VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

PŘÍMÁ KLASIFIKACE METAGENOMICKÝCH SIGNÁLŮ ZE SEKVENACE NANOPÓREM

DIRECT BINNING OF METAGENOMIC SIGNALS FROM NANOPORE SEQUENCING

SEMESTRÁLNÍ PRÁCE
SEMESTRAL THESIS

AUTOR PRÁCE
AUTHOR

Bc. Marko Lebó

VEDOUCÍ PRÁCE
SUPERVISOR

Mgr. Ing. Karel Sedlář

BRNO 2019

ABSTRAKT

Táto diplomová práca sa venuje klasifikačným metódam metagenomických signálov, nezávislých na referenčnej databáze, ktoré sú získané sekvenáciou pomocou sekvenátora MinION. Popisuje ako vznik a druh dát, s ktorými sa v metagenomike pracuje, tak už existujúce metódy klasifikácie metagenomických dát nezávislých na referenčnej databáze. Rovnako je v tejto práci popísané, aký veľký prielom v metagenomike znamenalo objavenie sekvenácie DNA tretej generácie a konkrétne sa táto práca špecializuje na funkciu sekvenátora MinION od spoločnosti Oxford Nanopore. Cieľom práce bolo na základe modelových dát metagenómu navrhnúť a zrealizovať vlastnú metódu klasifikácie metagenomických dát získaných zo sekvenátora MinION a na záver ju porovnať s už existujúcou metódou klasifikácie.

KLÍČOVÁ SLOVA

metagenomika; sekvenácia tretej generácie; Oxford nanopore; klasifikácia metagenomických dát; squiggle

ABSTRACT

This diploma thesis deals with taxonomy independent methods for classification of metagenomic signals, aquired by a MinION sequencer. It describes the formation and character of metagenomic data and already existing methods of metagenomic data classification and their development. This thesis also evaluates an impact of the third generation sequencing techniques in the world of metagenomics and further specialises on the function of the Oxford Nanopore MinION sequencing device. Lastly, a custom method for metagenomic data classification, based on data obtained from a MinION sequencer, is proposed and compared with an already existing method of classification.

KEYWORDS

metagenomics; third generation sequencing; Oxford nanopore; metagenomic data clasification; squiggle

LEBÓ, M. *Přímá klasifikace metagenomických signálů ze sekvenace nanopórem* . Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2019. 47 s. Vedoucí diplomové práce Mgr. Ing. Karel Sedlář, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma Přímá klasifikace metagenomických signálů ze sekvenace nanopórem jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a~jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

(podpis autora)

PODĚKOVÁNÍ

Rád by som poďakoval svôjmu vedúcemu práce Dr. Karel Sedlář. za jeho odborné rady a vedenie mojej práce aj napriek jeho pobytu mimo územie ČR, za jeho ústretový prístup pri organizácií konzultácií a predovšetkým za jeho odborné vedenie a rady poskytnuté pri tvorbe tejto diplomovej práce.

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

S pokrokom technológie sekvenácie, ktorá naberá na rýchlosti a presnosti, pri klesajúcej cene na jednu nasekvenovanú bázu, sa umožnil zrýchlený rozvoj v oblasti vedného oboru metagenomiky.

Cieľom tejto diplomovej práce preto je najskôr ozrejmiť čitateľovi, čo je metagenomika a prečo nachádza tak široké využitie, popísať druh dát, s ktorým sa v tejto oblasti vedy pracuje a ako vznikajú a aké sú hlavné problémy klasifikácie týchto dát. Ďalej sú v nej predstavené rôzne, už existujúce, metódy klasifikácie spracovania metagenomických dát a ich vývoj. Keďže sa práca sústredí na prácu so sekvenátorom MinION od spoločnosti Oxford Nanopore, je potrebné všeobecne uviesť vývoj sekvenácie tretej generácie, na ktorého konci je práve sekvenátor MinION a rovnako podrobne popísať spôsob jeho funkcie.

Podstatná časť tejto práce je venovaná návrhu a realizácií vlastnej metódy klasifikácie metagenomických signálov, ktorá používa nový prístup analýzy metagenomických signálov, použitím Poincarého máp. Jedná sa o nový prístup, ktorý doteraz nebol na tieto účely nikdy použitý a zároveň sa prejavil ako pomerne efektívny a výpočtovo nenáročný. Poincarého deskriptory sú v súčasnej dobe používané hlavne pri analýze variability srdečného rytmu, ale v tejto práci boli tie isté deskriptory, s minimálnou úpravou, aplikované na výstupné dáta sekvenátora MinION za účelom klasifikácie dát získaných zo sekvenácie modelového metagenómu. Po výpočte Poincarého deskriptorov bolo potrebné získané dáta zaradiť do zhlukov, na čo bolo potrebné vybrať správnu metódu zhlukovania vzhľadom na povahu získaných dát, kedy boli testované zhlukovacie metódy k-means a DBSCAN.

Na záver bola navrhnutá metóda porovnaná s už existujúcou metódou klasifikácie metagenomických dát - VizBin.

1 METAGENOMIKA V PRAXI

Metagenomika vznikla v dôsledku zistenia, že veľké množstvo mikroorganizmov nie je možné kultivovať v laboratórnych podmienkach klasickým spôsobom, keďže tento proces nie sú schopné prežiť [1]. Jedná sa o metódu analýzy komunít mikroorganizmov žijúcich v rôznych prostediach, využitím extrakcie DNA a nasekvenovania celých genómov organizmov, z ktorých sa dané komunity tohto prostredia skladajú.

Tento prístup umožnil objavenie nových druhových línií mikroorganizmov, ktoré dovtedy neboli známe práve v dôsledku ich straty pri snahe o izoláciu mikroorganizmov pri kultivácií [2] a navyše sa preukázalo, že komunity mikroorganizmov sa správajú vo svojom prirodzenom prostedí úplne inak ako po ich izolácií. Preto sa pomocou metagenomiky začali skúmať mikrobiálne komunity a ich funkcie v rôznych prostrediach. Ako jedny z najdôležitejších komunít, vplývajúcich priamo na zdravie človeka, sú mikrobiálne komunity čreva človeka, ktoré priamo oplyvňujú schopnosti trávenia a ich zloženie môže priamo súvisieť napríklad s výťažkom množstva energie z potravy, čo môže mať priamy súvis napríklad s obezitou [3], ich diverzita môže priamo súvisieť so zápalovými ochoreniami čriev [4] a nedávne štúdie dokonca potvrdili možnosť regulácie množstva železa v krvi, pomocou regulácie baktérie *E. coli* v mikrobiálnej komunite čreva, ktorá môže podporiť efektivitu vstrebávania železa z potravy do organizmu [5].

Všetky tieto štúdie teda dokazujú, aké môže byť štúdium zloženia bakteriálnych komunít v ľudskom čreve prínosné vo svete medicíny, no mikrobiálne komunity majú veľký význam aj v oblastiach rastlinnej biológie a v oblasti poľnohospodárstva, keďže sú schopné priamo ovplyvniť množstvo výživných látok v pôde a regulovať tak úrodnosť pôdy [6]. Mikróby zohrávajú svoju úlohu napríklad aj vo vodách morí, kedy sú schopné ovplyvniť hydrotermálne procesy vody, ktoré majú priamy vplyv ako na ekosystém žijúci v týchto vodách, tak aj na meteorologické procesy našej planéty [7].

Dôležitosť metagenomiky je teda veľmi veľká vo všetkých oblastiach vedy a jej napredovanie sa priamo odvíja od naších možností generovať metagenomické dáta cez sekvenáciu celých genómov organizmov, z ktorých sa skladajú jednotlivé mikrobiálne komunity a rovnako od schopnosti spracovávať tak obrovské množstvo dát, ktoré

sekvenácia generuje, aby sme potom boli schopní identifikovať organizmy, ktorým jednotlivé nasekvenované genómy patria.

Keďže sa v mikrobiálnych komunitách stretávame s veľkou diverzitou druhov v rámci jednej vzroky, znamená to vysoké množstvo genómov, ktoré musia byť nasekvenované a teda vysokú náročnosť na sekvenačné metódy. Spôsob sekvenácie, ktorý sa v metagenomike využíva je metóda shotgun sekvenovania, kedy sú jednotlivé genómy rozdelené na menšie časti, ktorých dĺžka záleží na ďalej použitej metóde sekvenácie a až tieto časti genómov sú nasekvenované [9].

Výstupom shotgun sekvenácie je teda veľmi veľký počet nasekvenovaných častí genómov, ktoré nie sú priradené k žiadnemu genómu zo skupiny genómov nachádzajúcich sa vo vzorke. Toto spôsobilo rozvoj metód klasifikácie metagenomických dát, ktoré sa všeobecne rozdeľujú do dvoch skupín – možnosti klasifikácie dát závislých a nezávislých na referenčnej databáze, ktorých účelom je priradiť jednotlivé časti genómov k daným genómom, rekonštuovať tieto genómy a podľa toho identifikovať organizmy, ktoré sa nachádzajú v analyzovanej vzorke. [9]

Alternatíva k celogenómovej sekvenácií, ktorá sa snažila vyhnúť sekvenovaniu celých genómov a obísť všetky problémy s tým spojené, bola metóda, ktorá sa zaraďuje do pseudo metagenomiky, kedy sa jednotlivé organizmy vo vzorke klasifikovali pomocou sekvenácie len konkrétnych fylogenetických génov (16S rRNA, rboP, atd.), ktoré slúžili ako identifikátory daných organizmov [10]. Toto uľahčilo a urýchlilo rozvoj metód klasifikácie metagenomických dát, pretože náhle nebolo potrebné sekvenovať celé genómy organizmov, ale stačili len jednotlivé časti genómov, ktoré zodpovedali daným génom. To spôsobilo obrovský nárast v počte identifikovaných organizmov, ktoré sa aktuálne nachádzajú v niekoľkých databázach. Pre predstavu, k roku 2012 boli na svete 3 databáze sekvencií génu 16S rRNA - Greengenes, SILVA a EzTaxon-e, ktoré obsahovali dohromady asi 1 700 122 bakteriálnych 16S rRNA sekvencií. [11]

2 MOŽNOSTI KLASIFIKÁCIE METAGENOMICKÝCH DÁT

Ako bolo už v predchádzajúcej kapitole načrtnuté, možnosti klasifikácie shotgun metagenomických dát sa všeobecne rozdeľujú na dve skupiny – možnosti klasifikácie závislé a nezávislé na referenčnej databáze. Klasifikácia shotgun metagenomických dát závislá na referenčnej databáze je metóda klasifikácie, ktorá sa pri vyhodnocovaní príslušnosti častí genómov opiera o porovnávanie častí genómov s genómami uloženými v referenčnej databáze, za pomoci rôznych metód zarovnávania sekvencií ako sú BLAST, BLAT, Bowtie, BWA, využíva rôzne Markovské modely, pozerá sa na daný problém z pohľadu obsahu GC párových báz v sekvencií, alebo výskytu jednotlivých oligonukleotidových vzorov v danom genóme a na základe zhody s jej zodpovedajúcemu genómu v databáze [8]. Takýto prístup má dva veľké nedostatky, ktorými sú vysoká výpočtová náročnosť a obmedzená veľkosť referenčnej databáze. Podľa zdroju [12] sa predpokladá, že v prírode je približne 10^{30} druhov prokaryotických organizmov, pri čom v databáze NCBI sa k roku 2018 nachádzalo len 11 829 kompletne nasekvenovaných genómov prokaryotických organizmov, čo je veľmi malý zlomok odhadovaného počtu prokaryotických organizmov. Preto metóda klasifikácie metagenomických dát, ktorá je závislá na referenčnej databáze, bude mať len veľmi obmedzené využitie, keďže takýto systém nebude schopný klasifikovať veľa častí genómov, prípadne získa veľa falošne pozitívnych výsledkov. Z tohto dôvodu sa v tejto práci zaoberáme hlavne metódou klasifikácie metagenomických dát, ktoré sú nezávislé na referenčnej databáze a teda sú v tomto ohľade všestrannejšie, keďže sa ich výsledky nemusia opierať o žiadnu referenčnú databázu.

2.1 Metódy klasifikácie nezávislé na referenčnej databáze

Oproti klasifikácií metagenomických dát za použitia referenčnej databáze, sa klasifikácia nezávislá na referenčnej databáze opiera o princíp, kedy táto metóda porovnáva sekvencie proti sebe navzájom, podľa čoho vytvára distančnú maticu, pre následné zhlukovanie jednotlivých častí genómov do zhlukov, alebo inak povedané do taxónov, ktoré reprezentujú jednotlivé genómy organizmov, nachádzajúcich sa vo vyšetrovanej vzorke [13]. Takáto metóda nie je odkázaná na nijakú referenčnú databázu a preto je schopná identifikovať aj prítomnosť dovtedy nepoznaných druhov organizmov vo vyšetrovanej vzorke.

Metódy klasifikácie shotgun metagenomických dát, ktoré sú nezávislé na referenčnej databáze, môžu byť rozdelené do troch skupín: metódy založené na kompozícií, hojnosti zatúpenia a hybridné metódy.

Metódy založené na kompozícií

Tieto metódy sú založené na predpoklade, že každý taxón sa od ostatných líši svojou genómovou kompozíciou a preto sú založené na porovnávaní obsahu jednotlivých nasekvenovaných úsekov genómov. Prvým krokom takejto analýzy je najskôr prevedenie znakových sekvencií na číselný vektor. [8]

Najpoužívanejšou metódou prevodu je normalizovaná frekvencia výskytu k-mérov o istej veľkosti [13]. Takýto výpočet sa opiera o najjednoduchší možný model genómu, kedy sa predpokladá, že genóm môže byť reprezentovaný ako jediná pravdepodobnost výskytu p_A, ktorá značí výskyt báze A na danom mieste v genóme. Potom podľa komplementarity báze vyplýva, že $p_A=p_T$ a ďalej $p_C=1/2 - p_A=p_G$. O niečo pokročilejší model potom je, že jeden genóm sa skladá z istého množstva k-mérov. Ak je teda $k=1$, tak je nukleotid na danej pozícií v genóme nezávislý od predchádzajúceho nukleotidu. Ak je $k=2$, tak je genóm náhodným súborom dimérov a tak ďalej. [15]

Obyčajne sa na výpočet využívajú tetra-méry, ktoré sa ukázali jako optimálne [16]. Takýmto najjednoduchším nástrojom, ktorý využíva porovnávanie tetera-mérov, je nástroj TETRA, ktorý využíva korelácie medzi všetkými pármi nasekvenovaných častí genómov a tie potom podľa tohoto parametra rozdeľuje k jednotlivým taxónom. Tento nástroj je jeden z prvých nástrojov klasifikácie metagenomických dát a z dôvodu jeho

vysokej výpočtovej náročnosti sa dnes už nepoužíva a nie je ani dostupný.

Ďalším nástojom, ktorý využíva porovnávanie kompozícií nasekvenovaných čítaní je LikelyBin, ktorý využíva prístup Markovových reťazcov Monte Carlo na efektívne zníženie dimenzionality priestoru a vypočítanie najpravdepodobnejšej kombinácie nasekvenovaných čítaní, ktoré tvoria jeden genóm. Tento nástroj má ale len obmedzený počet genómov, ktoré sa môžu nachádzať v súbore analyzovaných genómov a to od 2 do 10 genómov, pri minimálnej dĺžke čítaní 400 bp, s 90% presnosťou [15].

Nástrojom, ktorý na svôj výpočet využíva opäť Markovove modely, je nástroj SCIMM (Sequence Clustering with Interpolated Markov Models) [17]. Tento nástroj využíva Interpolované Markovove modely, aby určil pravdepodobnosť výskytu nukleotidu na danej pozícií v genóme, na základe *w* nukleotidov, ktoré sa nachádzajú v sekvencií pred ním. Takýmto spôsobom potom pomocou modifikovanej verzie k-means rozdeľuje jednotlivé čítania do zhukov, ktoré figurujú ako genómy. Táto metóda má ale nevýhodu v tom, že metóda k-means potrebuje na svôj výpočet daný počet zhlukov, do ktorých má jednotlivé čítania zaraďovať, čo musí byť uvedené buď ako vstupný parameter, alebo musí byť pred túto metódu zaradená iná metóda, ktorá určí aspoň približný počet zhukov, s ktorými potom metóda SCIMM má pracovať. Toto výrazne znižuje presnosť a použiteľnosť tejto metódy, keďže je priamo závislá na týchto parametroch.

Na riešenie problému zhlukovania sa neskôr začali využívať umelé neurónové siete a to kontrétne SOM (self-organizing maps) [18], ktoré boli adaptované na účely genómovej informatiky v nástroji BLSOM (Batch-Learning Self Organising Maps). Samoorganizujúce sa mapy sa ukázali ako vhodný nástroj na zhlukovanie a vizualizáciu vysokodimenzionálnych dát na dvojdimenzionálnu mapu, ale ich veľkou nevýhodou je vysoká náročnosť finálneho zhlukovania a vysoká časová náročnosť výpočtu.

