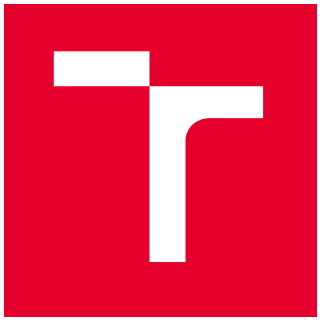
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2022 Veronika Plocková



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

VLASTNOSTI PROUDOVÝCH SIGNÁLŮ PŘI SEKVENACI NANOPÓREM

PROPERTIES OF CURRENT SIGNALS IN NANOPORE SEQUENCING

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — **Veronika Plocková**

AUTHOR

VEDOUCÍ PRÁCE — **Mgr. Ing. Karel Sedlář, Ph.D.**

SUPERVISOR

BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Veronika Plocková ***ID:*** 211575

Ročník: 3 ***Akademický rok:*** 2021/22

NÁZEV TÉMATU:

Vlastnosti proudových signálů při sekvenaci nanopórem

POKYNY PRO VYPRACOVÁNÍ:

1) Prostudujte princip sekvenace technologií Oxford Nanopore Technologies (ONT), zaměřte se na vlastnosti proudových signálů, tzv. squiggles, a principy dekódování jednotlivých bází DNA z těchto signálů. 2) Vypracujte literární rešerši o různých sekvenačních kitech a čipech, které lze s ONT použít. 3) Vytvořte vhodný testovací dataset, který bude obsahovat sekvenační data z různých organismů a při použití různých sekvenačních čipů a kitů. Signály v datasetu zhodnoťte několika vhodně zvolenými parametry pro popis signálů. 4) Najděte vhodný parametr, který dokáže rozlišit různé signály mezi sebou, a proveďte shlukování signálů. 5) Výsledky shlukování signálů porovnejte s výsledky shlukové analýzy provedené na dekódovaných signálech, tedy znakových sekvencích. 6) Rozdíly mezi různými přístupy diskutujte.

DOPORUČENÁ LITERATURA:

[1] LOMAN, N. J. a A. R. QUINLAN. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics. 2014, 30(23), 3399-3401.

[2] BAO, Yuwei, Jack WADDEN, John R ERB-DOWNWARD, Piyush RANJAN, Robert P DICKSON, David BLAAUW a Joshua D WELCH. Real-Time, Direct Classification of Nanopore Signals with SquiggleNet. bioRxiv [online]. 2021, 2021.01.15.426907.

Termín zadání: 7.2.2022 ***Termín odevzdání:*** 27.5.2022

Vedoucí práce: Mgr. Ing. Karel Sedlář, Ph.D.

doc. Ing. Jana Kolářová, Ph.D.
předseda rady studijního programu

ABSTRAKT

Technologie Oxford Nanopore přinesly novou a revoluční technologii v oblasti sekvenování DNA. Jejich sekvenační zařízení měří změny elektrického proudu protékajícího póry spolu s DNA. Tato práce si klade za cíl popsat rozdíly mezi nezpracovanými signály produkovanými různými sekvenačními soupravami a průtokovými komůrkami při sekvenování několika různých bakterií. K analýze různých statistických parametrů, které by byly vhodné pro popis signálů získaných z nanopórů, byly použity dva soubory dat kombinující pět různých organismů, dva sekvenační kity a dva typy průtokových kyvet. Následně byly vzorky klasifikovány pomocí algoritmu k-means a výsledky byly diskutovány.

KLÍČOVÁ SLOVA

Proudové signály, Oxford Nanopore Technologies, sekvenování, fylogenetická analýza, statistické parametry, k-means.

ABSTRACT

Oxford Nanopore technologies brought new and revolutionary technology in the field of DNA sequencing. Their sequencing device measures changes in the electric current flowing through pores together with DNA. This work aims to describe differences between raw signals produced by various sequencing kits and sequencing flowcells while sequencing several different bacteria. Two datasets combining five different organisms, two sequencing kits, and two types of flowcells were used to analyze various statistical parameters that would be suitable for the description of current signals gathered from nanopores. Finally, the samples were classified using the k-means algorithm and the results were discussed

KEYWORDS

Current signals, Oxford Nanopore Technologies, sequencing, phylogenetic analysis, statistical parameters, k-means.

PLOCKOVÁ, Veronika. *Vlastnosti proudových signálů při sekvenaci nanopórem*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2022, 64 s. Bakalářská práce. Vedoucí práce: Mgr. Ing. Karel Sedlář, Ph.D.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Veronika Plocková
VUT ID autora:	211575
Typ práce:	Bakalářská práce
Akademický rok:	2021/22
Téma závěrečné práce:	Vlastnosti proudových signálů při sekvenaci nanopórem

Prohlašuji, že svou závěrečnou práci jsem vypracovala samostatně pod vedením vedoucí/ho závěrečné 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 autorka uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského 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.

Brno ....

podpis autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucímu bakalářské práce panu Mgr. Ing. Karlu Sedláři, Ph.D. za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Sekvenování nanoporů je aktuálním tématem bioinformatiky, v současnosti je zmiňováno ve velkém množství vědeckých článků. V roce 2014 vydala společnost Oxford Nanopore Technologies (ONT) své první přenosné zařízení pro sekvenování nanoporů a umožnila tak revoluci v sekvenování především díky schopnosti sekvenovat DNA kdykoli a kdekoli bez potřeby laboratoře. Sekvenování nanoporů má potenciál nabídnout nákladově efektivní genotypizaci, vysokou mobilitu pro testování a rychlé zpracování vzorků v reálném čase.

Cílem teoretické části práce je podat obecné informace o sekvenaci, její historii a různými přístupy k sekvenování. Především se pak zaměřuje na seznámení čtenáře se sekvenátory třetí generace, konkrétně sekvenátory od společnosti Oxford Nanopore Technologies (ONT). V práci jsou představeny přístroje společnosti ONT, různé sekvenační chemie, kity a dále pak datové formáty ve kterých jsou data uložena.

