distribúcie hodnoty signálu pre jednotlivé k-mery poskytuje výrobca MinIONa [5]. Na základe tohto modelu sa Viterbiho algoritmom vypočíta najpravdepodobnejšia postupnosť báz, ktorá mohla vygenerovať pozorovaný signál.

Iné basecallery sú založené na rekurentných neurónových sieťach. Niektoré (napr. DeepNano [1]) pracujú so signálom rozdeleným na udalosti, iné pracujú s nerozdeleným signálom (napr. Chiron [6]).

Najlepšie súčasné basecallery majú pre jedno čítanie presnosť okolo 85% až 90%. Ak sa osekvenuje viac kópií rovnakej DNA, skombinovaním dostatočného počtu prekrývajúcich sa čítaní sa dá dosiahnuť presnosť okolo $99,9\%$ [7].

[2] z anglického *base calling* – určovanie báz.

1.2 Varianty v DNA

V prírode sa často vyskytujú dvojice DNA molekúl, ktoré obsahujú veľmi podobnú postupnosť báz, líšiacu sa iba v malých detailoch. Typickým príkladom sú DNA dvoch rôznych jedincov rovnakého druhu. Tieto malé odlišnosti v DNA voláme *varianty*. Najjednoduchšie druhy variantov sú nasledovné.

Jednonukleotidový polymorfizmus (SNP[3]). Jedna báza z prvej DNA postupnosti sa v druhej postupnosti nahradí inou bázou.

```
ACCACTGGACTTTCGA
ACCACTGCACTTTCGA
```

Inzercia. Do postupnosti je vsunutá skupina báz.

```
ACCACTG GACTTTCGA
ACCACTGATGACTTTCGA
```

Delécia. Z postupnosti vypadne súvislá skupina báz.

```
ACCACTGGACTTTCGA
ACCACT ACTTTCGA
```

[3] z anglického *single nucleotide polymorphism*

Kapitola 2

Ciele práce

Pri niektorých využitiach DNA sekvenovania sa sekvenuje vzorka, o ktorej je známe, že by sa mala podobať na inú, už osekvenovanú DNA. Cieľom sekvenovania je potom zistiť, ako sa tieto dve DNA postupnosti líšia. Jedným z takýchto využití je napríklad zisťovanie rezistencie baktérií na antibiotiká [2].

V našej práci sa budeme zaoberať nasledujúcim scenárom. Máme nejakú známu postupnosť dusíkatých báz, ktorú budeme nazývať *referencia*. Ďalej máme vzorku DNA, o ktorej vieme, že sa od referencie líši len veľmi málo. Táto vzorka bola spracovaná MinIONom, máme teda k dispozícii nameraný surový signál z jednotlivých čítaní. Naším cieľom je identifikovať varianty v sekvenovanej vzorke vzhľadom na referenciu. Ideálne by bolo vedieť s dobrou presnosťou určovať varianty už z jedného čítania.

Snažíme sa teda navrhnúť algoritmus s nasledovným vstupom a výstupom (neformálne):

Vstup: referenčná postupnosť dusíkatých báz,

postupnosť nameraných hodnôt surového signálu,

nepovinné: odhad očakávaného množstva variantov

Výstup: popis nájdených variantov (pozícia, typ, skóre)

Náš algoritmus bude nevyhnutne robiť chyby. V niektorých prípadoch nezvládne nájsť variant, ktorý vzorka obsahovala (falošné odmietnutie), v iných prípadoch nájde variant, ktorý neexistuje (falošné prijatie). V niektorých aplikáciách môže byť cena za chyby jedného druhu väčšia, než cena za chyby opačného druhu. Algoritmus preto vráti ku každému z nájdených variantov aj skóre, indikujúce istotu algoritmu, že naozaj ide o variant. Znižovaním minimálneho skóre, ktoré budeme vyžadovať, aby sme nájdený variant považovali za skutočný, bude možné znížit množstvo falošných odmietnutí za cenu zvýšenia množstva falošných prijatí, a obrátene.

2.1 Rozdiel oproti určovaniu báz

Jedným z možných riešení nášho problému je určiť zo signálu bázy pomocou niektorého z existujúcich basecallerov a následne už len zisťovať odlišnosti dvoch postupností báz. Problémom tohoto riešenia je nízka presnosť (ak nemáme veľa prekrývajúcich sa čítaní).

Pri tomto prístupe však basecaller vôbec nevyužíva fakt, že sekvenovaná postupnosť sa podobá na referenciu. Uvažuje teda podstatne väčší priestor možných výsledných sekvencií, než je nutné. V dôsledku toho upredňostňuje sekvencie, ktoré lepšie vysvetľujú pozorovaný signál, aj keď môžu byť výrazne vzdialené od referencie. Zmenšenie priestoru uvažovaných sekvencií môže navyše znížiť výpočtovú náročnosť určovania báz, prípadne umožniť použitie presnejších techník, ktoré by za normálnych okolností boli príliš časovo náročné.

Kapitola 3

Zarovnanie signálu k referencii

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

Kapitola 4

Identifikácia jednonukleotidových polymorfizmov

V tejto časti navrhneme techniku na odhaľovanie jednonukleotidových polymorfizmov v sekvenovaných dátach.

Kapitola 5

Identifikácia inzercií a delécií

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

Kapitola 6

Testovanie

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

Kapitola 7

Implementácia

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

Záver

Na záver zhrnieme výsledky, ktoré sa nám podarilo dosiahnuť.

Literatúra

[1] Vladimír Boža, Broňa Brejová, and Tomáš Vinař. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLOS ONE*, 12(6):1–13, 06 2017.

[2] Phelim Bradley, N. Claire Gordon, Timothy M. Walker, Laura Dunn, Simon Heys, Bill Huang, Sarah Earle, Louise J. Pankhurst, Luke Anson, Mariateresa de Cesare, Paolo Piazza, Antonina A. Votintseva, Tanya Golubchik, Daniel J. Wilson, David H. Wyllie, Roland Diel, Stefan Niemann, Silke Feuerriegel, Thomas A. Kohl, Nazir Ismail, Shaheed V. Omar, E. Grace Smith, David Buck, Gil McVean, A. Sarah Walker, Tim E. A. Peto, Derrick W. Crook, and Zamin Iqbal. Rapid antibiotic-resistance predictions from genome sequence data for staphylococcus aureus and mycobacterium tuberculosis. *Nat Commun*, 6:10063, Dec 2015. 26686880[pmid].

[3] Matei David, L. J. Dursi, Delia Yao, Paul C. Boutros, and Jared T. Simpson. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. *Bioinformatics*, 33(1):49–55, 2017.

[4] 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.

[5] Oxford Nanopore Technologies. kmer_models. `https://github.com/nanoporetech/kmer_models`.

[6] Haotian Teng, Minh Duc Cao, Michael B. Hall, Tania Duarte, Sheng Wang, and Lachlan Coin. Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning. *bioRxiv*, 2017.

[7] Ryan Wick. A comparison of different Oxford Nanopore basecallers. `https://github.com/rrwick/Basecalling-comparison`.