Ďalším podobným nástrojom je nástroj VizBin, ktorý opäť využíva na vizualizáciu mnohodimenzionálnych dát do dvojdimenzionálneho priestoru podobný princíp ako vyššie spomínané SOM a to síce BH-SNE (Barnes-hut stochastic neighbor embedding) [19]. Tento nástroj vyžaduje ako vstupné dáta množstvo parametrov ako dĺžka sekvencií, pokrytie, percentuálne zloženie GC párov, popisky jednotlivých zhukov a sekvencie, ktoré obsahujú niektorý zo značkovacích génov ako napríklad gén 16S

rRNA. Po zadaní vstupných parametrov sa spustí výpočet BH-SNE, ktorý na výstupe zobrazí reprezentáciu jednotlivých čítaní v 2D priestore, ktoré v ideálnom prípade vykazujú tendencie zhlukovania sa. Potom je to už na užívateľovi, aby presne vymedzil hranice jednotlivých zhlukov v 2D mape. Výhodou tohto nástroja je možnosť získania čítaní jednotlivých zhlukov znovu v jednom FASTA súbore.

Nástroj 2Tbinning (2-tier binning) zvolil prístup, kedy zaraďuje sekvenované ready do zhlukov podľa dvoch parametrov. Najskôr vypočíta zhlukovanie podľa parametru OFDEG (oligonucleotide frequency derived error) a potom tento nástroj urobí druhé zhlukovanie pomocou obsahu GC párov báz v čítaniach a výsledné zhluky sú priemerom týchto dvoch metód [20]. Nástrojom 2Tbinning bol zavedený nový parameter, ktorý nebol doteraz obsiahnutý v žiadnom z predschádzajucich nástrojov a to parameter OFDEF. Tento parameter sleduje odchýlku v oligonukleotidovej frekvencií celého genómu, ktorý sa rekonštruuje a jeho výsledná hodnota je pokles tejto odchýlky v priebehu postupného pridávania čítaní k danému genómu. Podľa tejto hodnoty potom sú jednotlivé čítania zaradené do zhlukov a priradené jednotlivým taxónom [21].

Metódy založené na hojnosti zastúpenia

Metódy založené na kompozícií majú jeden veľký nedostatok a zlyhávajú v prípade, keď sa v súbore genómov nachádza genóm s menšou hojnosťou zatúpenia čítaní. Jeho čítania sú preto často falošne priradené do zhlukov, do ktorých nepatria a preto hrozí zlyhanie detekcie takýchto genómov. Navyše druhým problémom tejto metódy je, že so znižujúcou sa dĺžkou čítaní, sa znižuje presnosť metódy, pretože je k dispozícií nedostatočný počet párov báz, kvôli príliž vysokej variabilite krátkych úsekov v sekvencií DNA.

Ako odpoveď na tento problém vznikli metódy založené na hojnosti zastúpenia k-mérov v čítaniach, miesto porovnávania podobnosti sekvencií. Takéto metódy umožňujú potom presne klasifikovať aj čítania o menších dĺžkach.[8]

Jedným z takýchto nástrojov je AbundanceBin, ktorý dokáže presne klasifikovať metagenomické dáta, u ktorých je dĺžka čítania väčšia ako 75 bp [22]. Nástroj AbundanceBin je postavený na predpoklade, že distribúcia nasekvenovaných čítaní sa riadi Lander-Watermanovým modelom, ktorý vypočítava pokrytie každého nukleotidu pomocou Poissonovho rozloženia. Každý jeden genóm vníma ako samostatné

Poissonovo rozloženie, kedy hľadá hojnosť populácie všetkých možných l-tupletov (častí) čítaní pre jednotlivé genómy v súbore metagenomických dát, ktorých výskyt sa podľa predpokladov riadi samotným Poissonovým rozložením. Najprv teda algoritmus zistí všetky kombinácie l-tupletov, ktoré je možné získať z daných nasekvenovaných čítaní a v nasledujúcom kroku je potom úlohou algoritmu optimalizovať logaritmus pravdepodobnosti získania konkrétneho vektora, ktorý je daným čítaním. Tento optimalizačný proces sa riadi algoritmom maximalizovaného očakávania (Expectation – Maximization - EM). [23]

Nátroj, ktorý spadá do rovnakej kategórie nástrojov, je aj MBBC (Metagenomic Binning Based on Clustering), ktorý využíva, podobne ako AbundanceBin, ku klasifikovaniu čítaní algoritmus EM a to hneď v prvom kroku klasifikácie, kedy tento algoritmus používa na vytvorenie skupín čítaní, ktoré sú rozdelené podľa frekvencií výskytu k-mérov v jednotlivých čítaniach [24]. Predpoklad, na ktorom si zakladá tento algoritmus je, že frekvencie výskytov k-mérov v čítaniach, budú podobné pre rovnaký druh organizmu a teda čítania s podobnými hodnotami tohoto parametru budú zaradené do jednej skupiny. Následne po prvotnom rozdelení čítaní do skupín vypočíta nástroj MBBC Markovove parametre pre všetky čítania zo skupiny, pričom stále zachováva predchádzajúci predpoklad a teda počíta s možnosťou, že čítania v danej skupine pochádzajú od jedného druhu organizmu a teda vysoko pravdepodobne sú súčasťou toho istého Markovovho reťazca. Na záver algoritmus rozdelí ready do definitívnych zhlukov iteratívnym výpočtom ich Markovových vlastností v rovnakej skupine, až kým tento výpočet úplne nekonverguje.

Posledným nástrojom, ktorý si predstavíme v tejto kategórií, je MetaCluster 5.0, ktorý sa špecializuje na vyhodnocovanie zašumených metagenomických dát s obsahom druhov s nízkym zastúpením čítaní vo vzorke [25]. MetaCluster 5.0 chápe ako šum vo vzorke čítania druhov organizmov, s extrémne nískym zástúpením, teda hojnosťou v súbore čítaní, čo spôsobuje ich pravdepodobné chybné zaradenie ku druhom s vyšším zastúpením a teda zašumenie ich genómov čítaniami, ktoré ku nim nepatria.

Odstránenie takéhoto šumu prebieha hneď v prvom kroku algoritmu, kedy algoritmus na základe pozorovania frekvencie k-mérov jednotlivých čítaní rozhoduje, či dané čítanie pochádza z genómu druhu, ktorý má extrémne nízku hojnosť a ak tomu tak

je, tak je toto čítanie automaticky odstránené. Rozhodovanie je položené na predpoklade, že ak čítanie pochádza z takéhoto genómu, frekvencia jeho k-mérov bude v súbore čítaní extrémne nízka, pretože ako už bolo pred tým spomenuté, predpokladá sa, že čítania pochádzajúce z rovnakého genómu budú obsahovať podobné zloženie k-mérov a navyše sa predpokladá, že frekvencia k-mérov má lineárny vzťah s hojnosťou zastúpenia čítaní jednotlivých druhov organizmov [22]. Takto odfiltrované dáta majú aj tú výhodu, že ich veľkosť sa týmto zmenší a teda sa znižuje aj výpočetná náročnosť algorimu. Súčasťou prvého kroku je aj určovanie čítaní, ktoré naopak patria k druhom organizmov s vysokou hojnosťou zatúpenia. Toto určovanie je postavené na výskyte dlhých w-mérov vo väčšom množstve čítaní, ktoré sa nachádzajú len v čítaniach pochádzajúcich od organizmov s vysokou hojnosťou.

Po identifikácií čítaní patriacich k organizmom s vysokou a extrémne nízkou hojnosťou zatúpenia, algoritmus zhlukuje čítania od druhov organizmov s nízkou hojnosťou. Zhlukovanie takýchto čítaní prebieha ako uzavretie medzery medzi detekciou organizmov s vysokou a extrémne nízkou hojnosťou zastúpenia, kedy sa veľkosť w-mérov detekovaných pri organizmoch s vysokou hojnosťou zastúpenia, začne postupne zmenšovať a teda sa začnú detekovať čítania prislúchajúce k druhom organizmov s menšou a menšou hojnosťou zastúpenia. Veľkosť w-mérov sa zmenšuje až kým sa nedosiahne stavu, kedy sa dostaneme na dĺžku q, ktorá je ešte dostatočne dlhá na to, aby sme mohli čítania priradiť do jedného zhluku, ale nie je príliž nízka na to, aby boli do týchto zhlukov zaradené čítania, ktoré patria k organizmom s extrémne nízkou hojnosťou zastúpenia. V prípade, že sa dostaneme do tohoto stavu, sú prevdepodobne identifikované všetky zhluky a algoritmus sa ukončí.

Hybridné metódy

Ako už z názvu kategórie hybridních metód vyplýva, ide o metódy klasifikácie metagenomických dát, kedy je na tento účel použitých niekoľko metód zároveň, čím sa zvyšuje presnosť vytvárania zhlukov.

Takýmto nástrojom je napríklad nástroj CompostBin. Keďže pri použití tetramérov, s ktorými sme sa prvýkrát stretli u nástroja TETRA, nastáva vytvorenie euklidovského kompozičného priestoru o rozmere 4^4, teda 256 dimenzií, nastala potreba tento rozmer znížiť, čím by sa znížila aj výpočtová náročnosť daného nástroja. Túto úlohu splnil práve nástroj CompostBin, ktorého postup je taký, že vypočíta najskôr frekvenciu k-mérov rôznych dĺžok a potom PCA analýzou, ktorá je váhovaná prevrátenou hodnotou pokrytia sekvencie, zníži kompozičný priestor na možné minimum. Stále tu ale ostával jeden nedostatok, ktorý CompostBin zdielal spolu s nástrojom TETRA, ktorým je závislosť efektivity oboch nástrojov na dĺžke sekvenovaných častí genómu, ktoré museli byť dlhé aspoň 700 bp. [26]

Nástroj, ktorý je svojou konštrukciou a použitím metód klasifikácie čítaní podobný nástroju CompostBin, je CONCOCT [27]. Tento nástroj zlučuje využitie frekvencie k-mérov a výpočtu pokrytia sekvencie, za využitia neváhovanéj PCA analýzy na zníženie dimenzionality kompozičného priestoru. Charakteristikou tohoto nástroja je, že jeho presnosť je priamo úmerná s počtom čítaní, pri čom najnižší počet čítaní, pri ktorom je tento nástroj dostatočne presný, bol experimentálne určený na 50 čítaní. Na vytvorenie samotných zhlukov čítaní potom kombinuje CONCOCT Bayesovský prístup s GMM (Gaussian mixture model), čo z neho robí nástroj nezávislý na ľudskom vstupe.

Nástroj, ktorý využíva väčšie množstvo podnástrojov je MyCC [28]. MyCC využíva na vytvorenie zhlukov jednak detekovanie markerových génov v genómoch a tak isto aj metódy podobné nástroju VizBin, kedy miesto užívateľom stanovaných zhlukov, je na ten istý účel použitý algoritmus Affinity Propagation (AP) [29]. Workflow tohoto programu je, že najskôr pomocou nástroja FetchMG[30], ktorý z metagenomických dát získa 40 markerových génov pomocou využitia skrytých Markovových modelov, ktoré sa neskôr použijú na spresnenie vytvorených zhlukov. Ďalším krokom je vytvorenie prvotného odhadu zhlukov pomocou vizualizácie metagenomických dát metódou BH-SNE, ktorá bola prvýkrát využitá nástrojom VizBin a redukuje 2D dáta na 1D zobrazenie.

Dvojdimenzionalita dát spočíva vo vyhodnocovaní čítaní pomocou dvoch parametrov a to výskytu jednotlivých k-mérov a ich reverzných komplementov v konsenzuálnej sekvencií. Vzniknutá vizualizácia sa v nasledujúcom kroku podrobí samotnému zhlukovaniu, kedy zhluky sú vytvorené pomocou algoritmu AP. V ďalšom kroku sú potom použité detekované markerové gény, kedy na základe výskytu čítaní, ktoré kódujú tie isté gény, sú zlučované zhluky s tými istými génmi a zároveň nachádzajúce sa blízko pri sebe. Takto vytvorené zhluky sú potom finálnymi zhlukmi čítaní patriacich k jednému druhu organizmov.

Nástroj, ktorý kombinuje využitie ako metód založených na kompozícií, tak metód založených na hojnosti zastúpenia a dopĺňa ich využitím pravdepodobnostných metód, je nástroj MetaBAT. [31] Tento nástroj vypočítava pravdepodobnostnú maticu medzi jednotlivými sekvenciami za využitia frekvencie k-mérov a na základe hojnosti zastúpenia sekvencií čítaniami. Pravdepodobnostné vzdialenosti tetranukleotidových frekvencií v tejto matici sú vypočítané na základe empirického modelu, ktorý bol získaný na základe vnútro- a medzidruhovými vzdialenosťami už známich genómov. Samotné zhlukovanie čítaní potom prebieha za využitia algoritmu EM, ktorý bol prvýkrát spomenutý pri nástroji MBBC, pričom je na odhadnutie prvotného počtu zhlukov využitá detekcia markerových génov, ktoré sú pre každý genóm jedinečné.

Nástroj s výrazne odlišným workflow od tých predchádzajúcich, je naštroj GroopM. Vstupnými dátami tohoto nástroja sú už assemblované metagenomické dáta, ktoré sú výstupom niektorého assemblovacieho nástroja. GroopM potom vypočítava pokrytie a tetramérovú frekvenciu, ktorá zapříčiní vznik 136 dimenzionálneho vektora, kvôli čomu GroopM pre zníženie dimenzionality využíva len principiálne komponenty tohoto vektora. Tieto dáta sú potom ďalej transformované do trojdimenzionálneho priestoru pre vizualizáciu všetkých contigov, ktoré boli obsiahnuté vo vstupných dátach. Všetky contigy sú potom zoradené podľa ich pokrytia a ich frekvencie k-mérov a zo zostavy contigov je vyradený každý druhý contig. Tento krok sa opakuje až kým zostava contigov dosiahne počet 1500, z ktorých potom vytvorí maticu vzdialenosí na základe ich pokrytia a ku každému contigu sú priradené jeho čítania. Druhým krokom algoritmu je vytvorenie prvotne odhadovaných zhlukov, kedy sú zhlukované contigy s vysoko podobnými profilmi pokrytia, frekvenciami k-mérov a dĺžkami, pričom sú použité len contigy s dĺžkou väčšou jako 1500 bp. Tieto prvotné odhady zhlukov slúžia ako „semená“

zhlukov, ktoré sa v ďalšej etape výpočtu zvanej recruitment, rozrastajú a postupne formujú finálne zhluky na základe vzdialeností prepočítavaných z už spomínaných parametrov contigov. Výstupné dáta sú potom exportované do bežných formátov ako napríklad FASTA atd. [32]

2.2 Sekvenátory tretej generácie

Vysoký rozvoj metód spracovania metagenomických dát by ale nebol možný vďaka súčasnému rozvoju metód sekvenácie molekúl nukleových kyselín, keďže v prípade metagenomických dát spracovávame genetické informácie pochádzajúce z niekoľkých celých genómov, ktoré musia byť najskôr nasekvenované.

Veľký pokrok v tomto smere znamenal vývoj sekvenácie tretej generácie, ktorá dokáže sekvenovať obrovské množstvo genetickej informácie, pochádzajúcej z niekoľkých úsekov tej istej molekuly naraz a preto potrebuje výrazne nižší čas na nasekvenovanie jednej molekuly nukleovej kyseliny ako metódy predchádzajúcich generácií [32]. Sekvenátory druhej generácie využívali prístup, kedy sa na templátový reťazec DNA syntetizoval komplementárny reťazec, pričom bola táto reakcia detekovaná, väčsinou pomocou detekcie vyžiarenia fluorescenčných značiek, ktorými boli označené báze, ktoré sa na templátový reťazec viazali [35]. Pomocou zaznamenávania poradia farieb emitovaných fluorescančných značiek sa potom zapisovalo poradie báz, ktoré sa na DNA viazali a z tejto sekvencie sa potom dala predpokladať sekvencia báz, ktorá tvorila molekulu DNA. Tento prístup bol oproti sekvenačným technológiám prvej generácie výrazným pokrokom v rýchlosti a v cene na jednu naskvenovanú bázu, ale po odštartovaní projektu na nasekvenovanie celého ľudského genómu (Human Genome Project), bol vytvorený tlak na ešte väčšie zefektívnenie sekvenácie, keďže sekvenácia druhej generácie stále využívala veľa reagencií (voľné označené nukleotidové báze, DNA polymeráza, atd.), ktoré reagovali s templátovou DNA za kontrolovaných podmienok pri Polymerázovej reťazovej reakcií – PCR [35]. Navyše asi najväčšou nevýhodou pre metagenomiku bolo, že pri sekvenovaní druhej generácie, vznikali čítania o dĺžkach niekoľko desiatok, prípade pár stoviek bp, čo je pre klasifikáciu a presnú rokonštrukciu genómov nedostatočná dĺžka.