Praktická část práce se věnuje především sestavení vhodného datasetu, následného předzpracování dat, filtraci, vlivu předzpracování a filtrace na signály a následnou analýzu. Data, ze kterých jsou vytvořeny celkem dva datasety pocházejí z databáze různých sekvenací na ústavu biomedicínského inženýrství FEKT VUT Brno.

První dataset je tvořen jedním organismem a dále analyzován. Cílem tohoto experimentu je zjistit, zdali se statistické parametry mění pro různé signály stejného organismu či nikoli. Další dataset obsahuje celkem 8 datových sad složených z 5 organismů. Signály byly vybrány tak, aby pokryly co největší množství průtokových komůrek a kitů. Každý signál je popsán vybranými parametry pro popis signálu a zobrazen pomocí krabicových grafů. Nejdříve byl proveden Shapiro-Wilkův test pro zjištění normality dat, dále pak Kruskal-Wallisovým testem byly získány parametry, které jsou vhodné pro odlišení alespoň jednoho organismu od ostatních. Poté byl proveden Tukeyho test kdy byly získány dvojice organismů, které od sebe lze spolehlivě odlišit. Pomocí této analýzy byly vybrány parametry vhodné jako příznaky ke klasifikaci organismů.

Závěrečná část práce je zaměřena na testování různých algoritmů strojového učení a vlivu sekvenačních nástrojů na klasifikaci.

1 Sekvenační techniky

1.1 Historie sekvenování

Prvním krokem k pochopení genetické informace bylo objevení struktury DNA Francisem Crickem a Jamesem Watsonem v roce 1953. To byl pro oblast bioinformatiky a molekulární biologie zásadní objev. Jako první na světě identifikovali stavbu deoxyribonukleové kyseliny, která je nositelem dědičných a biochemických vlastností všech živých organismů[1]. Schopnost měřit, nebo odvozovat takovéto sekvence je nezbytná pro biologický výzkum a napomáhá k pochopení genetické informace. Své uplatnění nachází především v medicíně v oboru diagnostiky nádorových a dědičných onemocnění[2].

Počátky sekvenování sahají do 70. let minulého století, kdy nezávisle na sobě dva vědecké týmy vyvinuly metody Maxam-Gilbert a Sanger. U Sangerovy metody se v průběhu let několikanásobně zvýšila efektivita a přesnost, a to mělo za následek, že se stala nejvyužívanější sekvenační metodou, až do konce 20 století. Sangerova metoda využívá replikace DNA a vlastností nukleotidů ddATP, ddCTP, ddGTP a ddTTP. Ty nemají na 3' uhlíku ribózy OH skupinu, a tudíž na tento konec už nemůže být navázán další nukleotid. Výsledkem sekvenačního procesu jsou 4 směsi oligonukleotidů různých délek, ty jsou poté vyhodnoceny pomocí elektroforézy[**?**]. Nevýhoda této metody je její časová, technická, a především ekonomická náročnost, což přirozeně vedlo k nahrazování jinými metodami, které jsou schopné generovat za kratší časový úsek více dat[4]. Dnes se používá jen k dosekvenování úseků, se kterými mají jiné technologie problém[5].

O třicet let později přicházejí první komerční produkty a svoji pozornost si získávají spíše jména velkých firem než jednotlivých vědců. Společným znakem metod 'Další generace' (z anglického Next Generation) je masivní paralelizace, což má za následek zrychlování celého procesu a snížení nákladů. Mezi nejpoužívanější metodu současnosti je považována metoda Illumina. Jedná se o metodu sekvenování druhé generace. Principem je fragmentace DNA na menší části a připojení adaptérů, namnožení fragmentů pomocí můstkové PCR a jejich paralelnímu sekvenování. Detekuje se fluorescence, ta je dále zpracovávána a převáděna nukleotidové sekvence[6].

Tato bakalářská práce se však věnuje především sekvenátorům třetí generace, konkrétně technologii vyvinutou firmou Oxford Nanopore Technologies, která bude představena v následující kapitole.

1.2 Třetí generace sekvenátorů

Sekvenátory třetí generace se vyznačují zcela novým přístupem k sekvenaci, než dříve vyvinuté, výše zmíněné metody. Hlavním principem je sekvenování jednotlivých molekul DNA. Mezi platformy třetí generace se řadí SMRT (z anglického single molecule real time sequencing) od společnosti PacBio a dále sekvenování pomocí nanopóru od společnosti ONT. Třetí generace se řadí mezi nejnovější na trhu, a proto také stále dochází k vývoji, jsou předmětem dalšího zkoumání a vylepšování.

S myšlenkou, že nukleotid může potenciálně způsobit specifickou blokádu iontového proudu při průchodu kanálem jako první přišel profesor David Deamer již v roce 1989. Tomuto nápadu se však naplno začal věnovat až v roce 1993, kdy se skupinou svých kolegů provedli první experimenty[4].

Společnost ONT se již od svého vzniku v roce 2005 zabývá vývojem přístojů umožňující sekvenování jednotlivých molekul DNA pomocí nanopóru, touto technologií je však možné sekvenovat i proteiny, nebo jiné molekuly. ONT se od svých předchůdců liší hned v několika faktorech. Jedná se o zatím jedinou komerčně dostupnou technologii. Dále jako jediná existující technologie umožňuje sekvenování libovolně dlouhých fragmentů v reálném čase. Dokáže analyzovat nativní DNA, nebo RNA a je oproti svým předchůdcům a konkurentům výrazně levnější a mnohem dostupnější.

1.3 Princip Oxford Nanopore Technologies (ONT)

Princip této technologie je založen na biologických vlastnostech nanopóru, ty mohou být biologické, nebo syntetické. Syntetické nanopóry se vyznačují svojí nízkou cenou a snadnou paralelizací. V důsledku horší kompatibility s modifikujícími enzymy, produkují hůře interpretovatelné a více zašuměné signály. Z toho důvodu jsou i přes své výhody stále více používány póry biologické[7]. Mezi nejpoužívanější biologické nanopóry patří např. alpha-hemolysin. Výhodou biologických nanopórů je možnost manipulace jejich struktury a snadná reprodukce jejich rozměrů. alpha-hemolysin je heptamerní proteinový pór s vnitřním průměrem 1 nm, přibližně 100 000krát menší než průměr lidského vlasu. Důležitou vlastností alpha-hemolysinu je jeho stabilita. Funkčnost zůstává zachována i při teplotách blízkých varu vody[8].

