METHODS OF PROCESSING OXFORD NANOPORE SEQUENCING DATA FOR METAGENOMICS

Lujza Barilíková

Bachelor Degree Programme (3), FEEC BUT

E-mail: xbaril02@stud.feec.vutbr.cz

Supervised by: Kristýna Kupková

E-mail: kupkova@feec.vutbr.cz

Abstract: The presented paper describes a new method of processing data produced by revolutionary sequencing technology introduced by Oxford Nanopore Technologies – MinION, which holds a great promise in the field of metagenomics. Low cost, produced long reads and portability, due to its small dimensions, represents only one of the many advantages of this technology. Despite of the benefits, there is a lack of available computational tools for handling the produced data and that is the reason, why a new method of processing such data should be created. In this study such method is created based on dimensionality reduction for data visualization.

Keywords: metagenomics, Oxford Nanopore, nanopore sequencing, dimensionality reduction

1 ÚVOD

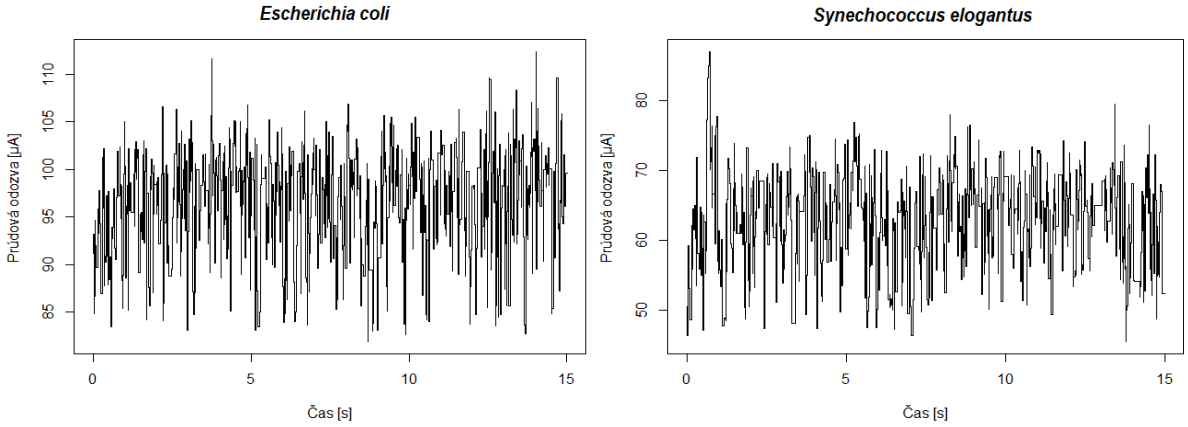
Mikroorganizmy sú neoddeliteľnou súčasťou nášho života, preto sa v snahe preskúmať komunity mikroorganizmov postupom času začali vyvíjať sekvenačné metódy a technológie študujúce ich genetickú informáciu. Vďaka obrovskému rozvoju dostupných metód sekvenovania sa problém množstva produkovaných dát stal minulosťou. V súčasnosti sa preto rozvinulo úsilie tieto dáta spracovať, čo predstavuje vývoj nových algoritmov slúžiacich k ich spoľahlivej analýze a klasifikácii, čomu sa venuje aj tento príspevok.

2 SEKVENAČNÉ DÁTA ZÍSKANÉ NANOPÓROVÝM SEKVENOVANÍM

Snahou spoločnosti Oxford Nanopore Technologies je proces sekvenovania zjednodušiť na úroveň, kedy využitím ich technológií bude možné kýmkoľvek analyzovať akékoľvek zvolené vzorky v príslušnom prostredí, či už v laboratóriu alebo aj v teréne. Sekvenované môžu byť extrémne dlhé fragmenty vo veľmi vysokej kvalite, pretože v procese dochádza k detekcii jednotlivých molekúl a vynechaný je krok predchádzajúcej amplifikácie, ktorá býva častým zdrojom skreslenia signálu. Základom nanopórového sekvenovania, ktoré je poskytované zariadením MinION, je priama elektrická detekcia jednovláknovej DNA, ktorá prichádza do kontaktu s nanopórom začleneným do nevodivej membrány, ku ktorej je priložené napätie indukujúce v nanopóre prúd. Pri prechode molekuly prostredníctvom póru dochádza k zmenám tohto prúdu, ktoré sú charakteristické pre jednotlivé bázy [1]. Prúd v nanopóre je meraný pomocou senzora niekoľko tisíc krát za sekundu a následné produkované dáta sa líšia predovšetkým v závislosti od použitej chémie. V súčasnosti sú vyvinuté dve významné chémie používané v prietokových komôrkach (tzv. flowcell) zariadenia. Jedná sa o chémiu R7 a R9, ktoré sa odlišujú v použití konkrétneho póru a od toho sa odvíjajúcich parametrov. V tomto článku sú zahrnuté dáta produkované predovšetkým chémiou R7, keďže sú aktuálne najviac dostupné. Konečným výstupom zariadenia MinION je súbor FAST5 z každého čítania. Podobne ako štandardný dátový formát HDF5, aj FAST5 súborový formát je založený na hierarchickom usporiadaní, v ktorom sú v prípade R7 chémie skladované metadáta odpovedajúce čítaniu produkovanému jedným z 512-tich kanálov prietokovej komôrky zariadenia a taktiež udalosti (tzv. events), obsahujúce sekvenátorom predspracované dáta o meranom prúde [2].