Preto obrovským krokom vpred bol práve vznik sekvenácie tretej generácie, ktorý dal za vznik novému pojmu, tzv. single molecule read, kedy v priebehu prípravy čítaní na sekvenáciu, bola molekula DNA/RNA na časti o dĺžkach v jednotkách kbp, ktoré boli potom v rôznych kanáloch sekvenačného čipu sekvenované naraz. Takýto prístup umožnil vynechanie amplifikácie pomocou PCR reakcie, ktorá bola zdrojom amplifikačných artefaktov a rovnako spôsobovali chyby pri pokrytí genómu. Ďalšou výhodou sekvenovania tretej generácie oproti sekvenovaniu novej generácie (NGS) je aj odstránenie častých chýb pri výskyte homopolymérov, alebo veľmi častých krátkych opakujúcich sa úsekov. Dlhé čítania navyše znamenajú aj veľkú výhodu pri spätnej rekonštrukcií genómov z čítaní, kedy s rastúcou dĺžkou čítaní sa tento proces výrazne uľahčuje [34].

Medzi sekvenátory tretej generácie patria riešenia od firiem Helicos, ktorá predstavila funkčný princíp single molecule sequencingu ako prvá, PacBio využívajúci princíp Single-molecule real-time sequencing (SMRT), Halcyon Molecular využívajúci priame snímanie sekvencií pomocou elektrónovej mikroskopie a MinION, ktorý používa metódu sekvenácie pomocou nanopórov (Oxford Nanopore).[36]

Sekvenátor od firmy Helicos bol prielomovým sekvenátorom, ktorý ako prvý úspešne aplikoval sekvenovanie genómu z jedinej molekuly DNA. V procese prípravy sekvenovania je potrebné rozdeliť molekulu DNA na čítania o dĺžkach niekoľkých stoviek bp, na konce ktorých sú naviazané poly-A konce a posledná adenínová báza je označená fluoroforom. Tieto čítania sa v prvej fázi sekvenovania naviažu na poly-T reťazce, ktoré sú prichytené na stene čipu, z ktorého sa budú sekvencie sekvenovať. Po prichytení čítaní na stenu čipu, sa spustí prvé čítanie, ktoré má za úlohu lokalizovať polohu aktívnych miest na čítanie. Fluorofory na poslednej báze adenínu sú postupne excitované pomocou laseru a ich emitované svetlo zachytáva CCD kamera umiestnená nad čipom. Po inicializačnej excitácií sa na čipe začne fáza vymývania a naväzovania, kedy sa vymyje fluorofor poslednej báze adenínu a na vrch polyT značky sa prichytí prvá báza, komplementárna prvému znaku sekvenoveného reťazca, ktorá je rovnako značená fluoroforom charakteristickým danému druhu báze. Po tejto fázi sa opakuje snímanie naviazaných fluoroforov a kroky vymývania a naväzovania, až získame sekvenciu všetkých čítaní v genóme. [37]

Sekvenátor firmy PacBio využíva podobný prístup, ktorý je ale aplikovaný iným spôsobom. Na sekvenáciu sa miesto čipu Helicosu využíva nanofotonická štruktúra (zero-mode waveguide - ZMW), ktorá sa skladá z komôr s priemerom 100 nm, na dne ktorých sa nachádzajú ukotvené molekuly DNA-polymerázy. Do týchto komôr sa potom privádzajú jednovláknové molekuly ssDNA čítaní, ktoré sú polymerované pomocou nukleiotidov, opäť označených fluoroforom, ktorý sa pri naviazaní na templátovú DNA excituje a potom uvôlni. Pod tenkou komôrkou sa nachádza kamera, ktorá sníma svetlo emitované fluoroformi a podľa farieb zábleskov sa potom skladá sekvencia znakov báz. Priemer komôrok je presne navrhnutý tak, aby cez neho prenikalo len svetlo aktuálne emitovaného fluoroforu, aby sa minimalizovalo svetlo pozadia a iné zdroje rušenia. [38]

Ďalší už spomínaný prístup sa snaží na sekvenáciu, respektíve na rozpoznanie molekúl reťazca DNA použiť trasmisnú elektónovú mikroskopiu (TEM). Tento nápad vznikol pri využití TEM na chemickú a štrukturálnu analýzu materiálov, kedy sa rôzne materiály dali pomocou TEM skúmať doslova atóm po atóme. [39] V roku 1959 bola táto metóda prvýkrat spojená so sekvenáciou, kedy bola navrhnutá metóda, pri ktorej sa molekula DNA natiahla pomocou veľmi tenkých ihiel a reťazec báz by potom mohol byť pomocou TEM prečítaný. Takáto metóda nebola potom nikdy použitá, až kým ju neskôr neaplikovala firma ZS genetics. Na túto aplikáciu ale bohužial nevznikli žiadne štúdie. Vedúce postavenie v tejto oblasti má v dnešnej dobe firma Halcyon Molecular, ktorá priamo zobrazuje a chemicky chemicky detekuje atómy, z ktorých sa skladajú nukleotidy a z ktorých sa skladá templátová DNA. [36]

Prístupov sekvenácie tretej generácie je mnoho. Do práce boli vybrané také prístupy, ktoré reprezentujú odlišne využité princípy s tým istým cieľom – sekvenovanie genómu z jedinej molekuly DNA/RNA. Zo všetkých týcho prístupov najväčší ohlas a popularitu získal prístup spoločnosti Oxford Nanopore, sekvenátor s názvom MinION, ktorý bude viac do detailov popísaný v nasledujúcej kapitole 3.

3 OXFORD NANOPORE

Vznik sekvenátoru MinION s technológiou Oxford nanopore vyvolal veľmi pozitívne ohlasy hlavne kvôli jeho kompaktným rozmerom (10 cm x 3 cm x 2 cm), nízkej cene sekvenátora a zároveň nízkej cene za detekovanú bázu, nízkej hardwareovej náročnosti na pripojený počítač a jednoduchú a rýchlu konektivitu pomocou portu USB3.

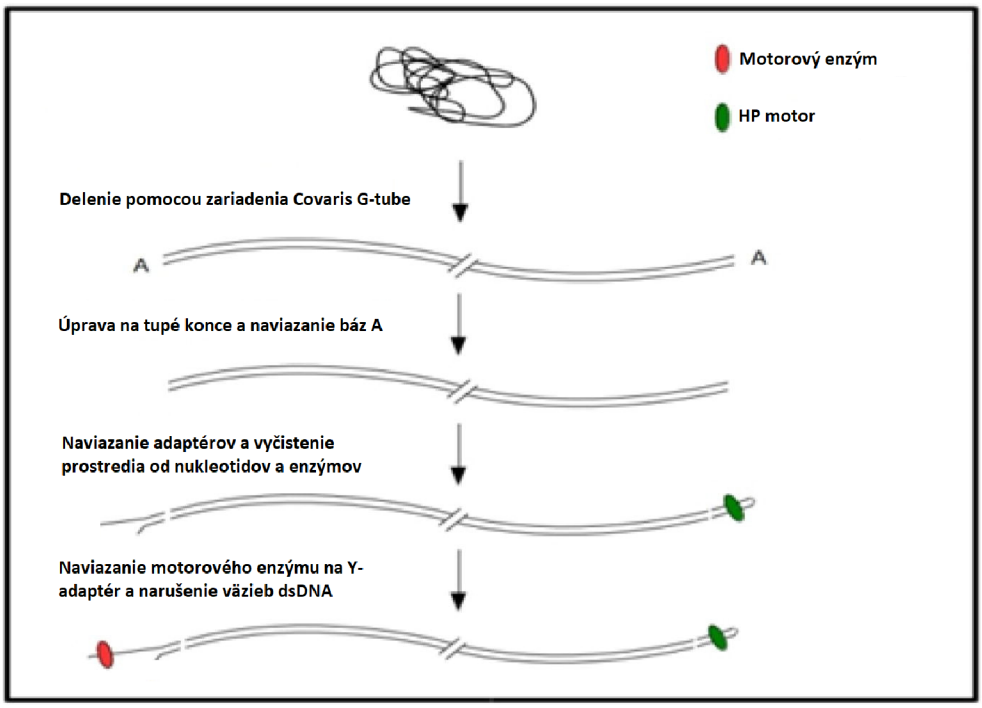
Sekvenácia prostredníctvom sekvenátoru MinION pozostáva z dvoch krokov, kedy prvým krokom je príprava knižnice fragmentov DNA, ktoré sa budú sekvenovať a samotná sekvenácia.

3.1 Príprava knižnice sekvenovaných fragmentov DNA

Príprava knižnice fragmentov sa veľmi neodlišuje od prípravy knižnice ostatných sekvenačných metód tretej generácie. Väčšinou sa sekvenuje dvojitý reťazec DNA (dsDNA), kedy sa sekvenujú oba reťazce v jednom behu.

Prvým krokom prípravy DNA na sekvenáciu je rozdelenie dsDNA na menšie fragmenty, (čítania), ktoré budú sekvenované zároveň v rôznych nanopóroch. Toto rozdelnie reťazca DNA je za využitia zariadenia Covaris g-TUBE, ktoré je schopné reťazec DNA rozdeliť na časti s dĺžkou 6 kbp až 20 kbp. Po rozdelení reťazca DNA nasleduje voliteľný krok, kedy sa molekula upravuje tak, aby oba jej konce boli tupé, teda aby oba reťazce DNA mali presne rovnaký počet báz. Na 3′-koniec molekuly sa potom naviaže báza adenínu a na oba konce sa naviažu tzv. adaptéry. Jeden z adaptérov sa volá Y adaptér kvôli jeho tvaru a tento adaptér je vedúci adaptér, ktorý bude molekulu DNA „prevádzať“ nanopórom, kedy sa sekvenácia začína na jeho 5′-konci. Druhý adaptér sa nazýva HP-adaptér, alebo „hairpin“ adaptér, ktorý má za úlohu prepojiť dva reťazce dsDNA tej istej molekuly do jedného. Na záver prípravy knižnice čítaní je urobená purifikácia, kedy sa prostredie očistí od všetkých zbytkových nukleotidov a enzýmov. Po ukončení prípravy knižnice čítaní tzv. motorový enzým naruší väzby medzi dvoma reťazcami sekvenovaného čítania a vznikne z neho tzv. 2D-read. Použitie HP-adaptéru je voliteľné a v prípade, že ho nepoužijeme, v poslednom kroku, kedy sa prerušia väzby medzi dvoma reťazcami DNA, je druhý, komplementárny reťazec stratený a vnikne tzv.

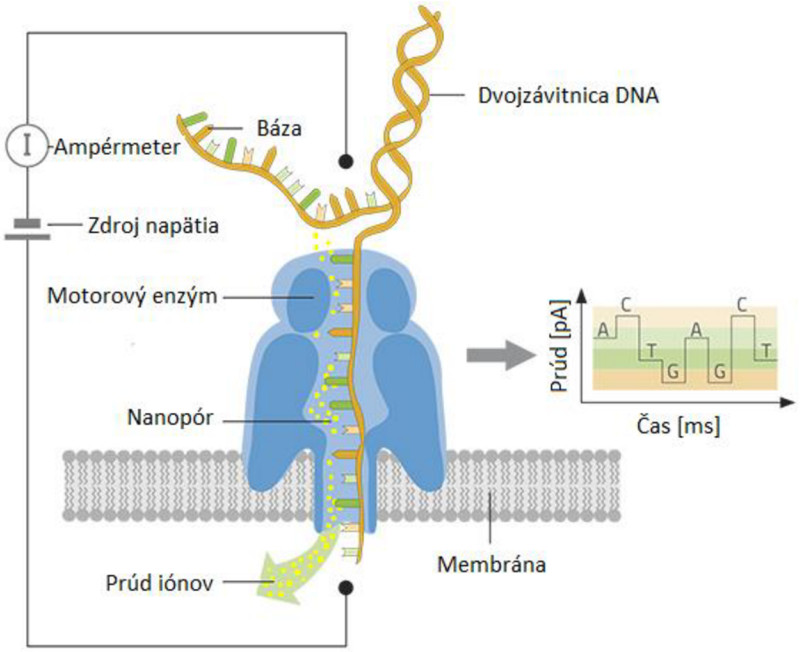
1D-read. Zároveň s posledným krokom sa potom Y-adaptér navazuje na nanopór a sekvenácia je pripravená.[40]



Obr. 1: Schematické znázornenie priebehu prípravy knižnice readov na sekvenáciu. Upravené z [40]

3.2 Proces sekvenácie

Na samotnú sekvenáciu DNA sa využíva sústava nanopórov, ktoré sú usporiadané v prietokovej bunke, tzv. flowcell a na ich vnútornej strane sú kovalentne naviazané syntetické cyklodextrínové senzory. Princípom sekvenácie za využitia nanopórov je, že na oboch stranách prietokovej bunky sú rozdielne potenciály a teda nanopórmi danej flowcell preteká prúd iónov s konštantným prietokom. V priebehu sekvenácie potom jednotlivé molekuly nukleových báz prechádzajú vplyvom tohto prúdu na druhú stranu membrány cez nanopóry a v závislosti na báze, ktorá sa v danom okamihu nachádza v nanopóre, bude vyvolaná zmena pretekajúceho prúdu iónov daným nanopórom, čo vyvolá pokles úrovne prúdu systému na úroveň, ktorá je iná pre každú bázu a teda podľa úrovňe prúdu sme schopní určiť bázu, ktorá sa v póre nachádzala v daný okamih. Celý tento proces na znázornený na Obr. 2.



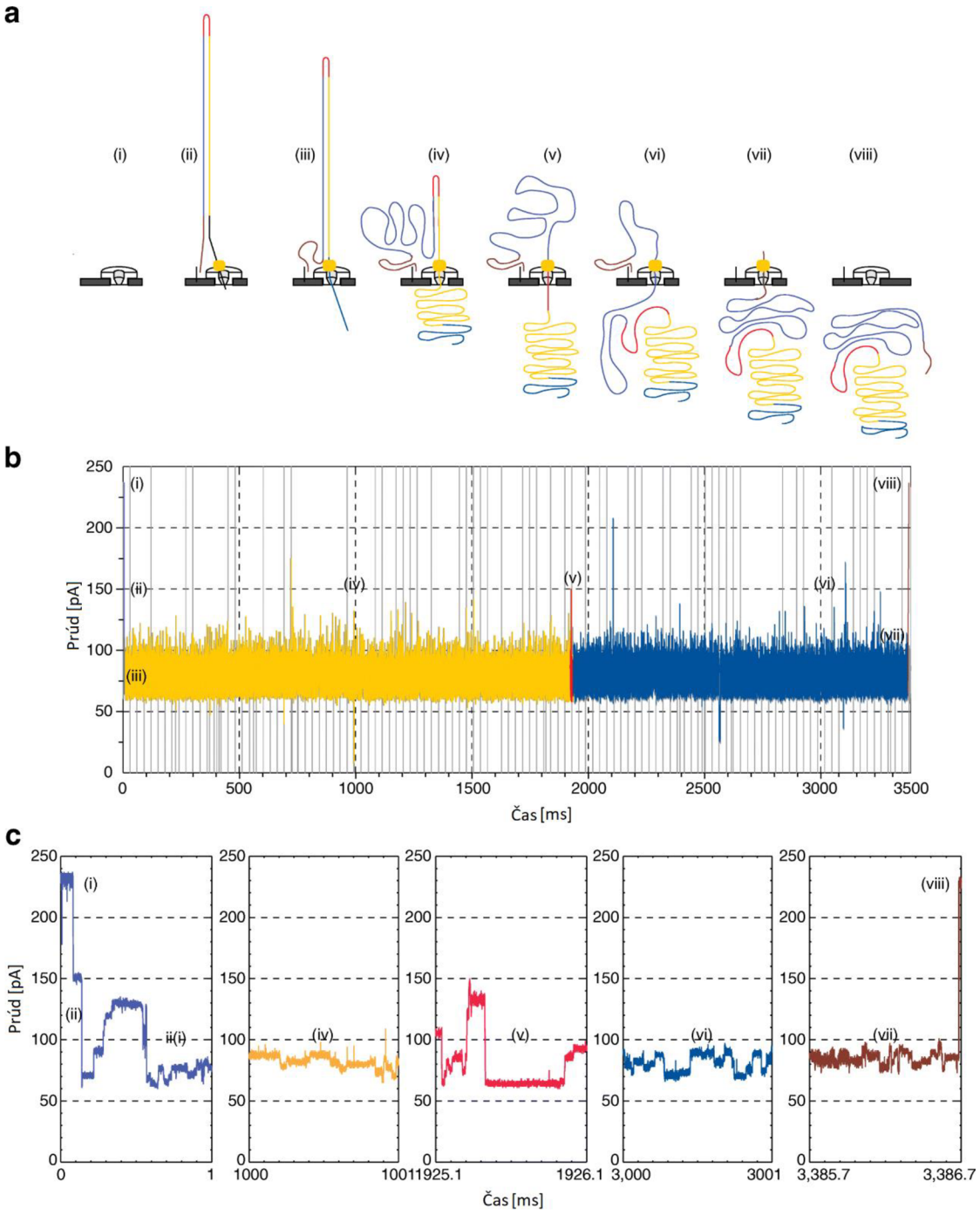
Obr. 2: Ilustrácia priebehu sekvenovania za použitia nanopóru. Upravené z [41].