Základem je průtoková komůrka (z anglického flow cell), která obsahuje proteinové nanopóry, uložené v nevodivé membráně ze syntetického polymeru, která je ponořena do iontového roztoku. Každý pór je připojen ke své elektrodě na array čipu, vyrobeného z polovodičového materiálu. Samotný nanopór, jak již bylo výše zmíněno, je tvořen alpha-hemolysinem, nebo jinou, synteticky vyrobenou látkou[9].

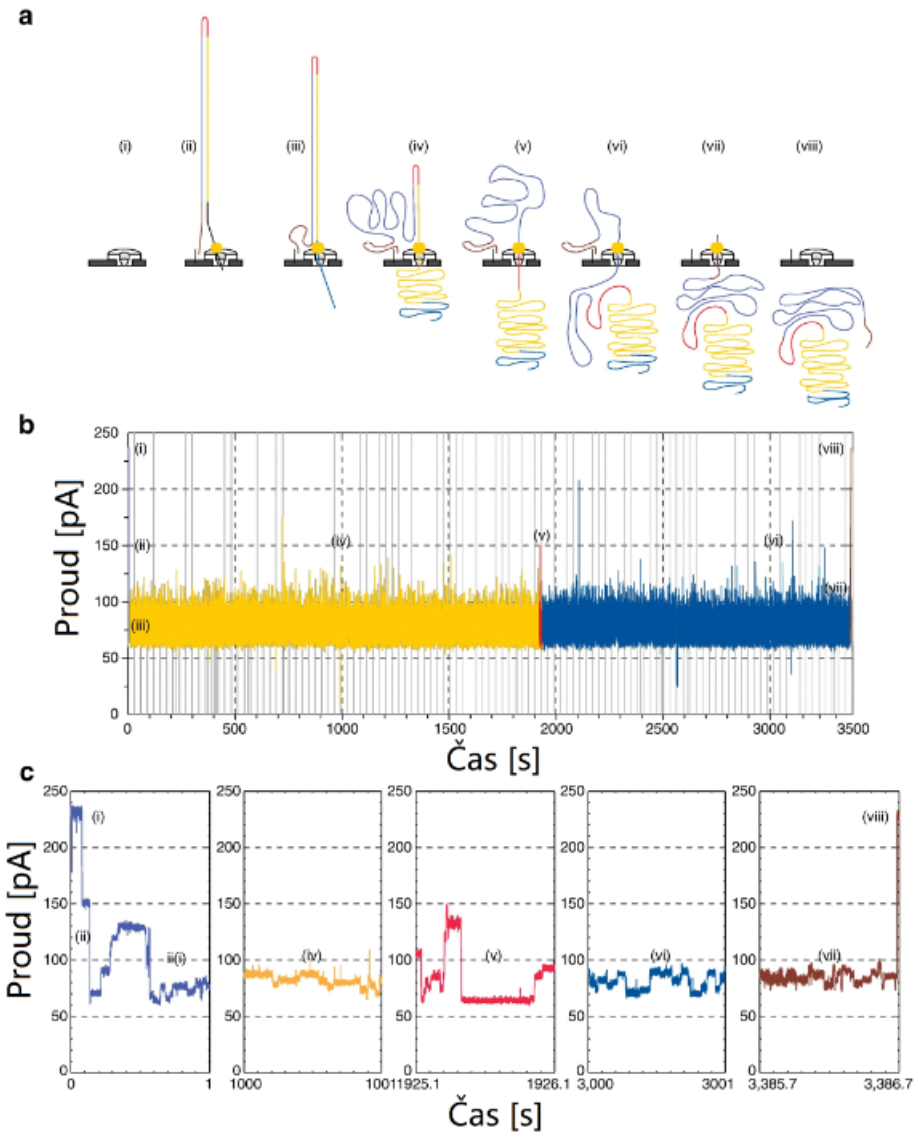
Průtoková komůrka se vkládá do sekvenátoru např. GridION nebo MinION. Toto zařízení je připojeno k počítači přes USB port, kde poté probíhá sekvenace a ná-

sledná analýza dat. Zařízení obsahuje 512 kanálů, kde po vložení vzorku, je na jedné průtokové komůrce schopen sekvenovat 512 různých molekul DNA zároveň. Z toho tedy vyplývá, že každý pór čte jednu sekvenci a generuje jeden signál, představující jedno čtení. Rychlost, jakou projde molekula DNA, určuje helikázový enzym. Ten je schopen translokovat až 450 bází za sekundu. Ne vždy ale tento parametr platí, protože nanopór detekuje molekuly nezávisle na ostatních a aktivita nanopórů se od sebe liší, tudíž výkon jednotlivých kanálů není stejný. Proud iontů protékajících spolu s DNA nanopórem vytvoří napětí, jehož hodnota se liší pro každý k-mer bází[10].

Průběh sekvenování DNA molekuly nanopórem je detailněji zobrazen a popsán na obrázku1.1.

Po aplikaci napětí začne molekula DNA procházet nanopórem a začne generovat iontový proud. Změny proudu odpovídají jednotlivým nukleotidovým bázím. To umožňuje rychlé sekvenování dlouhých jednotlivých molekul DNA, přípravu vzorku bez amplifikace a přímou detekci epigenetických modifikací, jako je metylace bází. Průchod (volného translokačního DNA) by však byl příliš rychlý na to, aby se odlišily změny proudu specifické pro nukleotidy od šumu, z tohoto důvodu bylo nutné její průtok nějak zpomalit. Jednoduché techniky pro snížení rychlosti translokace DNA, jako je snížení experimentální teploty, nebo zvýšení viskozity roztoku snižují aktuální signál, nejedná se tedy o vhodné řešení tohoto problému. Jako nejvhodnější způsob se ukázalo využití polymeráz. K aktivaci polymerázy může docházet předčasně, zabránit tomu lze použitím syntetického oligonukleotidu, který je komplementární k templátovému vláknu. Po aplikaci napětí dojde k 'rozepnutí' blokujícího oligonukleotidu a zahájení syntézy[11].