2.1 SPRACOVANIE DÁT POMOCOU VIZUALIZAČNÝCH METÓD

Cieľom príspevku je vytvorenie vhodného algoritmu pre tzv. binning (roztriedenie) metagenomických dát, získaných spomínanou sekvenačnou technológiou. Výstupom sekvenácie sú však krátke úseky DNA, u ktorých nie je známe z akého organizmu pochádzajú. Z toho dôvodu je tu snaha vytvoriť algoritmus, ktorý bude schopný uskutočniť binning dát na základe vektoru príznakov, odvodených z každej sekvencie, za použitia vhodného zhlukovacieho algoritmu. Prvým krokom je načítanie dát s dopracovaním sa k surovým dátam meraného prúdu, ktoré sú spracované a konvertované do sekvencie jednotlivých udalostí (tzv. events). Tie obsahujú informácie o priemernej hodnote prúdu, príslušnej odchýlke a dĺžke trvania, čo predstavuje určitú nevýhodu tohto formátu, pretože dochádza k strate signálu nášho záujmu, ktorý je potrebné následne rekonštruovať. Rekonštrukcia signálu vychádza z dopočítania počtu vzoriek pripadajúcich na každú priemernú hodnotu vyskytujúcu sa v zložke udalostí. Ukážku takto zrekonštruovaných signálov je možné vidieť na Obrázok 1:.



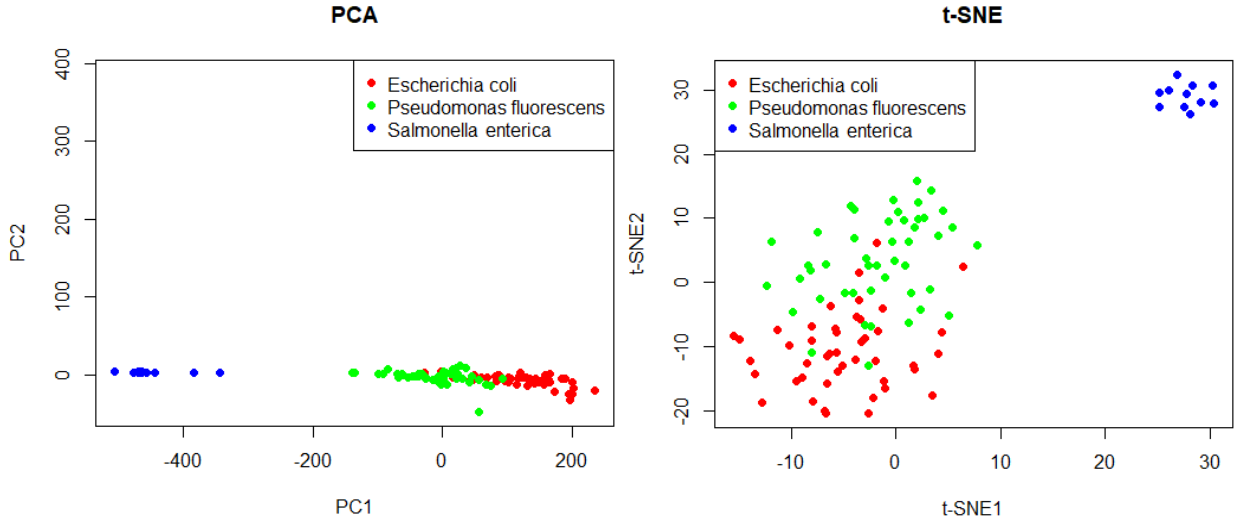
Obrázok 1: Ukážka zrekonštruovaných signálov pochádzajúcich zo sekvencií *Escherichia coli* a *Synechococcus elogantus* získaných nanopórovou technológiou