Ako môžme vidieť na Obr. 2, výstupnými surovými dátami je vlastne závislosť prúdu na čase. Po zaznamenaní tohto signálu potom nasleduje posledná fáza sekvenovania, ktorá je tiež naznačená na pravej strane Obr. 2 a to spracovanie surového signálu a následné rozhodovanie na základe úrovňe spracovaného signálu, aká báza sa na danej pozícií v sekvencií nachádza. Tento proces určovania báz podľa úrovne signálu sa nazýva basecalling. Problém, ktorému sekvenátor MinION zo začiatku čelil bol práve v tejto fázi sekvenácie, kedy v roku 2015 bola odhadnutá presnosť sekvenovania bakteriálnych genómov na 28,6% - 70,7% [42]. Keďže ale MinION v sebe ukrýval veľký potenciál, odštartoval sa proces optimalizácie algoritmov určujúcich báze zo signálu a v dnešných dňoch sa vyšplhala až na 99% [43], preto sa táto metóda ukazuje dnes už ako veľmi spoľahlivá.

3.3 Výstupné dáta sekvenátoru MinION

Ako bolo v predchádzajúcej kapitole spomenuté, surové výstupné dáta skvenátoru MinION sú v podobe závislosti prúdu, ktorý sa vytvára na membráne a na čase. Takýto signál (squiggle signal) sa získava z 512 kanálov, z ktorých každý je pripojený ku 4 čítacím detektorom, ktoré zaznamenávajú daný signál. Pred samotnou sekvenáciou sa vykoná test týchto detektorov, kedy sa zistí najaktívnejší detektor tejto štvorice, ktorý pri samotnom sekvenovaní dostáva prednosť a prvých 24 hodín sekvenácie sa signál získava len z tohto detektora a až potom sa zvyšný čas rozvrhne rovnako medzi ostatné detektory danej skupiny. Každý kanál produkuje svoje vlastné dáta, ktoré sú uložené do zvlášť súboru. Dátový formát, ktorý MinION používa, je formát FAST5, ktorý je vlastne len adaptovaný HDF5 (Hierarchical Data Format 5) formát, ktorý je prispôsobený na tak vysoko objemné dáta ako sú dáta získané zo sekvenácie pomocou MinIONu. Formát FAST5 obsahuje totiž ako surové dáta sekvenácie, tak aj všetky priložené metadáta o jednotlivých čítaniach, čo spolu tvorí desiatky, až stovky GB dát. [43]

Ukážka procesu získania výstupných surových dát sekvenátoru sú zobrazené na Obr. 3. Na Obr. 3a môžme vidieť proces, ako sa jedna molekula nukleovej kyseliny, ktorá je prichytená na nanopóre, postupne dostáva na druhú stranu membrány, pričom sú farebne odlíšené všetky jej súčasti, ktoré vznikli vo fáze prípravy (viď kapitolu 3.1 o príprave knižnice fragmentov DNA). V časti obrázku Obr. 3b potom vidíme, že ako molekula nukleovej kyseliny prechádza cez nanopór, sqiggle signál je produkovaný v reálnom čase, takže v tomto signále môžme vidieť všetky časti molekuly v poradí, ktoré zodpovedá zostaveniu fragmentu molekuly, popísanému v kapitole 3.1. Na Obr. 3c potom už vídíme len surový signál s väčším časovým rozlíšením, kde je očividné, ako si signál počas prestupu jednej báze cez nanopór udržiava rovnaké hladiny napätia, čo zodpovedá teoretickému predpokladu uvedenému v predchádzajúcej kapitole.



Obr. 3: Proces tvorby výstupných dát sekvenátoru MinION. Upravené z [44].

Surové squiggle signály sú potom v závislosti na verzii použitej chémie uložené buď len ako záznamy tzv. udalostí (pri verzii R7), ktoré reprezentujú len priemernú hodnotu prúdu po jeho náhlej zmene, dobu trvania udržania tejto hodnoty na danej úrovni, smerodatnú odchýlku výkyvov od danej úrovne a polohu danej udalosti na časovej osi celkového signálu. V prípade potreby získania squiggle signálu je preto potrebná najskôr

jeho rekonštrukcia zo spomenutých dát. Príchodom novej verzie chémie R9 sa ale radikálne mení proces ukladania dát a viacero parametrov squiggle signálov, ako aj sekvenácie samotnej. Prvým veľkým rozdielom je, že sa surový squiggle signál už neukladá do udalostí, ale ukladá sa kompletne celý jeho priebeh v celku, tak ako bol nasnímaný. Má to výhodu zachovania presnej podoby signálu a menšiu pravdepodobnosť skreslenia dát pri rekonštrukcií. Ďalším rozdielom medzi verziami chémie R7 a R9 je, že v prípade chémie verzie R7, je aktuálna hodnota prúdu ovplyvnená piatimi, alebo šiestimi nukleotidmi, ktoré sa v danom monente nachádzajú v póre. Pri verzií R9 sa podarilo počet týchto nukleotidov znížiť na 3 a teda pomocou chémie R9 je dosiahnutá vyššia rozlišovacia schopnosť a počet možných detekovaných k-mérov výrazne narastá. [46] Prechodom z chémie R7 na R9 rovnako prichádza do hry aj možnosť rýchleho režimu prietoku nukleotidov pórom, kedy sa rýchlosť prietoku nukleotidov zvyšuje zo 70 báz/sekundu na 250 báz/sekundu pri zachovaní rovnakej presnosti pri znížení ceny jedného Gb genomickej informácie. Vzhľadom na túto zmenu musela byť zvýšená aj vzorkovacia frekvencia signálu zo 4000 Hz (R7) na 5000 Hz (R9) pre zanechanie aspoň rovnakej. Squiggle signály sú v oboch prípadoch uložené do súborov formátu FAST5. [47]

Takéto surové dáta sú potom štandardne nahrané na cloudové úložisko Metrichor [45], kde je vykonaný samotný proces určovania báz a po dokončení tohto procesu si užívateľ môže spracovanú sekvenciu stiahnuť naspäť z tohto úložiska. Sekvencia s určenými bázami je uložená v súbore formátu FASTQ [48]. Tento formát je štvorriadkovým formátom, kedy sa v hlavičke začínajúcej znakom @ nachádzajú anotačné dáta, ktoré identifikujú a popisujú danú sekvenciu. V druhom riadku sa nachádza samotná sekvencia báz, ziskaná zo surových dát, bez akéhokoľvek úvodného značenia a v poslednom riadku sa opäť nachádza identifikátor sekvencie s popisom. Obsah tretieho riadku je voliteľný a v prípade, že ostáva prázdny, bez opakovaného zadania identifikátora, je povinné do neho vložiť aspoň jeho inicializačný znak, ktorým je symbol „+“. V poslednom riadku sa potom nachádzajú znaky kvality pre sekvenciu z druhého riadku, ktoré určujú kvalitu, s ktorou bola báza na zodpovedajúcej pozícií určená. Tieto symboly majú dané poradie, ktoré zodpovedá kvalite určenia, kedy toto poradie je presne špecifikované pre formát FASTQ. Príklad krátkej sekvencie uloženej vo FASTQ formáte môžme vidieť na ďalšej strane, na Obr. 4.

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#""""""""""""7F@71,'";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EA0D@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

Obr. 4: Príklad FASTQ formátu. Prevzané z [48].

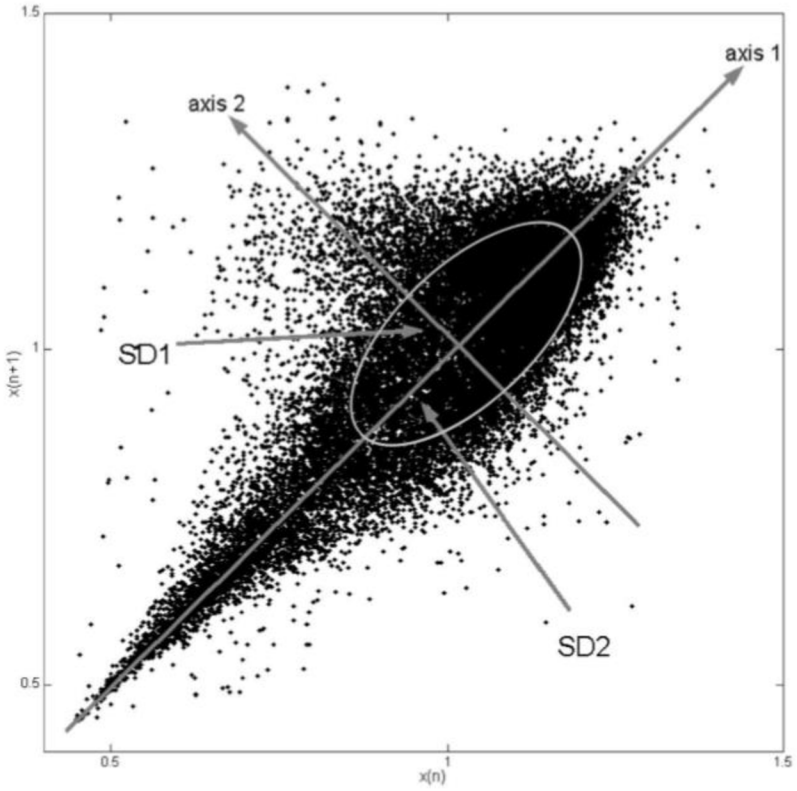
4 NAVRHNUTÁ METÓDA

Cieľom tejto práce je navrhnúť vlastnú metódu klasifikácie metagenomických dát, so zameraním na sekvenačnú platformu MinION. V článku [51] sa pojednáva o Poincarého mapách ako o deskriptore variability biologických signálov, ktorú popisujú celkovo tri parametre: SD1, SD2 a pomer SD2/SD1, ktoré sú vypočítané podľa vzťahov (1). Tieto parametre popisujú smerodatnú odchýlku bodov Poincarého mapy v smere osy zhluku a v smere kolmom na os zhluku. Tieto parametre sú názorne zobrazené na Obr. 5.

$$SD1 = \frac{\sqrt{2}}{2}\sigma(x_n - x_{n-1}),$$

$$SD2 = \sqrt{2\sigma(x_n)^2 - \frac{1}{2}\sigma(x_n - x_{n-1})^2}, \tag{1}$$

Kde SD1 je hodnota smerodatnej odchýlky v smere osi 2 a SD2 je hodnota smerodatnej odchýlky v smere osi 1 elipsy, ktorá je zobrazená na Obr. 5, hodnota x_n je hodnota aktuálneho vzorku výpočtu, hodnota x_{n-1} je hodnota predchádzajúceho vzorku výpočtu a σ je potom smerodatná odchýlka z hodnôt x_n a x_{n-1}.



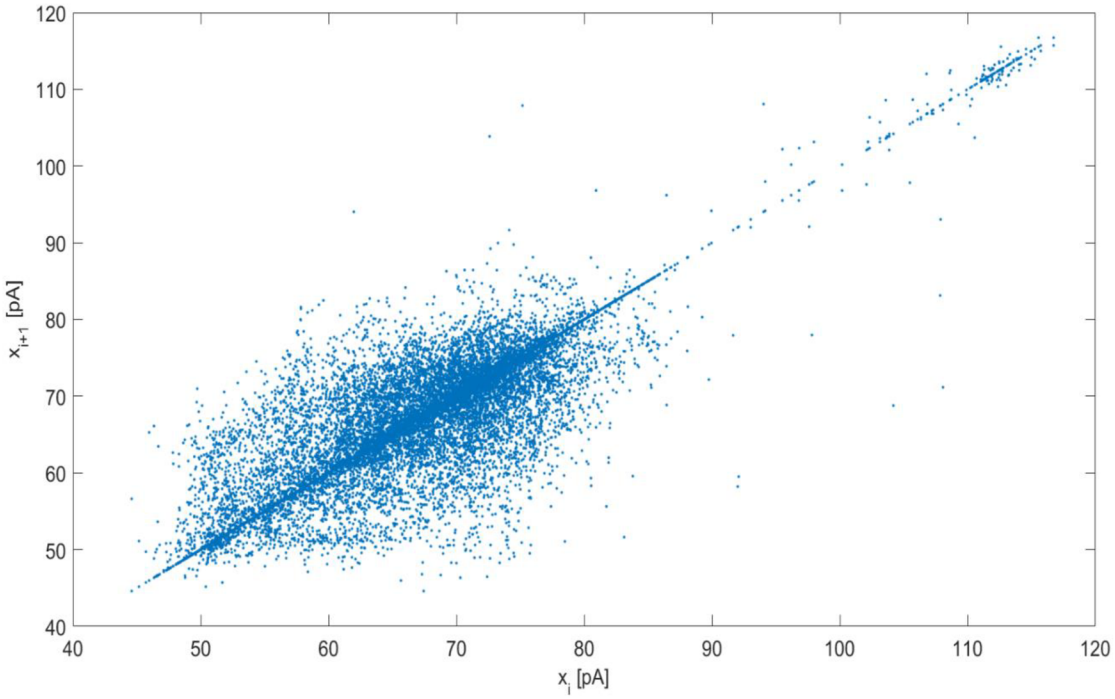
Obr. 5: Príklad Poincarého mapy s vyznačenými parametrami SD1 a SD2. Upravené z [51]

Navrhnutá metóda sa opiera o tento všeobecný fakt, keďže metóda popisu variability biologického signálu sa už dlhú dobu používa pri vyhodnocovaní variability srdečného rytmu z EKG signálov a preto je jej fungovanie pochopené a veľmi dobre popísané [52].

Pre použitie Poincerého máp a ich deskriptorov je ale potrebné niektoré detaily prispôsobiť pre použitie na squiggle signály získané zo sekvenátora MinION. Prvým rozdielom použitia Poincarého máp na vyhodnocovanie variability srdečného rytmu a vyhodnocovanie squiggle signálov je, že pri EKG signáloch používame ako súradnice v mape veľkosti intervalov medzi dvomi R vrcholmi v signále – hodnota R_i a nasledujúceho intervalu ďalšieho R vrcholu – hodnota R_{i+1}.

Keďže v squiggle signáloch takéto intervaly nie sú definovateľné, v navrhnutej metóde sú súradnice máp tvorené z každého jedného vzorku, čiže za súradnicu x je považovaný aktuálny vzorek iterácie x_i a za súradnicu y považujeme hodnotu nasledujúceho vzorku x_{i+1}. Len takto zostrojená Poincarého mapa poskytovala dostatočnú variabilitu signálu na to, aby boli squiggle signály rôznych organizmov od seba navzájom odlíšiteľné pomocou Poincarého deskriptorov. Pri výpočte Poincarého deskriptorov z máp pozostávajúcich len z reprezentácie udalostí, ako sú uložené v súboroch daných squiggle signálov pri použití verzie chémie R7, rozlíšitelnosť poinarého máp klesla pod detekovateľnú úroveň.

Preto úspešnosť tejto metódy je zospodu ohraničená istou vzorkovacou frekvenciou, ktorá v prípade, že je príliž nízka, môže spôsobiť nefunkčnosť metódy. V prípade sekvenátora MinION táto možnosť ale nehrozí, keďže vzorkovacia frekvencia má hodnotu 5000 Hz pri použití chémie verzie R9 a pri použití chémie verzie R7 je vzorkovacia frekvencia 4000 Hz. Čo sa týka rozdielnej frekvencie pri použití rôznych verzií chémie, tak vplyv by mal byť na použitú metódu minimálny, pretože ajkeď má verzia R9+ vyššiu vzorkovaciu frekvenciu, je zvýšený aj prietok cez nanopór a teda udalosti v signále by mali mať síce kratšie trvanie, no to je vykompenzované väčšou vzorkovacou frekvenciou a teda variabilita signálu by mala byť ovplyvnená len minimálne. Tento fakt ale nie je potvrdený z dôvodu neprístupnosti celogenómových dát sekvenovaných za použitia chémie verzie R9.