Příliš dlouhý sekvenační běh může způsobit ucpání póru. Nabízí se tedy možnost nevyužívat průtokovou komůrku ve svém plném rozsahu, ale čtení přerušit a póry propláchnout roztokem.



Obr. 1.1: Sekvenování nanopórem MinION. a) Průchod vlákna DNA nanopórem: (i) Prázdný, otevřený nanopór. (ii) dsDNA začíná prostupovat s vedoucím vláknem (modrá), je zachycena nanopóry; po zachycení následuje translokace. (iii) Následuje jej templátové vlákno (oranžová). (iv) Dále prostupuje vlásenkový adaptér (červená). (v) Dále prostupuje komplementární vlákno. (vi) Ukončení sekvenace koncovým adaptérem (hnědá). (vii) Po ukončení sekvenace se nanopór vrátí do počátečního stavu. b) Surový proud. c) Záznamy surového proudu pro konkrétní příklady (i-vii) Převzato z:[12]

1.4 Vlastnosti proudových signálů

Zařízení společnosti ONT fungují tak, že měří změny iontovového proudu (tzv. squiggles) při průchodu pórem. Změny elektrického proudu indukované molekulou DNA, nebo RNA závisí na specifických chemických vlastnostech nukleotidů, včetně interakcí sekundární struktury a epigenetických modifikací, jako je např. methylace. Nanopórový sekvenátor je schopen sekvenovat pouze část molekuly (v angličtině označované jako ON-demand) a tím umožňuje cílené sekvenování.

1.5 Přiřazování bází

Po sekvenaci jsou signály uloženy ve formátu FAST5. Nejnovější verze pórů tzv. R9 signály ukládá jako posloupnost čísel. U starších verzí nanopórů, jako je třeba chemie R7 se signál nezapisuje posloupností čísel, ale popisuje tzv. události, které obsahují informaci o začátku a konci sekvenace daného fragmentu, směrodatné odchylce a průměrné hodnotě. V současné době se s tímto typem zápisu není možné běžně setkat, protože byl plně nahrazen chemií R9 a i ta je nyní postupně nahrazována vylepšenou chemií R10. Důvodem nahrazení tohoto způsobu popisu, byla nutnost rekonstrukce signálu, kdy docházelo ke ztrátě důležitých informací[13].

Aby bylo možné signály analyzovat je třeba přeložit tuto informaci do znakové podoby. Jelikož se nejedná o triviální úkol, vyskytuje se zde např. oproti jiným sekvenačním technikám, jako je Illumina výrazně vyšší chybovost. Ta se v závislosti na zvoleném nástroji pohybuje mezi 5 až 15% [14]To může být způsobeno neschopností nástrojů správně detekovat báze, vysokou mírou šumu v signálech, stejně jako nízkému poměru signálu k šumu[15]. K nízkému poměru signálu k šumu může přispět nerovnoměrná rychlost průchodu nukleotidů nanopórem, strukturní podobnost nukleotidů a vlivem více nukleotidů na výsledný signál[16]. Ke zlepšení kvality čtení taktéž přispěla možnost sekvenování templátových i komplementárních vláken zároveň bez nutnosti fyzické ligace nazývaných jako $1D^2$. Zrcadlené čtení je poté dekódováno zároveň a tím jsou opraveny chyby v sekvenování[7]. Oproti tomu 1D čtení funguje na principu rozbalení DNA a následnému čtení jednoho vlákna[16].

Protože změny proudu jsou pro jednotlivé báze příliš malé a nanopór je schopen pojmout současně k- nukleotidů. Nejčastěji se jedná o využívá se 5- nebo 6-mer[17]. Aktuální změny tedy ukazují k-mery, které procházejí nanopory. Taková aktuální měření lze použít k identifikaci sekvencí bází řetězců DNA/RNA[18].

V současnosti je k dispozici mnoho nástrojů umožňující rozkódování surového signálu. Mezi nejpoužívanější patří Guppy[19]. Jedná se o nástroj poskytovaný společností ONT. Poskytuje dvě architektury, jedna využívající GPU, jejíž přesnost je

okolo 90%-96%. Hlavní nevýhodou je výpočetní náročnost a nemožné použití v reálném čase. Ten k rozkódování signálu využívá hluboké rekurentní neuronové sítě (RNN). Ty je třeba naučit na reálných datech, v důsledku toho je ovlivněn daty použitými k trénování, které mohou obsahovat modifikace[17].

1.6 Read-until algoritmus

Read Until algoritmus umožňuje cílené sekvenování, to znamená, že je schopen v reálném čase data analyzovat a odmítat fragmenty DNA, která považuje za neinformační. To je výhodné, hledáme-li pouze nějakou část genu, jako je například gen pro rezistenci bakterií k vybraným druhům antibiotik. K tomu už zapotřebí znalost přípravy primerů a knihovny, navíc PCR zesiluje fragmenty a pro vlákno je tedy těžší projít pórem. Fragmenty DNA jsou fyzicky vysunuty z póru změnou polarizace na velmi krátkou dobu. Jedním z algoritmů ze skupiny 'Read Until' je SquiggleKit. Jedná se o nástroj využívající neuronové sítě a konvoluci. Díky tomu je schopen klasifikovat molekuly, které nás zajímají, na základě statistických vzorců vodivosti nanopórů[20].

1.7 Sekvenační platformy ONT

V této kapitole budou představeny různé sekvenační platformy a průtokové komůrky společnosti ONT. Vzhledem k rozdílnosti požadavků na sekvenaci společnost Oxford Nanopore Technologies vyvinula několik druhů sekvenačních platofem, které jsou uvedeny níže.