Je známe, že sekvencie pozostávajúce z postupnosti jednotlivých nukleotidov sú charakterizované viac ako troma parametrami, to znamená, že nie je vždy jednoduché ich spracovávať a následne zrozumiteľne a ľudsky prezentovať. Obrovskou výhodou takto vzniknutej signálovej reprezentácie sekvencií, oproti znakovej podobe, je možnosť použiť na tieto dáta už existujúce metódy slúžiace k spracovávaniu signálov. Z toho dôvodu je v tomto príspevku predstavený nový prístup spracovania sekvenačných dát využívajúci aplikáciu metód redukcie dimenzií priamo na signálovú reprezentáciu. Vo všeobecnosti ide o získanie vektorov signálových sekvencií, ktoré sú následne zobrazené v nízko-dimenzionálnom priestore. Využitím metód redukcie dimenzií získavame prístup vizualizovať obrovské súbory dát v snahe zachovať ich pôvodnú informáciu.

3 HODNOTENIE VÝSLEDKOV

Po získaní matice dát boli v súčasnej dobe k vizualizácii použité metódy PCA a t-SNE, obe vychádzajúce z princípu redukcie dimenzií. K spracovaniu bol použitý simulovaný metagenomický dataset, pozostávajúci z dát získaných z archívu ENA (European Nucleotide Archive), v ktorom sú zastúpené tri odlišné organizmy, konkrétne *Escherichia coli, Pseudomonas fluorescens* v archíve pod prístupovým číslom PRJEB8716, a taktiež *Salmonella enterica* s prístupovým číslom PRJEB7205. Dáta sa podarilo vizualizovať, kde každý bod v grafe reprezentuje jednu konkrétnu sekvenciu, na základe čoho je možné pozorovať skryté vzťahy medzi analyzovanými objektami.

Z grafu na Obrázok 2: môžeme vidieť, že zhluk predstavujúci *Salmonella enterica* sa v oboch prípadoch separuje od zvyšných, vďaka čomu by bolo teoreticky možné ju detegovať v prípade jej prítomnosti v reálnom metagenomickom datasete.



Obrázok 2: Ukážka vizualizácií PCA a t-SNE simulovaného metagenomického datasetu

Po jej odfiltrovaní by sme sa ďalej mohli zaoberať zvyšnými zhlukmi, ktoré už nie sú natoľko vzájomne separované, no napriek tomu, sme schopní ich určitým spôsobom rozlíšiť. V ďalšom prípade by bolo možné použiť iné algoritmy k rozdeleniu týchto organizmov do kategórií na základe ich príslušnosti k taxónu. Takýmto prístupom by sme sa mohli dopracovať k identifikácii všetkých prítomných taxonomických jednotiek. Rovnako by bolo vhodné preskúmať efektívnosť týchto metód na dáta poskytované novšou chémiou R9.

4 ZÁVER

V danom príspevku je predstavená metóda spracovania surových dát poskytnutých nanopórovým sekvenovaním, ktorá by mala umožniť efektívnu klasifikáciu, vrátane vizualizácie metagenomických datasetov. Na základe doposiaľ získaných výsledkov je možné vyvodiť záver, že použité algoritmy a predspracovanie dát poskytujú dostatočne jasné rozlíšenie sekvencií, pochádzajúcich z troch rôznych organizmov. Na základe tohto prístupu by jasným diferencovaním zhodných sekvencií blízko príbuzných taxónov mohlo byť poskytnuté lepšie pochopenie komunít a taktiež súborov komunít, ktoré vytvárajú biosféru ako nejaký vnorený systém v systéme, ktorého človek je súčasťou a od ktorého prežitie ľudí závisí. Na poli metagenomiky by takýto algoritmus mohol predstavovať významný prelom v oblasti klasifikácie sekvencií získaných predovšetkým technológiou nanopórového sekvenovania

REFERENCIE

[1] IP, Camilla L.C., Matthew LOOSE, John R. TYSON, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research [online]. , - [cit. 2017-11-10]. DOI: 10.12688/f1000research.7201.1. ISSN 2046-1402.

[2] LU, Hengyun, Francesca GIORDANO a Zemin NING. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics & Bioinformatics [online]. 2016, 14(5), 265-279 [cit. 2017-11-10]. DOI: 10.1016/j.gpb.2016.05.004. ISSN 16720229.