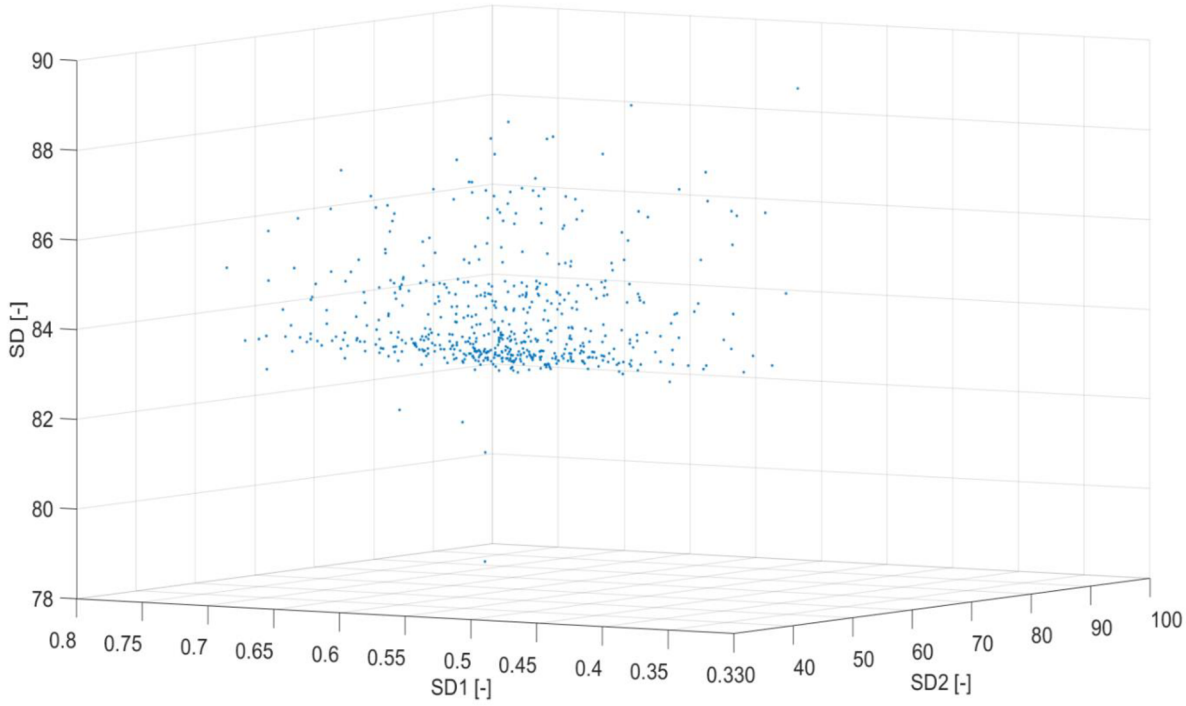
Obr. 6: Poincarého mapa pre jedno čítanie

Takýmto spôsobom vznikne Poincarého mapa pre každé čítanie v behu a teda pre každé čítanie behu dostávame rovnako tri vyššie spomínané deskriptory SD1,SD2 a ich pomer SD^{-1}. Ponicarého mapa na Obr. 6 je mapa získaná z jedného z čítaní patriaheho organizmu *E. coli.* z modelového genómu, ktorý je popísaný v kapitole 4.1. Pri určovaní variability signálu pomocou pomeru hodnôt SD1 a SD2 je zaužívané tento pomer počítať podľa vzťahu (2), ale lepšiu rozlišovaciu schopnosť preukázal výpočet prevrátenej hodnoty tohto pomeru, teda výpočet podľa vzťahu (3), v čom spočíva ďalšie prispôsobenie danej metódy pre využitie klasifikácie čítaní získaných zo sekvenátora MinION.

$$SD = \frac{SD1}{\mathrm{SD2}} \tag{2}$$

$$SD^{-1} = \frac{SD2}{\mathrm{SD1}} \tag{3}$$

V ďalšom kroku analýzy je potom zo všetkých troch deskriptorov vytvorený 3D priestor, kedy na ose x sa nachádzajú hodnoty SD2, na ose y sa nachádzajú hodnoty SD1 a na ose z sa nachádzajú hodnoty SD^{-1}. Takýto priestor vytvorený zo všetkých squiggle signálov získaných zo sekvenácie genómu organizmu *Microcystis aeruginosa* je zobrazený na Obr. 7.



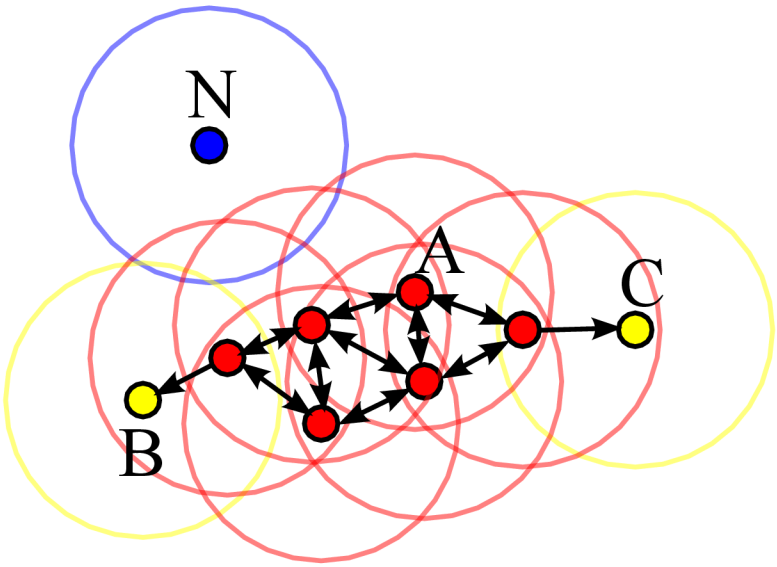
Obr. 7: 3D priestor Poincarého deskriptorov organizmu *Microcystis aeruginosa*

Keďže má navrhnutá metóda na klasifikovať dáta pochádzajúce z metagenómu, je potrebné vybrať spávnu metódu zhlukovej analýzy, ktorá získané čítania roztriedi podľa ich pôvodu do zhlukov patriacich jednotlivým organizmom. Výberom vhodnej zhlukovej analýzy sa budeme zaoberať v kapitole 5. Po tejto záverečnej fázi sú potom výstupnými dátami metódy roztriedené čítania metagenómu, získané zo sekvenácie metagenomických dát, ktoré by potom užívateľ po ďalšej analýze mal byť schopný priradiť konkrétnym organizmom.

Ako prvý bol testovaný algoritmus zhlukovej analýzy k-means. Tento algoritmus funguje na princípe náhodného umiestnenia centroidov zhlukov, stanovenia euklidovských vzdialeností jednotlivých bodov dát k centroidom a iteračným postupom sa potom snaží upraviť umiestnenie centroidov tak, aby minimalizoval priemernú vzdialenosť k zhlukom prislúchajúcim k daným centroidom. Na konci iteračného postupu sú potom konečné zhluky stanovené na základe toho, ku ktorému centroidu sú najbližšie.

Ako druhý bol testovaný algoritmus Density-based spatial clustering of applications with noise (DBSCAN) [53]. Tento zhlukovací algoritmus sa výrazne líši od fungovania algoritmu k-means, pretože funguje na základe hustoty umiestnenia dát. Na vstupe je

treba algoritmu zadat dva parametre – MinPoints a ε. Parameter MinPoints určuje počet bodov, ktorý analyzovaný bod musí vo svojom susedstve o veľkosti ε obsahovať, aby bol považovaný za bod niektorého zhluku a nie za bod šumu. Body, ktoré tieto podmienky nespĺňajú, sú potom považované za šum. Samotné zhluky sú potom tvorené len z bodov, ktoré spĺňajú dané podmienky a tvoria sa na princípe dosiahnuteľnosti, kedy za bod toho istého zhluku sa považujú body priamo susediace s niektorými z bodov zhlukov, alebo ku ktorým vedie cesta skrz iné body zhluku. Tento postup je znáznornený na Obr. 1.



Obr. 8: Ilustrácia algoritmu DBSCAN. Prevzaté z [54]

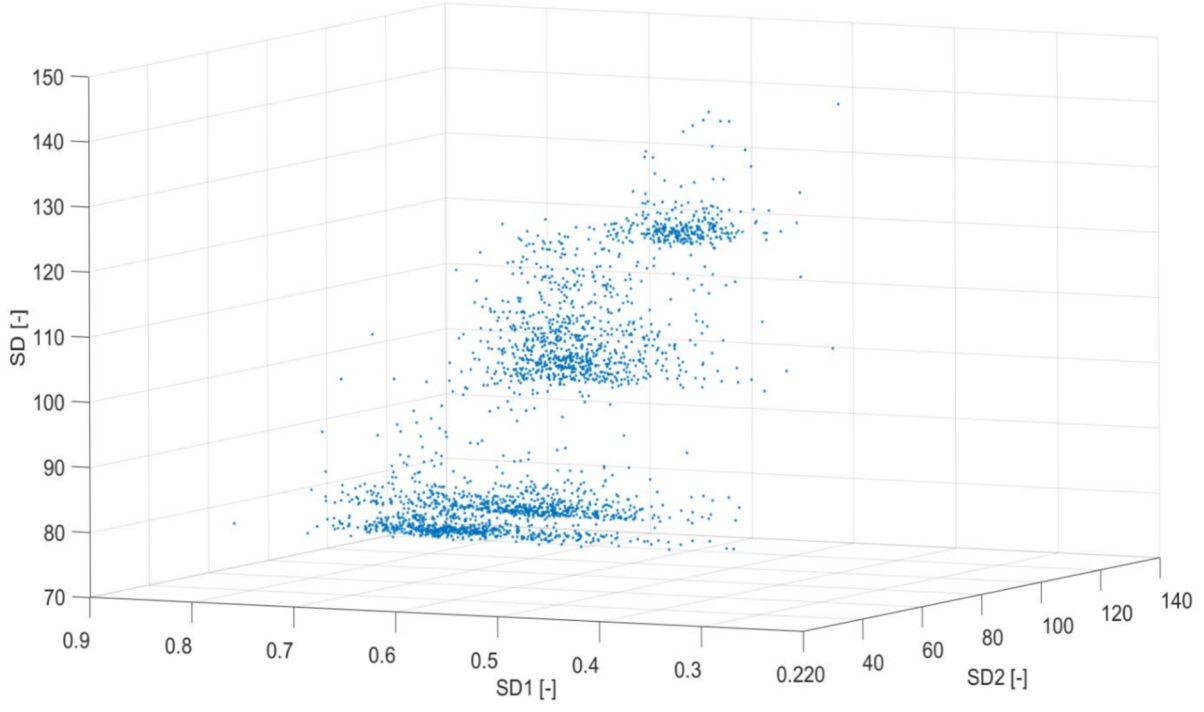
Pozn.: V práci bola použitá implementácia algoritmu DBSCAN v Matlabe, ktorá bola vytvorená užívateľom Yarpiz a nahraná na oficiálnu stránku spoločnosti Mathworks, pre zdielanie užívateľsky vytvorených funkcií pre Matlab. [55]

4.1 Modelový metagenóm

Pre účely vytvorenia modelu umelého metagenómu boli vybrané genómy troch organizmov - *Escherichia coli, Microcystis aeruginosa* a *Pseudomonas fluorescens*, ktoré boli súčasťou umelo vytvoreného metagenómu štúdie [49] a dáta zo sekvenovania boli nahrané do voľne prístupnej databáze ENA (European Nucleotide Archive) na portáli www.ebi.ac.uk spravovanom EMBL-EBI, pod štúdiou s kódom PRJEB8672. Genómy patria organizmom, ktoré boli kultivované v spoločnej kultúre, kde mali každý iné početné zastúpenie a preto sa bude líšiť aj počet získaných čítaní na jeden organizmus.

Po kultivácií bola z organizmov vyextrahovaná DNA nasekvenovaná pomocou sekvenátora MinION, pri použití chémie verzie R7.

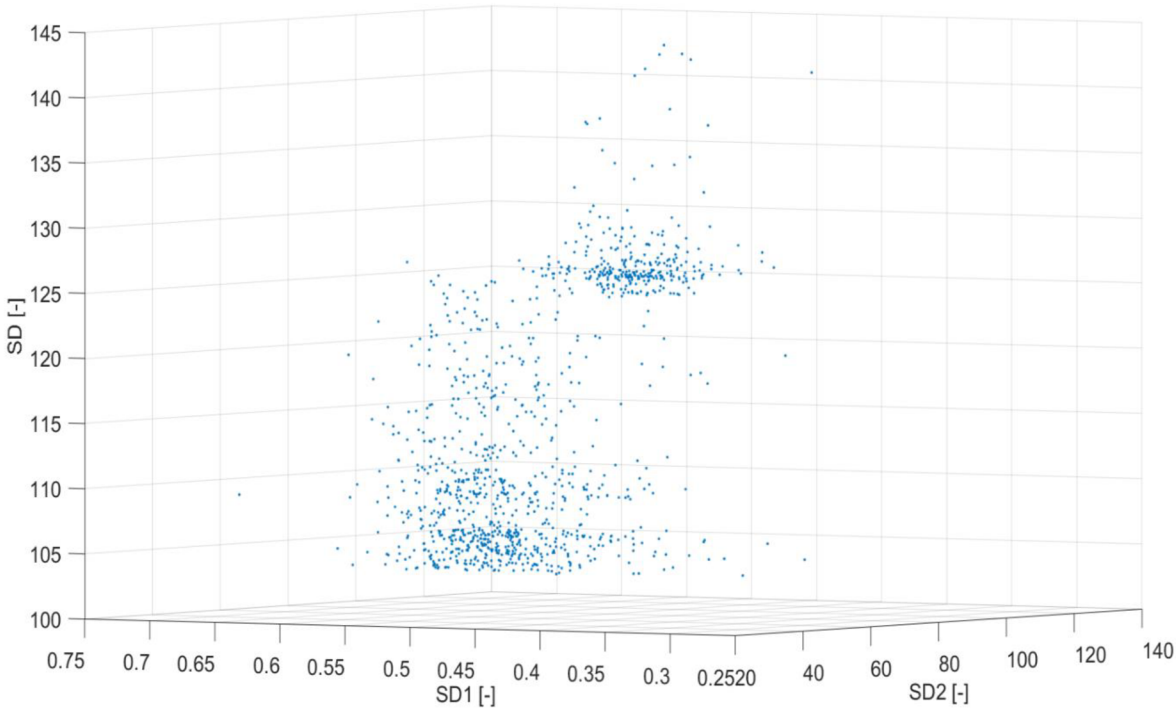
Modelový metagenóm bol potom vytvorený načítaním dát každého genómu zvlášť a podľa postupu v predchádzajúcej kapitole 4, boli dáta vložené do rovnakého 3D priestoru, kde sa vytvorili dáta metagenómu pre ďalšiu klasifikáciu. Tento priestor je zobrazený na Obr. 9.



Obr. 9: 3D priestor Poincareho deskriptorov pre modelový metagenóm

5 VÝSLEDKY NAVRHNUTEJ METÓDY

Ako je už na Obr. 9 viditeľné, v 3D priestore vytvoreného z troch organizmov, máme vytvorené 4 viditeľné zhluky, pričom najlepšiu rozlišovaciu schopnosť nám poskytuje práve osa z, na ktorej sú hodnoty SD. Keď sa ale pozrieme na 3D priestor obsahujúci dáta genómu organizmu *Escherichia coli*, je zrejmé, odkiaľ sa vo vytvorenom priestore objavil jeden zhluk navyše. Hodnoty deskriptorov tohto organizmu totiž vytvárajú dva rôzne zhluky a teda pre jeden organizmus získavame dva zhluky. Pôvod tohto dvojitého zhluku nie je zrejmý, keďže v popise štúdie [49] bolo uvedené, že genómy pochádzajú od organizmov z jednej zmiešanej kultúry, z ktorých bola po istej dobe kultivácie vyextrahovaná DNA.



Obr. 10: 3D priestor Poincarého deskriptorov genómu *Escherichia coli*

Z toho vyplýva, že vo vytvorenom priestore sú rozlíšiteľné všetky tri organizmy metagenómu podľa osi SD a medzi nimi sú jasne viditeľné hranice. Preto úloha klasifikácie dát pomocou zhlukovej analýzy by pri správne zvolenej metóde mala fungovať spoľahlivo.

Na analýzu správnosti fungovania zhlukovania boli použité tri kritéria – Accuracy (presnosť), ktorá popisuje pomer medzi správne identifikovanými dátam a celkovými dátami, Recall (spoľahlivosť), ktorá popisuje pomer medzi správne identifikovaným pozitívnymi detekciami a celkovymi detekciami daného zhluku a Precision (precíznosť), ktorá popisuje pomer medzi správne detekovanými pozitívnymi dátami a celkovým počtom pozitívne detekovaných dát. Tieto tri parametre boli vypočítané samostatne pre každý detekovaný zhluk podľa vzťahov (4).