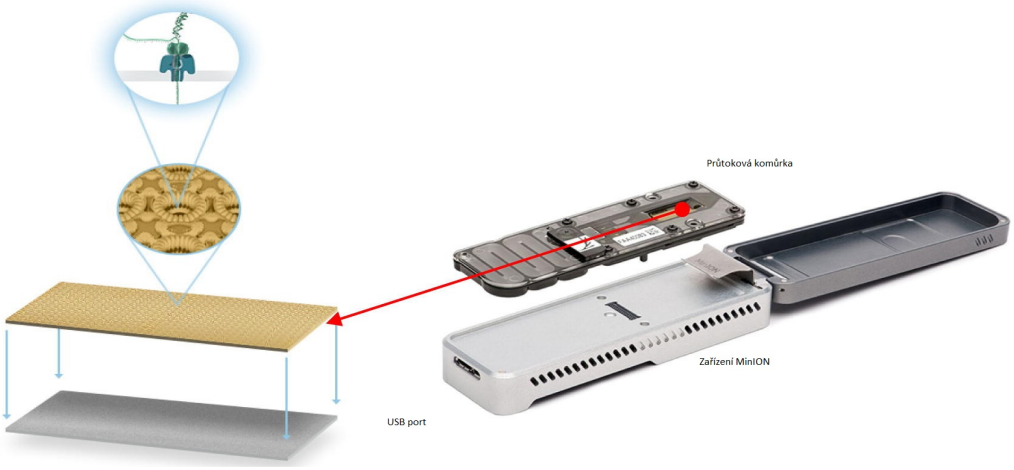
1.7.1 Průtokové komůrky

MinION

Průtoková komůrka MinION je kompatibilní se sekvenačními přístroji společnosti ONT. Jedná se o spotřební zařízení používající chemii R10, která postupně nahradila chemii R9, ta dokáže při správném použití generovat data s přesností až 99% Zařízení se skládá s ASIC obvodu a polem senzorů. Data jsou zaznamenávána v reálném čase a běh může být ukončen, kdykoli uživatel uzná za vhodné. Maximální doba běhu je 72 hodin, nebo do doby ucpání póru[4].

PromethION

PromethION je průtoková komůrka kompatibilní se zařízeními PromethION, standartně dodávána v balíčku po čtyřech. Průtoková komůrka PromethION obsahuje



Obr. 1.2: Sekvenační zařízení MinION s průtokovou komůrkou MinION. Převzato z [22]

3000 nanopórů a její životnost se pohybuje v průměru okolo 64 hodin[21].

Flonge

Flongle je jednorázový adaptér pro přenosná sekvenční zařízení MinION a stolní GridION X5. Využívá stejnou jádrovou technologii sekvenování nanopórů jako ostatní zařízení Oxford Nanopore, produkující data v reálném čase, přímé sekvenování a dlouhé čtení. Zároveň je velmi cenově dostupný, a tudíž umožňuje sekvenování v jakémkoli prostředí, od vědecké laboratoře, továrny až po práci v terénu. Pracovní postup je rychlý a jednoduchý, příprava knihovny zabere pouhých 10 minut. Malé průtokové cely Flongle mohou v současné době produkovat až 1,8 Gb sekvenčních dat. To umožňuje mnoho typů experimentů sekvenování, včetně sekvenování malých genomů, cílené sekvenování amplikonů, metagenomika pro identifikaci virů, bakterií nebo charakterizaci mikrobiomů nebo ID druhu. Může být také použit pro kontrolu kvality pro větší experimenty s nanopory[4][23].

1.7.2 Sekvenační přístroje

MinION

Jedná se o komerčně dostupné kapesní zařízení od společnosti ONT spojující výkonné sekvencování v reálném čase s přenosností. Hmotnost tohoto zařízení se pohybuje okolo 87 g a přes svoje malé rozměry je schopno sekvenovat čtení delší než 4mb s maximálním výkonem 420 bází za sekundu. K počítači je připojeno přes USB port a je kompatibilní s průtokovými komůrkami MinION, ale i Flonge[4].

GridION

GridION je jedno z nejmodernějších zařízení pro sekvenaci nanopórem. Jedná se o velmi schopný počítač s úložnou pamětí a možností analyzovat vzorky už během sekvenace, ve kterém může běžet 1 až 5 průtokových komůrek MinION, což značně zvyšuje výkon až na 30 miliardu bází za 72 hodin[4].

PromethION

Jedná se o stolní zařízení, které je schopno provozovat 24 až 48 průtokových kyvet zároveň, což poskytuje výtěžky až neuvěřitelných 9600 Gb. Jedná se o vysoce výkonný přístroj, který je schopen poskytovat v reálném čase velký objem dat z počtu zpracovávaných vzorků. Jeho značnou výhodou je, že každý průtokový článek je možné spustit samostatně, tudíž je možné sekvenovat čtení o námi zvolené délce a vyhovuje většině experimentálním požadavkům[21].

1.7.3 Další zařízení

VolTRAX

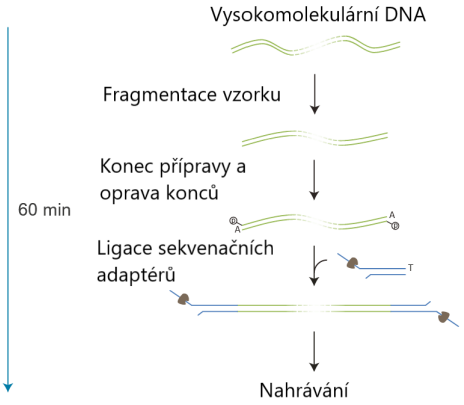
VolTRAX je malé zařízení s 15 vstupními porty určené k provádění automatické přípravy vzorků pro analýzy nanopórů. Díky zahrnutým funkcím mohou uživatelé provést komplexní přípravu knihovny, reverzní transkripci, amplifikaci nebo sekvenování pouze stisknutím tlačítka bez nutnosti dalšího zásahu člověka[24].

1.8 Druhy kitů

Před začátkem sekvenování, je velmi důležitým krokem příprava knihovny. K tomu slouží sady na přípravu knihoven od společnosti Oxford nanopore technologies. Níže jsou představeny některé z nich.

Rapid barcoding kit

Rapid barcoding kit je sada ligačních adaptérů na bázi transposázi, která současně štěpí molekuly templátu a připojuje značky ke štěpeným koncům, pro snadnou, levnou a rychlou přípravu knihovny. Její nevýhodou je menší propustnost a je optimalizovaná spíše pro sekvenování delších fragmentů[25].