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN},$$

$$Recall = \frac{TP}{TP+FN},$$

$$Precission = \frac{TP}{TP+FP}, \tag{4}$$

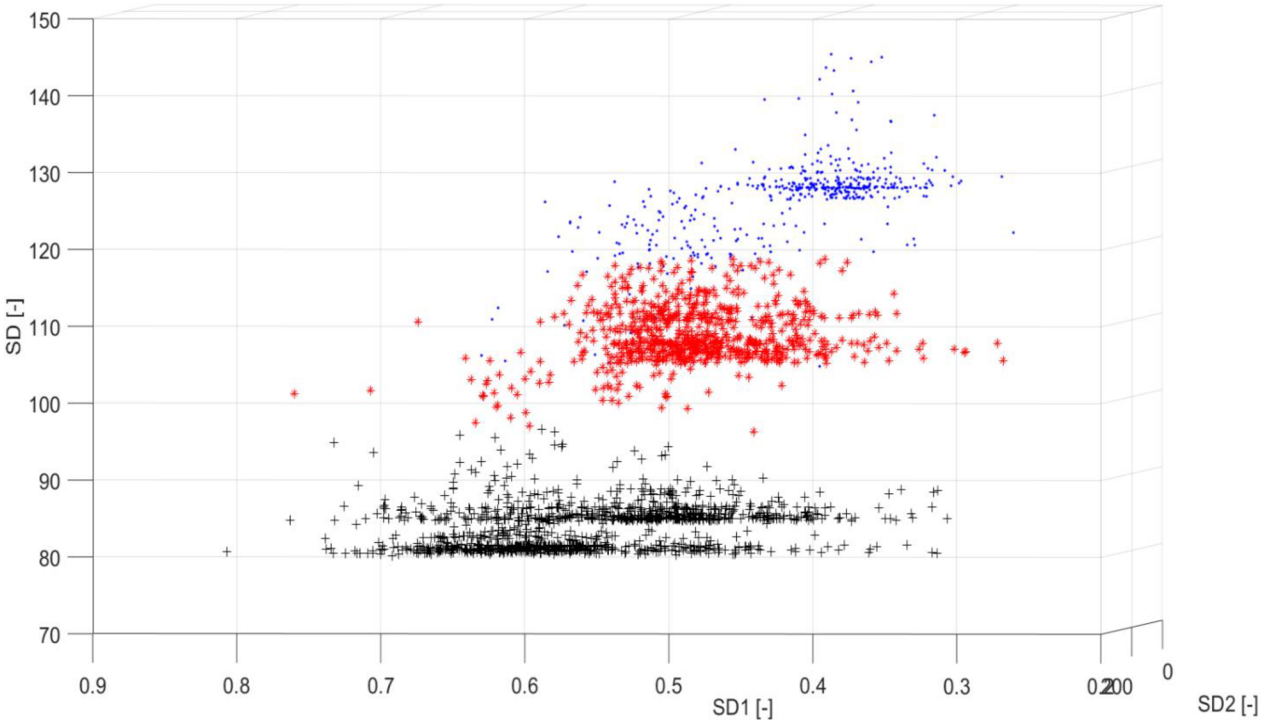
kde TP – počet pravdivo pozitívne vyhodnotených dát, TN – počet pravdivo negatívne vyhodnotených dát, FP – počet falošne pozitívne vyhodnotených dát a FN – počet falošne negatívne vyhodnotených dát.

Vyhodnotenie dát bolo urobené pre každý jeden zhluk zvlášť, kedy medzi TP hodnoty boli zaradené také dáta, ktoré reálne patrili medzi dáta organizmu nachádzajúceho sa na pozícií detekovaného zhluku. Medzi hodnoty TN boli zaradené také dáta, ktoré boli počas zhlukovej analýzy vyhodnotené ako dáta neprislúchajúce danému zhluku a zároveň reálne naozaj pochádzali od iného organizmu, ako toho, ku ktorému detekovaný zhluk patril. Za hodnoty FP boli považované také hodnoty, ktoré reálne patrili organizmu inému, ako tomu, ku ktorému daný zhluk prislúcha, ale napriek tomu boli vyhodnotené tak, že do tohto zhluku patria. Na záver za hodnoty FN boli považované dáta, ktoré reálne prislúchali organizmu daného zhluku, ale boli zhlukovou analýzou vyhodnotené negatívne, teda do daného zhluku neboli zahrnuté.

Ako prvá metóda zhlukovania bola použitá metóda k-means, ktorej výsledky sú uvedné v tabuľke Tab. 1 a rozdelenie zhlukov je zobrazené na Obr. 11.

Tab. 1: Tabuľka výsledkov zhlukovania pre metódu k-means.

	Pôvodný počet čítaní	**Detekovaný počet čítaní**	**Accuracy [%]**	**Precision [%]**	**Recall [%]**
Escherichia coli	1112	1286	73.08	99.56	40.65
Microcystis aeruginosa	569	454	70.84	44.25	100
Pseudomonas fluorescens	778	719	43.92	8.21	7.58

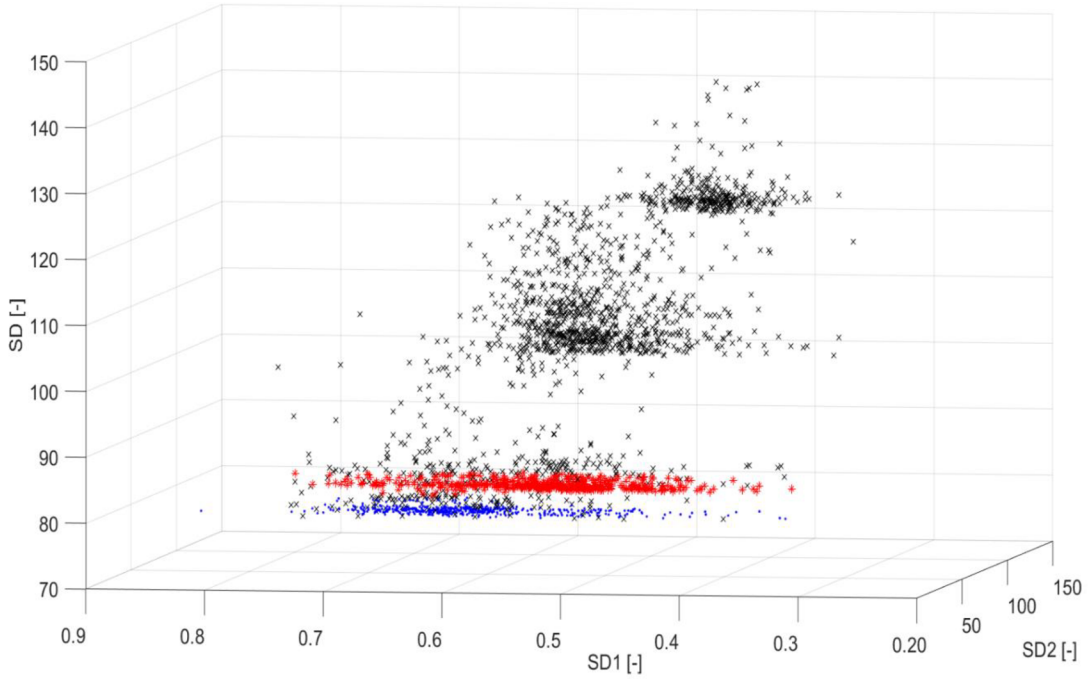


Obr. 11: Výsledok zhlukovania pre metódu k-means.

Ako druhá testovaná metóda zhlukovania bola metóda DBSCAN, ktorá sa od k-means líši tým, že sa jedná o metódu zhlukovania, ktorá je založená na zhusťovaní dát v jednom zhluku a teda pri pohlade na vstupné dáta na Obr. 9 je zrejmé, že by táto metóda mala byť efektívnejšia. Výsledky jej použitia sú uvedné v tabuľke a rozdelenie zhlukov je zobrazené na Obr. 11.

Tab. 2: Tabuľka výsledkov zhlukovania pre metódu DBSCAN.

	Pôvodný počet čítaní	**Detekovaný počet čítaní**	**Accuracy [%]**	**Precision [%]**	**Recall [%]**
Escherichia coli	1112	1516	83.41	73.16	100
Microcystis aeruginosa	569	458	95.74	91.43	99.76
Pseudomonas fluorescens	778	485	87.96	99.79	62.08



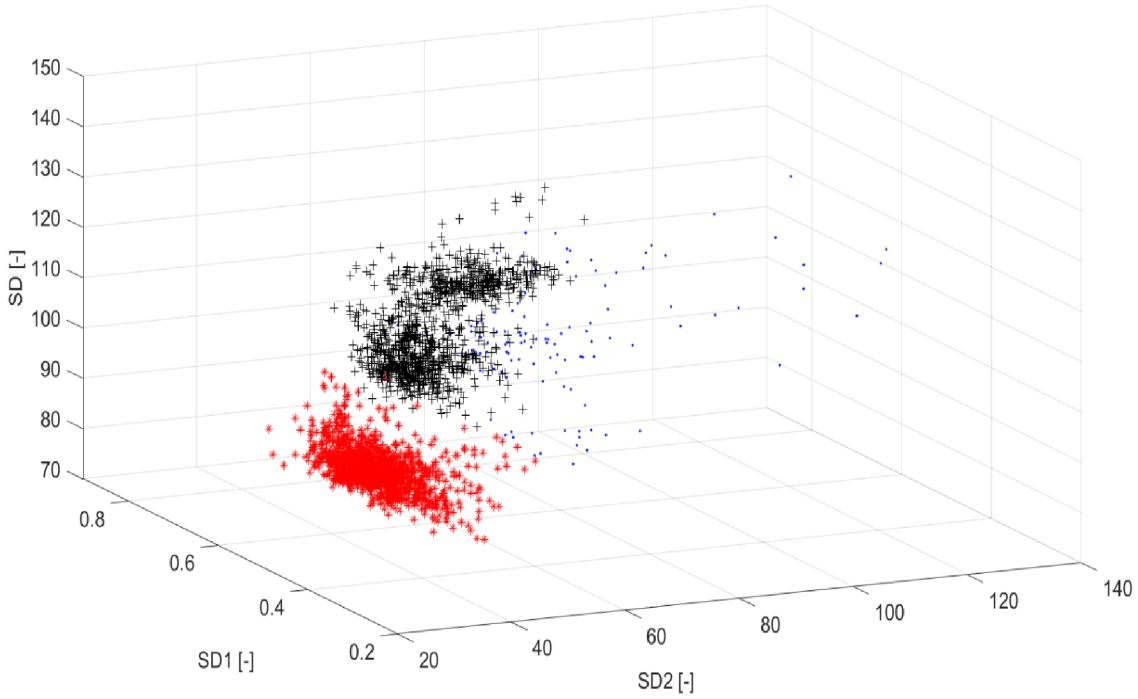
Obr. 12: Výsledok zhlukovania pre metódu DBSCAN.

5.1 Diskusia výsledkov navrhnutej metódy

Podľa výsledkov uvedených v predchádzajúcej kapitole 5 je vybraná ako efektívnejšia a teda vhodnejšia metóda zhlukovej analýzy metóda DBSCAN, ktorá reprezentovala zhlukovacie metódy založené na hustote dát. Táto metóda totiž dokázala zhluky organizmov identifikovať s oveľa väčsou presnosťou, precíznosťou a spoľahlivosťou ako metóda k-means. To je rovnako viditeľné aj pri vizuálnom porovnaní dát, kedy sa metóde DBSCAN podarilo správne odlíšiť zhluky reprezentujúce organizmy *Microcystis aeruginosa* a *Pseudomonas fluorescens*, ktorých zhluky sú veľmi blízko seba. Z tohto dôvodu metóda k-means zlyhala práve v tomto bode, pretože neberie do úvahy hustotu zhlukov, a jej parameter zhlukovania je len minimalizácia euklidovských vzdialeností v rámci zhlukov a keďže v modelových dátach sa nachádzajú organizmy s väčšími vzdialenosťami medzi jednotlivými dátami – *Escherichia coli* a organizmy s malými euklidovskými vzdialenosťami medzi dátami – *Microcystis aeruginosa* a *sseudomonas fluorescens*, k-means tieto rozdiely spriemerovala a tým pádom zaradila do jedného zhluku dva organizmy a genóm *Escherichie coli* rozdelila na dva zhluky.

Matóda DBSCAN nopak s rozdielnymi vzdialenosťami v rámci jedného zhluku počíta a preto, ako je to z výsledkov viditeľné, zhluky identifikovala s vysokou presnosťou, pri jej počiatočnom nastavení vstupných parametrov MinPoints = 41 a $\varepsilon = 1$.

Ďalšou nevýhodou použitia metódy k-means bola aj časť jej metódy, kedy počiatočné umiestnenie centroidov funguje na princípe náhody. Z tohto dôvodu boli výsledky zhlukovania pri každom behu algoritmu výrazne odlišné od seba navzájom a s každým behom boli preto obdržané výrazne odlišné výsledky. Príklad iného behu zhlukovania pomocou algoritmu k-means je zobrazený na Obr. 13, kedy je viditeľné chybné rozdelenie 3D priestoru v smeroch ôs SD a SD2, pri čom smer osi SD2 nemá vôbec žiadnu rozlišovaciu schopnosť a výsledné zhluky teda vôbec nereprezentujú odlišné organizmy.



Obr. 13: Výsledok opakovaného behu zhlukovacej metódy k-means.

Metóda DBSCAN oproti tejto metóde funguje spoľahlivo, keďže pri opakovaných behoch vždy poskytla rovnaké výsledky a zároveň ako je to z tabuľky Tab. 2 viditeľné, jej detekcia je vysoko efektívna, kedy do zhluku organizmu *Escherichie coli* priradila len pomerne malú časť čítaní z ostatných organizmov a jednotlivé zhluky sú od seba stále dostatočne odlíšiteľné aj po vizuálnej stránke.

V tabuľke potom vidíme spriemerované hodnoty Accuracy, Precision a Recall pre metódu využívajúcu zhlukovaciu metódu DBSCAN a aj z týchto hodnôt je viditeľná vysoká efektivita vyhodnotenia navrhnutej metódy. Preto je do navrhnutej metódy priamej klasifikace metagenomických signálov zo sekvenácie nanopórom zaradená zhluková metóda DBSCAN.

Tab. 3: Celkové spriemerované hodnoty Accuracy, Precision a Recall pre metódu s využitím metódy zhlukovania DBSCAN.

Accuracy [%]	**Precision [%]**	**Recall [%]**
89,04	88,13	87,28

5.2 Porovnanie výsledkov

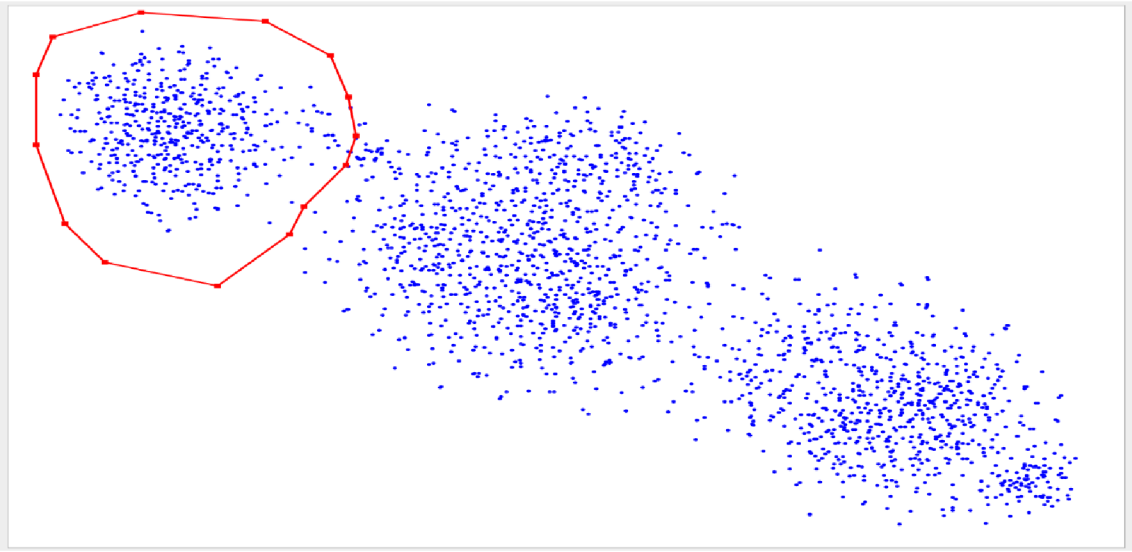
V nasledujúcej kapitole bude navrhnutá metóda porovnaná s už existujúcou metódou klasifikácie metagenomických dát – VizBin. Software VizBin je napísaný v programovacom jazyku Java a je spustiteľný v operačnom systéme Windows. Z hardwarového hľadiska ide o pomerne jednoduchý software, ktorý je schopný metagenomické dáta pomerne rýchlo a efektívne klasifikovať. Ako vstupné dáta požaduje software súbory čítaní, ktoré sú vo formáte FASTA, teda dáta zo sekvenátoru MinION musia najskôr prejsť tzv. basecallingom, v ktorom sú zo surového signálu squiggle získané znakové postupnosti báz. Po basecallingu získavame sekvenciu báz v súbore formátu FASTQ, ako je to popísané v kapitole 3.3. Tu sa stretávame s prvou nevýhodou z užívateľského hľadiska, kedy software VizBin nie schopný pracovať s dátami vo formáte FASTQ a užívateľ je donútený dáta z formátu FASTQ preformátovať do formátu FASTA, pričom pri navrhnutej metóde táto potreba nevzniká, keďže pracujeme už so surovými squiggle signálmi. Toto zamedzuje vzniku rôznych potenciálnych chýb či už v priebehu basecallingu, alebo preformátovania dát a navyše je to užívateľsky oveľa jednoduchšie.

Dáta, ktoré sú uvedené v kapitole 4.1 boli vyhodnotené softwareom VizBin rovnakým spôsobom, ako pri navrhnutej metóde v predchádzajúcej kapitole a výsledky tohto vyhodnotenia sú uvedené v tabuľke Tab. 4.

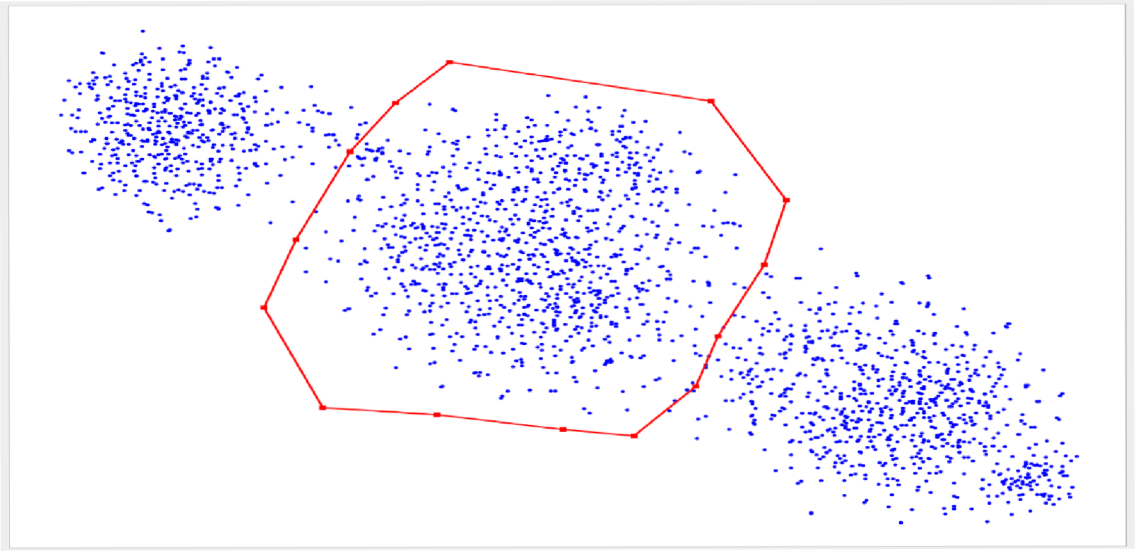
Tab. 4: výsledky vyhodnotenia modelových dát softwarom VizBin.

	Pôvodný počet čítaní	**Detekovaný počet čítaní**	**Accuracy [%]**	**Precision [%]**	**Recall [%]**
Escherichia coli	1112	1038	95,50	92,77	97,08
Microcystis aeruginosa	569	491	95.24	96,54	83,60
Pseudomonas fluorescens	778	783	95,84	91,83	95,74

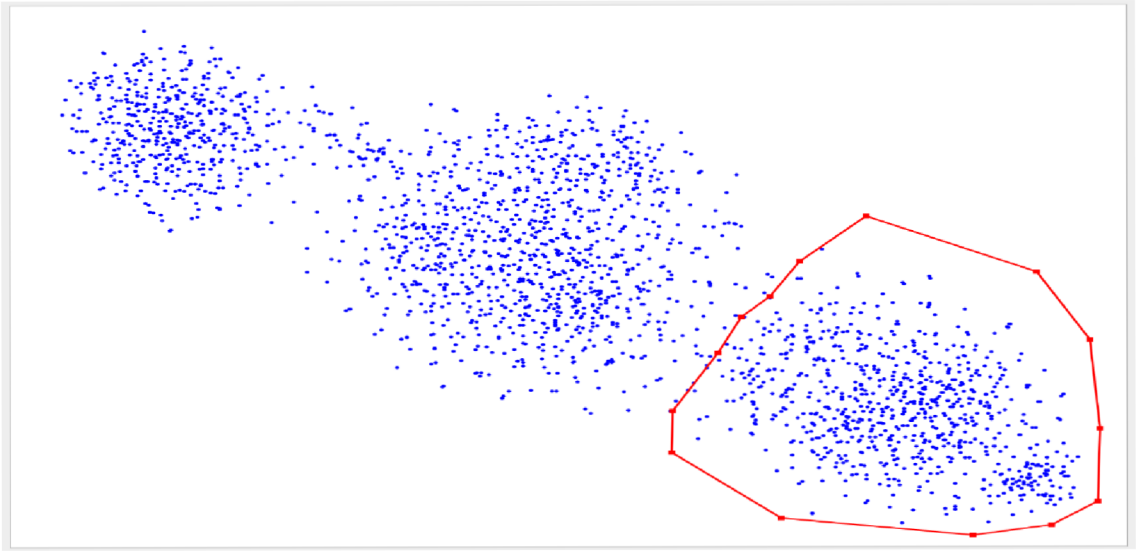
Ako je z tabuľky Tab. 4 viditeľné, modelové dáta boli vyhodnotené s vysokou efektivitou, ktorá bola ale od odladenej metódy používanej v praxi očakávaná. Pozitívne je, že efektivita klasifikácie je porovnateľná s navrhnutou metódou a teda po dlhšom odlaďovaní navrhnutej metódy by sa pravdepodobne mohla svojou efektivitou priblížiť k už existujúcim metódam. Po bližšom preskúmaní ale bol odhalený ďalší nedostatok metódy VizBin. Po zobrazení reprezentácie čítaní v 2D priestore je užívateľ nútený manuálne ohraničiť zhluky, ktoré v 2D priestore vidí a po ich označení sa mu automaticky vygeneruje FASTA súbor s čítaniami, ktoré sám označil za jeden zhluk. Ohraničenie zhlukov v prípade tejto klasifikácie je zobrazené na obrázkoch Obr. 14, Obr. 15 a Obr. 16



Obr. 14: Prvý ohraničený zhluk v 2D priestore reprezentácie čítaní metódy VizBin.



Obr. 15: Druhý ohraničený zhluk v 2D priestore reprezentácie čítaní metódy VizBin.



Obr. 16: Tretí ohraničený zhluk v 2D priestore reprezentácie čítaní metódy VizBin.

Takáto metóda má dve veľké nevýhody. Prvou je subjektívne vnímanie rozloženia bodov ako zhluk, ktoré môže mať za následok, že zhluky budú ohraničené nesprávne v prípade, kedy by šlo o nejednoznačnosť ohraničenia zhlukov, čo je pri biologických dátach vždy veľmi pravdepodobné. V prípade modelových dát sú zhluky od seba navzájom dobre rozoznateľné, ale aj v tomto prípade bolo ťažke exaktne určiť hranicu hlavne medzi druhým a tretím zhlukom.

Druhou nevýhodou takéhoto prístupu je možnosť vynechania niektorých čítaní, keďže po označení jedného zhluku a exportu čítaní do FASTA súboru, hranice tohto zhluku zmiznú a vzniká nebezpečenstvo, že niektoré čítania budú zaradené do rôznych zhlukov dvakrát, alebo ako sa to stalo v tomto prípade, niektoré čítania nebudú do zhlukov zaradené vôbec. V tomto prípade bolo zo zhlukov úplne vynechaných 147 čítaní kvôli užívateľskej chybe. V navrhnutej metóde možnosť vzniku takejto chyby neexistuje, keďže je proces kvalifikácie plne automatický, bez vonkajšieho zásahu úžívateľa.

S vybraným prístupom metódy VizBin sa spája aj získavanie rôznych výsledkov pri opakovaní kvalifikácie tých istých dát. Existuje vysoká pravdepodobnosť, že užívateľ pri opakovanom výbere ohraničenia zhlukov nevyberie presne tie isté hranice a teda čítania každým ďalším pokusom budú do zhlukov vybrané mierne odlišné skupiny čítaní. Tento fakt ešte posilňuje aj to, že metóda klasifikácie VizBinu je založená na princípe BH-SNE, ako je to spomenuté v kapitole 2.1, kedy tato metóda pri každom behu získava mierne odlišné dáta. Pri deterministickom výpočte Poincarého deskriptorov čítaní a pri zvolenej metóde zhlukovania dát DBSCAN, ktorá je rovnako deterministická, pri každom opakovanom behu na tých istých dátach určuje zhluky rovnako, a tento nedostatok je teda rovnako odstránený.

Zaujímavým pozorovaním je, že zhluk 3 na obrázku Obr. 16 javí znovu známky dvoch častí. Jedná sa o zhluk organizmu *Escherichi coli* a jej genóm, ktorý je rozdelený na dva zhluky aj u navrhnutej metódy a je viditeľný na obrázku Obr. 10. Navrhnutá metóda poskytuje pre tieto dva zhluky lepšiu rozlišovaciu schopnosť a je schopná ich oddeliť výraznejšie. Rovnako sa vizuálne javia zhluky jednotlivých organizmov ako lepšie a výraznejšie ohraničené u navrhnutej metódy, preto by sa dalo konštatovať, že navrhnutá metóda má vizuálne lepšiu rozlišovaciu schopnosť ako metóda VizBin.

Veľkou nevýhodou metódy VizBin je aj jej grafický výstup. Ako je na obrázkoch Obr. 14, Obr. 15 a Obr. 16 vidieť, grafický výstup klasifikácie metódou VizBin má nepopísané osi a preto pôsobí veľmi stroho a užívateľ preto nemusí mať istotu o aké dáta sa konkrétne jedná. Rovnako v prípade potreby ďalšej analýzy dát je znemožnené získať presné súradnice bodov čítaní, čo síce pre potreby zistenia obsahu zhlukov nemusí byť pre užívateľa problém, ale pre hlbšiu analýzu funkcie metódy sa to ako problém prejavilo. Preto bol medzi výstupné dáta navrhnutej metódy zaradený aj presný zoznam názvov

čítaní priradených do jednotlivých zhlukov ako aj ich presné hodnoty SD1, SD2 a SD^{-1}, čo sa rovnako zaraďuje medzi výhody tejto metódy, ktorá prispieva k poskytnutiu exaktnejších výsledkov ako metóda VizBin.

Záverom porovnania navrhnutej metódy klasifikácie metagenomických dát a už existujúcej metódy VizBin teda je, že tieto metódy sú porovnateľné. Pri vyššej efektivite klasifikovania dát u metódy VizBin, ktorá je ale ešte stále porovnateľná s navrhnutou metódou, má samotný software niekoľko designových nedostatkov, ktoré sú jednak užívateľsky nepohodlné a zároveň môžu byť zdrojom zbytočných chýb pri klasifikácií dát. Tieto chyby klasifikácie sú pri navrhnutej metóde eliminované a klasifikácia je, ajkeď s menšou efektivitou, užívateľsky oveľa priateľnejšia a jej detekcia je opakovateľná stále s rovnakým výsledkom.

6 ZÁVER

V tejto diplomovej práci boli splnené všetky ciele vytýčené v úvode a zadaní práce. Na začiatku teoretickej časti práce bol popísaný vznik vedného oborou metagenomiky, spôsob jeho uplatnenia a bol popísaný druh dát, s ktorými sa v metagenomike pracuje.

V druhej časti práce bol popísaný problém klasifikácie metagenomických dát a zároveň boli predstavené už existujúce metódy klasifikácie metagenomických dát nezávislých na referenčnej databáze, spolu s ich postupným vývojom.

Ďalej bola predstavená technológia sekvenácie tretej generácie a jej výhody a rozdiely od sekvenácie druhej generácie. Jedným z najpopulárnejších sekvenátorov tretej generácie je práve sekvenátor MinION od spoločnosti Oxford Nanopore, ktorý je stredobodom tejto práce. Preto je ďalej detailnejšie vysvetlený princíp jeho funkcie a rovnako sú popísané výstupné dáta tejto platformy.

Po popise výstupných dát sekvenátora MinION, je navrhnutá vlastná metóda klasifikácie metagenomických dát, ktorá na klasifikáciu metagenomických signálov zo sekvenácie nanopórom využíva reprezentáciu pomocou deskriptorov Poincarého máp a na samotnú klasifikáciu využíva zhlukovaciu metódu DBSCAN.

Výber zhlukovacej metódy je v predposlednej kapitole 5.1 odôvodnený porovnaním výsledkov klasifikácie zhlukovacých metód DBSCAN a k-means, kde sa metóda DBSCAN ukázala ako efektívnejšia, vzhľadom k povahe dát.

Na vyhodnotenie efektivity klasifikácie boli použité parametry Accuracy, Recall a Precision, kedy navrhnutá metóda dosiahla vysokej efektivity Accuracy = 89,04%, Precision = 88,13% a Recall = 87,28%. Pre lepší prehľad sú dané parametre v tabuľke Tab. 2 uvedené aj pre každý jeden organizmus zvlášť. Na záver práce bola potom navrhnutá metóda porovnaná s už existujúcou metódou klasifikácie VizBin a dosiahla porovnateľných výsledkov a nie je ovplyvnená užívateľským vstupom. Keďže ale navrhnutá metóda pracuje priamo so squiggle signálmi, je robustná voči chybám, ktoré môžu nastať pri procese basecallingu a pri prevode dát z formátu FASTQ na FASTA a z toho istého dôvodu je aj uživateľsky priatelnejšia. Na základe skúseností z používania nástroja VizBin boli potom poskytnuté také výstupné dáta, ktoré sú oveľa exaktnejšie a presnejšie ako výstupné dáta nástroja VizBin.

LITERATÚRA

[1] National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Washington (DC): National Academies Press (US); 2007. 1, Why Metagenomics? Dostupné také z: https://www.ncbi.nlm.nih.gov/books/NBK54011/

[2] HANDELSMAN, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiology and Molecular Biology Reviews. 2004, 68(4), 669-685. DOI: 10.1128/MMBR.68.4.669-685.2004. ISSN 1092-2172. Dostupné také z: http://mmbr.asm.org/cgi/doi/10.1128/MMBR.68.4.669-685.