Obr. 1.3: Princip Ligation sequencing kitu. Převzato z [25]

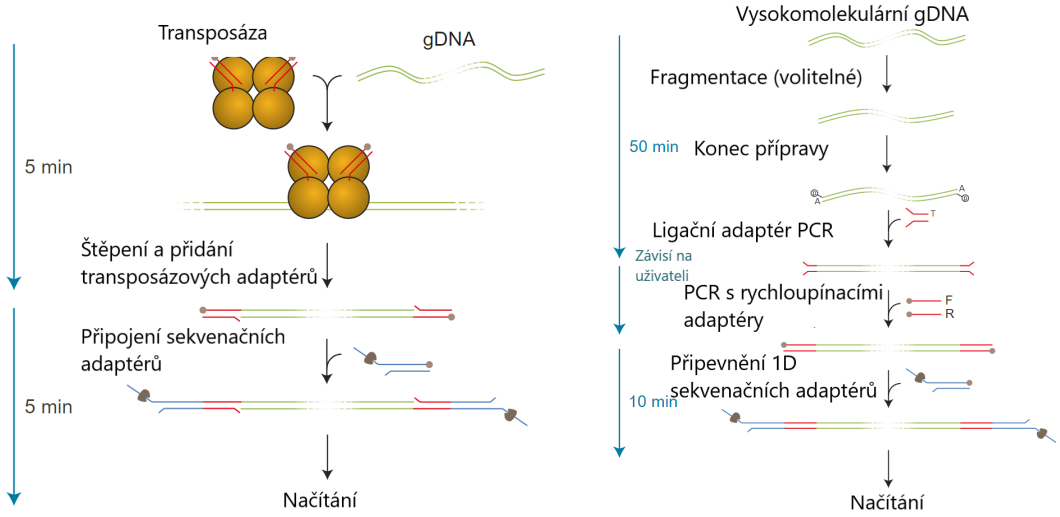
.

Ligation Sequencing kit

Ligation sequencing kit je sada vhodná pro delší čtení a vyznačuje se vysokou propustností a výkonem. Sekvenační knihovna může být připravena z dsDNA (např. gDNA, cDNA nebo amplikonů). Principem je fragmentace dsDNA, opravení konců fragmentů a následné ligování sekvenčních adaptérů na připravené konce [25].

PCR Barcoding Kit

Pokud jsme omezeni množstvím vzorku, chceme sekvenovat pouze vybrané úseky, nebo chceme vzorky multiplexovat, a tím snížit náklady na sekvenaci, je vhodné použít PCR Barcoding kit. Obvyklým postupem je fragmentace vzorů, stříhané konce jsou následně opraveny a poté se na opravené konce ligují PCR adaptéry. Ty obsahují vazebná místra pro primer. Amplifikované a barcodované vzorky se poté spojí dohromady a do směsi se přidají adaptéry pro rychlé sekvenování [25].



Obr. 1.4: Princip PCR sequencing kitu vpravo a Rapid barcoding kit vlevo. Převzato z [25]

.

1.9 Formáty dat používaných při sekvenaci ONT

FAST5

Pochopit, co je uvnitř souboru FAST5, vyžaduje určité porozumění souborům HDF5, protože FAST5 formát je jeho specifikací. HDF5 formáty slouží k ukládání velkých objemů dat a umožňuje ukládat různé typy dat do jednoho souboru. Jedná se o hierarchický formát s vnořenými prvky, což mu dodává obrovskou flexibilitu ve struktuře uložených dat. FAST5 formáty jsou produkovány Nanopore sekvenátory. Ten ukládá naměřený proudový signál, který je výsledkem sekvenování jedné molekuly DNA. Tento datový formát je obsahuje informace o surovém signálu, nebo alespoň data potřebná pro jeho rekonstrukci [26]. Dále také např. informaci o délce jednotlivých čtení, jeji kvalitě, vzorkovací frekvenci a mnoho dalšího. Nevýhodou tohoto

datového typu, je že společnost ONT formát neustále aktualizuje a struktura jednotlivých verzí formátu není konzistentní [27].

FASTQ

Jedná se o textový soubor s příponou .fastq podobný formátu FASTA obsahující data generovaný po přiřazení znakové sekvence signálu z datového souboru FAST5. Narozdíl od FASTA formátu je obohaceno o PHRED skóre. To používá jako zápis znaky tabulky ASCII. První řádek obsahuje hlavičku a začíná znakem'@'. Druhý řádek obsahuje standartně zapsanou sekvenci pomocí IUPAC kódu. Na třetím řádku se nachází pouze znak'+'[28].

FASTA

Formát FASTA je jednoduchý typ formátu, který bioinformatici používají k reprezentaci nukleotidových, nebo proteinových sekvencí. Skládá se z identifikátoru a samotné sekvence. Je napsán v textovém formátu, což umožňuje nástrojům pro zpracování data snadno analyzovat. Její směr je vždy od 5‘ ke 3‘ konci. Obecná přípona souboru je.fasta, ale můžou být použity také přípony .txt nebo .fa [28].

2 Analýza sekvenačních signálů

2.1 Práce se signálem v časové oblasti

Při analýze signálů jsou v podstatě dvě možnosti, jak s nimi pracovat. První a výpočetně méně náročnější je analýza v časové doméně. K té byly vybrány parametry uvedené níže.