[3] Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease, Sharon Greenblum, Peter J. Turnbaugh, Elhanan Borenstein, Proceedings of the National Academy of Sciences Jan 2012, 109 (2) 594-599; DOI:10.1073/pnas.1116053109, Dostupné také z: http://www.pnas.org/content/109/2/594.full

[4] Ott SJ, Musfeldt M, Wenderoth DF*, et al.* Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut* 2004;**53:**685-693. Dostupné také z: https://gut.bmj.com/content/53/5/685

[5] QI, Bin a Min HAN. Microbial Siderophore Enterobactin Promotes Mitochondrial Iron Uptake and Development of the Host via Interaction with ATP Synthase. Cell. 2018, 175(2), 571-582.e11. DOI: 10.1016/j.cell.2018.07.032. ISSN 00928674. Dostupné také z: https://linkinghub.elsevier.com/retrieve/pii/S0092867418309590

[6] Solanki, M.K.; Li, C.; Wang, F.; Wang, Z.; Lan, T.; Singh, R.K.; Singh, P.; Yang, L.; Li, Y. Linkages of Soil Nutrients and Diazotrophic Microbiome under Sugarcane-Legume Intercropping. Preprints 2018, 2018100382 (doi: 10.20944/preprints201810.0382.v1), Dostupné také z: https://www.preprints.org/manuscript/201810.0382/v1

[7] SOGIN, M. L., H. G. MORRISON, J. A. HUBER, D. M. WELCH, S. M. HUSE, P. R. NEAL, J. M. ARRIETA a HERNDL. Proceedings of the National Academy of Sciences. 2006, 103(32). DOI: 10.1073/pnas.0605127103. ISSN 0027-8424. Dostupné také z: http://www.pnas.org/cgi/doi/10.1073/pnas.0605127103

[8] REUTER, JASON A., SPACEK, DAMEK V. and SNYDER, MICHAEL P., 2015, High-Throughput Sequencing Technologies. Molecular Cell. 2015. Vol. 58, no. 4, p. 586-597. DOI 10.1016/j.molcel.2015.05.004. Elsevier BV Dostupné také z: https://www.sciencedirect.com/science/article/pii/S1097276515003408

[9] SEDLAR, Karel, Kristyna KUPKOVA a Ivo PROVAZNIK. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Computational and Structural Biotechnology Journal, 2016, 15(1), 48-55 Dostupné také z: https://www.sciencedirect.com/science/article/pii/S2001037016300678

[10] CLARRIDGE, J. E. Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. Clinical Microbiology Reviews. 2004, 17(4), 840-862. DOI: 10.1128/CMR.17.4.840-862.2004. ISSN 0893-8512. Dostupné také z: http://cmr.asm.org/cgi/doi/10.1128/CMR.17.4.840-862.2004

[11] SANTAMARIA, M., FOSSO, B., CONSIGLIO, A., DE CARO, G., GRILLO, G., LICCIULLI, F., LIUNI, S., MARZANO, M., ALONSO-ALEMANY, D., VALIENTE, G. and PESOLE, G., 2012, Reference databases for taxonomic assignment in metagenomics. Briefings in Bioinformatics. 2012. Vol. 13, no. 6, p. 682-695. DOI 10.1093/bib/bbs036. Oxford University Press (OUP) Dostupné také z: https://www.ncbi.nlm.nih.gov/pubmed/22786784

[12] TURNBAUGH, PETER J. and GORDON, JEFFREY I., 2008, An Invitation to the Marriage of Metagenomics and Metabolomics. Cell. 2008. Vol. 134, no. 5, p. 708-713. DOI 10.1016/j.cell.2008.08.025. Elsevier BV, Dostupné také z: https://www.cell.com/cell/fulltext/S0092-8674(08)01070-2?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867408010702%3Fshowall%3Dtrue

[13] SUN, Y., CAI, Y., HUSE, S. M., KNIGHT, R., FARMERIE, W. G., WANG, X. and MAI, V., 2011, A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Briefings in Bioinformatics. 2011. Vol. 13, no. 1, p. 107-121. DOI 10.1093/bib/bbr009. Oxford University Press (OUP) Dostupné také z: https://academic.oup.com/bib/article/13/1/107/218565

[14] BOOTHROYD, Peter a Xuân Nam PHẠM. Socioeconomic renovation in Viet Nam: the origin, evolution, and impact of doi moi. Clinical Microbiology Reviews. 2004, 2000, 17(4), 840-862. DOI: doi.org/10.1186/gb-2009-10-8-r85. ISSN 0893-8512. Dostupné také z: http://cmr.asm.org/cgi/doi/10.1128/CMR.17.4.840-862.2004

[15] KISLYUK, ANDREY, BHATNAGAR, SRIJAK, DUSHOFF, JONATHAN and WEITZ, JOSHUA S, 2009, Unsupervised statistical clustering of environmental shotgun sequences. BMC Bioinformatics. 2009. Vol. 10, no. 1, p. 316. DOI 10.1186/1471-2105-10-316. Springer Nature Dostupné také z: https://bmcbioinformatics.biomedcentral.com/track/pdf/10.1186/1471-2105-10-316

[16] PRIDE, D. T., 2003, Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases. Genome Research. 2003. Vol. 13, no. 2,

p. 145-158. DOI 10.1101/gr.335003. Cold Spring Harbor Laboratory Dostupné také z: https://genome.cshlp.org/content/13/2/145.short

[17] KELLEY, DAVID R and SALZBERG, STEVEN L, 2010, Clustering metagenomic sequences with interpolated Markov models. BMC Bioinformatics. 2010. Vol. 11, no. 1, p. 544. DOI 10.1186/1471-2105-11-544. Springer Nature, Dostupné také z: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-544

[18] KOHONEN, T. The self-organizing map. Proceedings of the IEEE. 78(9), 1464-1480. DOI: 10.1109/5.58325. ISSN 00189219. Dostupné také z: http://ieeexplore.ieee.org/document/58325/

[19] VAN DER MAATEN, Laurens. Barnes-hut-sne. arXiv preprint arXiv:1301.3342, 2013, Dostupné také z: https://arxiv.org/abs/1301.3342

[20] SAEED, ISAAM, TANG, SEN-LIN and HALGAMUGE, SAMAN K., 2011, Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. Nucleic Acids Research. 2011. Vol. 40, no. 5, p. e34-e34. DOI 10.1093/nar/gkr1204. Oxford University Press (OUP) Dostupné také z: https://academic.oup.com/nar/article/40/5/e34/1101400

[21] SAEED, ISAAM and HALGAMUGE, SAMAN K, 2009, The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. BMC Genomics. 2009. Vol. 10, no. Suppl 3, p. S10. DOI 10.1186/1471-2164-10-s3-s10. Springer Nature Dostupné také z: https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-S3-S10#Sec5

[22] WU, YU-WEI and YE, YUZHEN, 2011, A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-tuples. Journal of Computational Biology. 2011. Vol. 18, no. 3, p. 523-534. DOI 10.1089/cmb.2010.0245. Mary Ann Liebert Inc, Dostupné také z: https://www.liebertpub.com/doi/10.1089/cmb.2010.0245

[23] MOON, T.K. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*. **13**(6), 47-60. DOI: 10.1109/79.543975. ISSN 10535888. Dostupné také z: http://ieeexplore.ieee.org/document/543975/

[24] WANG, YING, HU, HAIYAN and LI, XIAOMAN, 2015, MBBC: an efficient approach for metagenomic binning based on clustering. BMC Bioinformatics. 2015. Vol. 16, no. 1. DOI 10.1186/s12859-015-0473-8. Springer Nature Dostupné také z: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0473-8

[25] WANG, Y., LEUNG, H. C. M., YIU, S. M. and CHIN, F. Y. L., 2012, MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. Bioinformatics. 2012. Vol. 28, no. 18, p. i356-i362. DOI 10.1093/bioinformatics/bts397. Oxford University Press(OUP), Dostupné také z: https://academic.oup.com/bioinformatics/article/28/18/i356/248051

[26] CHATTERJI, SOURAV, YAMAZAKI, ICHITARO, BAI, ZHAOJUN and EISEN, JONATHAN, 2018, CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. Arxiv.org [online]. 2018. [Accessed 17 November 2018]. Dostupné také z:: https://arxiv.org/abs/0708.3098

[27] ALNEBERG, JOHANNES, BJARNASON, BRYNJAR SMÁRI, DE BRUIJN, INO, SCHIRMER, MELANIE, QUICK, JOSHUA, IJAZ, UMER Z, LAHTI, LEO, LOMAN, NICHOLAS J, ANDERSSON, ANDERS F and QUINCE, CHRISTOPHER, 2014, Binning metagenomic contigs by coverage and composition. Nature Methods. 2014. Vol. 11, no. 11, p. 1144-1146. DOI 10.1038/nmeth.3103. Springer Nature, Dostupné také z: https://www.nature.com/articles/nmeth.3103

[28] LIN, HSIN-HUNG and LIAO, YU-CHIEH, 2016, Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Scientific Reports. 2016. Vol. 6, no. 1. DOI 10.1038/srep24175. Springer Nature, Dostupné také z: https://www.nature.com/articles/srep24175

[29] FREY, B. J. and DUECK, D., 2007, Clustering by Passing Messages Between Data Points. Science. 2007. Vol. 315, no. 5814, p. 972-976. DOI 10.1126/science.1136800. American Association for the Advancement of Science (AAAS), Dostupné také z: http://science.sciencemag.org/content/315/5814/972

[30] KULTIMA, JENS ROAT, SUNAGAWA, SHINICHI, LI, JUNHUA, CHEN, WEINENG, CHEN, HUA, MENDE, DANIEL R., ARUMUGAM, MANIMOZHIYAN, PAN, QI, LIU, BINGHANG, QIN, JUNJIE, WANG, JUN and BORK, PEER, 2012, MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. PLoS ONE. 2012. Vol. 7, no. 10, p. e47656. DOI 10.1371/journal.pone.0047656. Public Library of Science (PLoS), Dostupné také z: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0047656

[31] KANG, DONGWAN D., FROULA, JEFF, EGAN, ROB and WANG, ZHONG, 2015, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015. Vol. 3, p. e1165. DOI 10.7717/peerj.1165. PeerJ, Dostupné také z: https://peerj.com/articles/1165/

[32] IMELFORT, MICHAEL, PARKS, DONOVAN, WOODCROFT, BEN J., DENNIS, PAUL, HUGENHOLTZ, PHILIP and TYSON, GENE W., 2014, GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014. Vol. 2, p. e603. DOI 10.7717/peerj.603. PeerJ, Dostupné také z: https://peerj.com/articles/603/

[33] BEHJATI, Sam a Patrick S TARPEY. *What is next generation sequencing?*. 2013, **98**(6), 236-238. DOI: 10.1136/archdischild-2013-304340. ISSN 1743-0585. Dostupné také z: http://ep.bmj.com/lookup/doi/10.1136/archdischild-2013-304340

[34] SHIN, SEUNG CHUL, AHN, DO HWAN, KIM, SU JIN, LEE, HYOUNGSEOK, OH, TAE-JIN, LEE, JONG EUN and PARK, HYUN, 2013, Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. PLoS ONE. 2013. Vol. 8, no. 7, p. e68824. DOI 10.1371/journal.pone.0068824. Public Library of Science (PLoS), Dostupné také z: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068824

[35] MCGINN, STEVEN and GUT, IVO GLYNNE, 2013, DNA sequencing – spanning the generations. *New Biotechnology*. 2013. Vol. 30, no. 4, p. 366-372. DOI 10.1016/j.nbt.2012.11.012. Elsevier BV, Dostupné také z: https://www.sciencedirect.com/science/article/pii/S1871678412008692

[36] SCHADT, E. E., TURNER, S. and KASARSKIS, A., 2010, A window into third-generation sequencing. Human Molecular Genetics. 2010. Vol. 19, no. R2, p. R227-R240. DOI 10.1093/hmg/ddq416. Oxford University Press (OUP), Dostupné také z: https://academic.oup.com/hmg/article/19/R2/R227/641295

[37] HARRIS, T. D., P. R. BUZBY, H. BABCOCK, et al. Single-Molecule DNA Sequencing of a Viral Genome. *Science*. 2008, **320**(5872), 106-109. DOI: 10.1126/science.1150427. ISSN 0036-8075. Dostupné také z: http://www.sciencemag.org/cgi/doi/10.1126/science.1150427

[38] EID, J., A. FEHR, J. GRAY, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*. 2009, **323**(5910), 133-138. DOI: 10.1126/science.1162986. ISSN 0036-8075. Dostupné také z: http://www.sciencemag.org/cgi/doi/10.1126/science.1162986

[39] KRIVANEK, ONDREJ L., CHISHOLM, MATTHEW F., NICOLOSI, VALERIA, PENNYCOOK, TIMOTHY J., CORBIN, GEORGE J., DELLBY, NIKLAS, MURFITT, MATTHEW F., OWN, CHRISTOPHER S., SZILAGYI, ZOLTAN S., OXLEY, MARK P., PANTELIDES, SOKRATES T. and PENNYCOOK, STEPHEN J., 2010, Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy. Nature. 2010. Vol. 464, no. 7288, p. 571-574. DOI 10.1038/nature08879. Springer Nature Dostupné také z: https://www.nature.com/articles/nature08879

[40] LU, HENGYUN, GIORDANO, FRANCESCA and NING, ZEMIN, 2016, Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics & Bioinformatics. 2016. Vol. 14, no. 5, p. 265-279. DOI 10.1016/j.gpb.2016.05.004. Elsevier BV, Dostupné také z: https://www.sciencedirect.com/science/article/pii/S1672022916301309

[41] GÖPFRICH, KERSTIN and JUDGE, KIM, 2018, Decoding DNA with a pocket-sized sequencer | www.scienceinschool.org. Scienceinschool.org [online]. 2018. [Accessed 25 November 2018]. Dostupné také z: https://www.scienceinschool.org/content/decoding-dna-pocket-sized-sequencer

[42] LAVER, T., HARRISON, J., O'NEILL, P.A., MOORE, K., FARBOS, A., PASZKIEWICZ, K. and STUDHOLME, D.J., 2015, Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular Detection and

Quantification. 2015. Vol. 3, p. 1-8. DOI 10.1016/j.bdq.2015.02.001. Elsevier BV, Dostupné také z: https://www.sciencedirect.com/science/article/pii/S2214753515000224

[43] RANG, Franka J., Wigard P. KLOOSTERMAN a Jeroen DE RIDDER. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*. 2018, **19**(1). DOI: 10.1186/s13059-018-1462-9. ISSN 1474-760X. Dostupné také z: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1462-9

[44] JAIN, MITEN, OLSEN, HUGH E., PATEN, BENEDICT and AKESON, MARK, 2016, The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology. 2016. Vol. 17, no. 1. DOI 10.1186/s13059-016-1103-0. Springer Nature, Dostupné také z: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1103-0

[45] IP, Camilla L.C., Matthew LOOSE, John R. TYSON, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*. 2015, **4**. DOI: 10.12688/f1000research.7201.1. ISSN 2046-1402. Dostupné také z: https://f1000research.com/articles/4-1075/v1

[46] MAGI, Alberto, Roberto SEMERARO, Alessandra MINGRINO, Betti GIUSTI a Romina D'AURIZIO. Nanopore sequencing data analysis: state of the art, applications and challenges. *Briefings in Bioinformatics*. 2017. DOI: 10.1093/bib/bbx062. ISSN 1467-5463. Dostupné také z: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx062

[47] BROWN, Clive. *Oxford Nanopore Technologies "...a wafer thin update"* [online]. 09.2016 [cit. 2019-05-09]. Dostupné z: https://nanoporetech.com/resource-centre/wafer-thin-update

[48] COCK, PETER J. A., FIELDS, CHRISTOPHER J., GOTO, NAOHISA, HEUER, MICHAEL L. and RICE, PETER M., 2009, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research. 2009. Vol. 38, no. 6, p. 1767-1771. DOI 10.1093/nar/gkp1137. Oxford University Press (OUP), Dostupné také z: https://academic.oup.com/nar/article/38/6/1767/3112533

[49] BROWN, Bonnie L., Mick WATSON, Samuel S. MINOT, Maria C. RIVERA a Rima B. FRANKLIN. MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. GigaScience. 2017, 6(3). DOI: 10.1093/gigascience/gix007. ISSN 2047-217X. Dostupné také z: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/gix007/3051932

[50] WATSON, MICK, THOMSON, MARIAN, RISSE, JUDITH, TALBOT, RICHARD, SANTOYO-LOPEZ, JAVIER, GHARBI, KARIM and BLAXTER, MARK, 2014, poRe: an R package for the visualization and analysis of nanopore sequencing data. Bioinformatics. 2014. Vol. 31, no. 1, p. 114-115.

DOI 10.1093/bioinformatics/btu590. Oxford University Press (OUP), Dostupné také z: https://academic.oup.com/bioinformatics/article/31/1/114/2365693

[51] GOLIŃSKA, Agnieszka Kitlas. Poincaré Plots in Analysis of Selected Biomedical Signals. *Studies in Logic, Grammar and Rhetoric*. 2013, **35**(1), 117-127. DOI: 10.2478/slgr-2013-0031. ISSN 0860-150X. Dostupné také z: http://content.sciendo.com/view/journals/slgr/35/1/article-p117.xml

[52] MARCIANO, F., M.L. MIGAUX, D. ACANFORA, G. FURGI a F. RENGO. Quantification of Poincare' maps for the evaluation of heart rate variability. *Computers in Cardiology 1994*. IEEE Comput. Soc. Press, 1995, , 577-580. DOI: 10.1109/CIC.1994.470126. ISBN 0-8186-6570-X. Dostupné také z: http://ieeexplore.ieee.org/document/470126/

[53] SCHUBERT, ERICH, SANDER, JÖRG, ESTER, MARTIN, KRIEGEL, HANS PETER and XU, XIAOWEI, 2017, DBSCAN Revisited, Revisited. ACM Transactions on Database Systems. 2017. Vol. 42, no. 3, p. 1-21. DOI 10.1145/3068335. Association for Computing Machinery (ACM), Dostupné také z: https://dl.acm.org/citation.cfm?doid=3129336.3068335

[54] DBSCAN. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2019-04-26]. Dostupné z: https://en.wikipedia.org/wiki/DBSCAN

[55] Yarpiz. *DBSCAN Clustering Algorithm: Implementation of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) in MATLAB* [online]. 06.09.2015 [cit. 2019-05-13]. Dostupné z: https://www.mathworks.com/matlabcentral/fileexchange/52905-dbscan-clustering-algorithm

SEZNAM PŘÍLOH

1. Zdrojový kód navrhnutej metódy
2. Zdrojový kód analýzy motódy VizBin