Střední hodnota

Střední hodnota (někdy také je používán pojem aritmetický průměr) je určena vztahem, kde N je počet vzorků.

$$S(x) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

Směrodatná odchylka

Směrodatná odchylka ve statistice udává míru rozptylu v souboru hodnost. Pokud je směrodatná odchylka malá, ukazuje to, že hodnoty se blíží průměru, pokud je vysoká, jedná se o hodnoty s vyšším rozsahem. je určena vztahem, kde x je střední hodnota signálu a N je počet vzorků.

$$\sigma(x) = \sqrt{\frac{1}{1-N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2.2}$$

Rozptyl

Rozptyl se vypočítá jako aritmetický průměr čtverců odchylek jednotlivých vzorků od průměru celého signálu. N je počet vzorků.

$$R(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x_i})^2 \tag{2.3}$$

Variační koeficient

Variační koeficient udává míru rozptýlení vzhledem k průměru. S_x je směrodatná odchylka signálu.

$$V(x) = \frac{s_x}{x} \tag{2.4}$$

Koeficient šikmosti

Koeficient šikmosti určuje jsou-li hodnoty kolem zvoleného středu rozloženy souměrně, nebo je-li rozdělení hodnot zešikmeno.

$$\gamma_1(x) = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^3 \tag{2.5}$$

Koeficient špičatosti

Koeficient ostrosti udává míru ostrosti dat. Porovnává se s hodnotou 0. Hodnota větší než 0 označuje vrcholové rozdělení a hodnota menší než 0 označuje ploché rozdělení.

$$\gamma_2(x) = (\frac{1}{N\sigma^4}\sum_{i=1}^{N}(x_i - \mu)^4) - 3 \tag{2.6}$$

Hjorthovy deskriptory

Hjorthovy parametry, někdy popisovány jako deskriptory jsou jeden ze statistických ukazatelů vlastností během zpracování signálů v časové oblasti. Hjorth popsal celkem tři parametry aktivitu, mobilitu a komplexitu. Tyto parametry jsou schopny charakterizovat víceméně jakýkoli signál a jeho deriváty [29]. Hjorthovy parametry jsou odvozeny z následujících rovnic. Kde m_n je moment spektra řádu n, $S(\omega)$ představuje spektrum hustoty výkonu a $f(t)$ je biosignál vyjádřen jako funkce v časové oblasti [30].

$$m_0 = \int_{\infty}^{-\infty} S_{xx}(\omega)d\omega = \frac{1}{T}\int_{t-T}^{t} f^2(t)dt \tag{2.7}$$

$$m_2 = \int_{\infty}^{-\infty} \omega^2 S_{xx}(\omega)d\omega = \frac{1}{T}\int_{t-T}^{t} (\frac{\mathrm{d}f}{\mathrm{d}t})^2 dt \tag{2.8}$$

$$m_4 = \int_{\infty}^{-\infty} \omega^4 S_{xx}(\omega)d\omega = \frac{1}{T}\int_{t-T}^{t} (\frac{\mathrm{d}^2 f}{\mathrm{d}t^2})^2 dt \tag{2.9}$$

Aktivita udává sílu signálu a je definována jako rozptyl signálu. Ve frekvenční oblasti může ukázat plochu výkonového spektra.

$$h_0 = m_0 \tag{2.10}$$

Mobilita je definována jako rozptyl první derivace signálu, kde m_2 představuje spektrální moment 2. řádu. Je uvedena rovnicí:

$$h_1 = \sqrt{\frac{m_2}{m_0}} \tag{2.11}$$

Komplexita představuje rozptyl druhé derivace signálu, kde m_4 je spektrální moment 4. řádu.

$$h_2 = \sqrt{\frac{m_4}{m_2} - \frac{m_2}{m_0}} \tag{2.12}$$

2.2 Tvorba datasetu a statistická analýza

Data pro tuto bakalářskou práci byla poskytnuta z databáze různých sekvenací na ústavu biomedicínského inženýrství FEKT VUT Brno.

Poskytnutá data byla ve formátu FAST5. Z kompletních dat byl vždy vybrán jeden FAST5 soubor reprezentující část čtení daného organismu. Pro následnou analýzu byly vytvořeny dva datasety.

První dataset byl vytvořen z organismu *Clostridium beijerinckii*. Druhý dataset byl složen z celkem osmi vzorků obsahující data z 5 organismů. Organismy byly vybrány tak, aby bylo zastoupeno co největší množství sekvenačních kitů a typů průtokových komůrek.

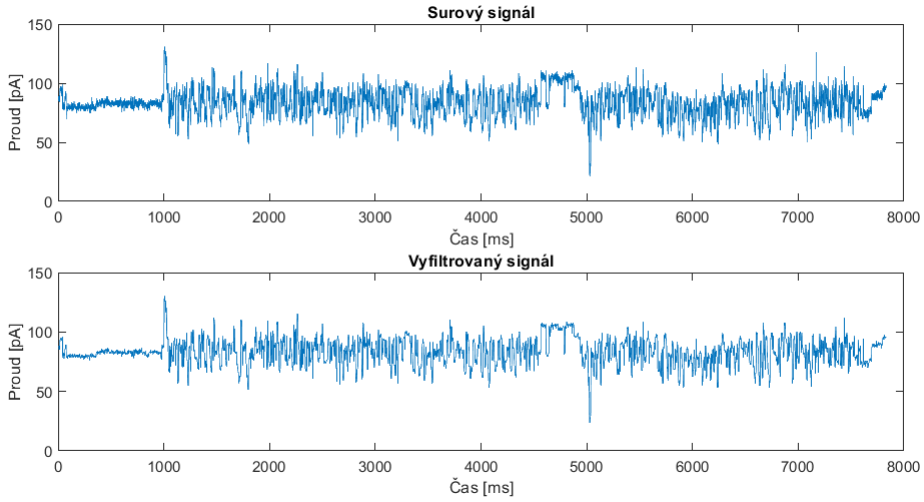
Protože se jedná o data, nasnímána stejným měřítkem, není třeba je normalizovat, např. z-scorem. Nutným krokem je však restaurace signálu. Signály jsou ve FAST5 souborech uloženy jako hladiny signálu. Signály jsou původně měřeny jako hodnoty pA (pikoampéry) a dále převedeny a uloženy jako 16bitové celočíselné hodnoty. Důvodem této transformace je komprese dat. Transformace zpět na původní hodnoty pA vyžaduje informaci o offsetu a škálování, které jsou uloženy ve FAST5 souborech v poli atributy. To je provedeno pomocí vztahu:

$$pA\_val = scale * (raw + offset) \tag{2.13}$$

Dalším krokem předzpracování signálů byla mediánová filtrace, a to i přes to, že společnost ONT na svých stránkách uvádí, že signály již není třeba dále filtrovat[34]. Signály se ukázaly být zarušené impulsním šumem, tudíž bylo potřeba alespoň se pokusit impulsní rušení vyfiltrovat. Pokud by data nebyla filtrována, mohla by být ovlivněna následná analýza odlehlými hodnotami a výpočty by mohly udávat chybné výsledky, to by mohlo vést ke zkreslení shlukové analýzy. K filtraci signálů byl použit mediánový filtr, délka okna byla nastavena na 5, kdy vzorky byly v okně setřízeny podle velikosti a na výstup byl poslán medián. Tato velikost okna byla vybrána především z důvodu, že větší okno by mohlo vyfiltrovat i jednotlivé k-mery. Výhodou mediánových filtrů mimo jiné je i to, že zachovávají hrany. Extrakce dat byla provedena v programovém prostředí MATLAB [32] a data byla uložena ve formátu .mat. Tento typ datového formátu byl zvolen především z důvodu, že následná analýza probíhá taktéž v programovém prostředí MATLAB a je tedy vhodnější.

Všechny surové signály byly vzorkovány vzorkovací frekvencí 4kHz. Důvodem takto vysoké vzorkovací frekvence je zachycení všech k-mer.

Níže na obrázku je příklad surového signálu a filtrovaného signálu. 2.1.



Obr. 2.1: Surový a filtrovaný signál

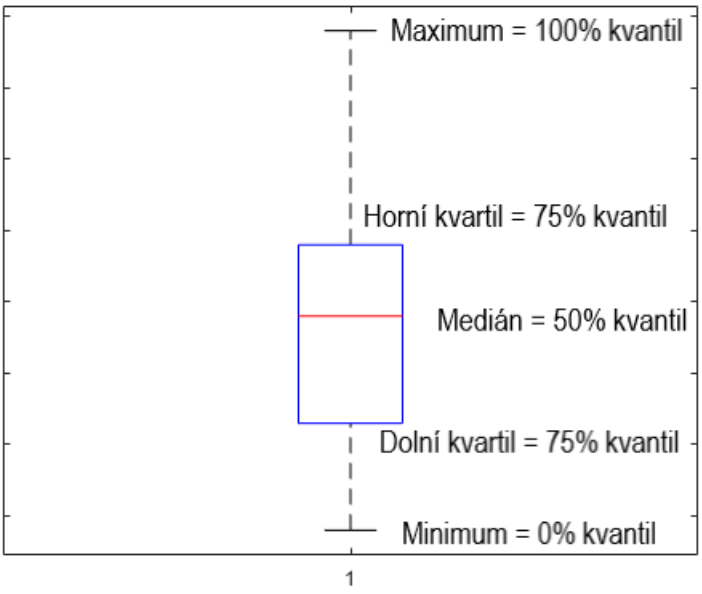
.

Krabicové grafy (boxploty)

K vykreslení a analýze dat byly použity krabicové grafy. Jedná se o standardizované způsoby zobrazení kvantitativních dat. Obsahuje informace o rozptylu dat, jejich symetričnosti, seskupení a zkreslení. Skládá se z krabice a "fousků". Čára uprostřed obdélníkového útvaru značí medián, to je 50 % kvartilu. Horní a dolní kraje obdélníku vyjadřují polohu horního a dolního kvartilu. Kvartil je hodnota, která z hodnot odděluje nejnižší čtvrtinu hodnot. Fousky označují variabilitu dat pod prvním a nad třetím kvartilem. Odlehlé hodnoty jsou vyznačovány jako červené křížky [31].

Shapiro-Wilkův test

Jedná se o test normality, který testuje nulovou hypotézu kdy vzorek pochází z normálně distribuované populace. Pokud je tedy hodnota p menší, než zvolená hladina alfa, pak je nulová hypotéza zamítnuta a existuje důkaz, že testovaná data nejsou normálně distribuována. [33] [34]



Obr. 2.2: Krabicový graf s vyznačením významných hodnot.

.

Kruskal-Wallisův test

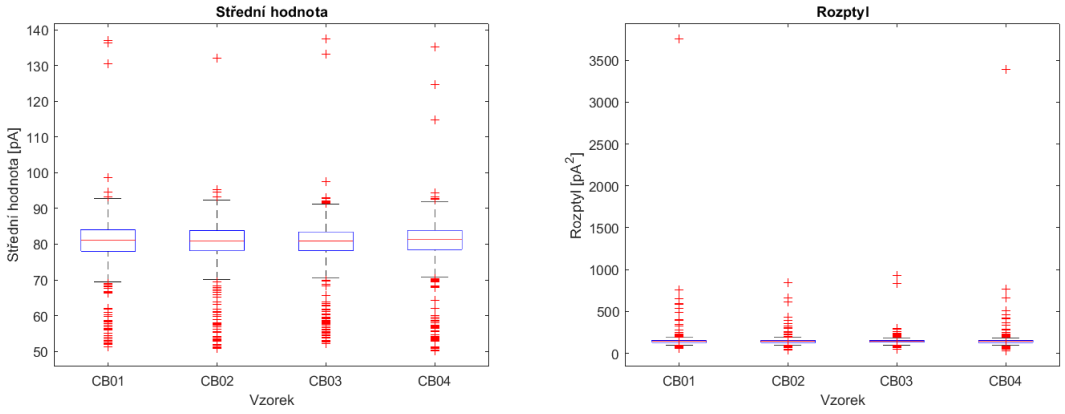
Kruskal-Waliss test je neparametrický statistický test používaný k posouzení existence statisticky významných rozdílů mezi dvěma a více skupinami nezávislé proměnné. Tento test nepředpokládá normální rozdělení dat, jeho nevýhodou je však nižší citlivost na odlehlé hodnoty. Předpokládá zisk dat náhodným výběrem. Výsledek testu je závislý na p-hodnotě, jejíž hodnota je ve většině případů nastavena na 0,05. Pokud je p-hodnota větší než 0,05, nulová hypotéza H_o předpokládá, že mediány souborů dat jsou stejné. Naproti tomu alternativní hypotéza H_1 předpokládá, že mediány souboru dat nejsou stejné.

2.3 Dataset z organismu *Clostridium beijerinckii*

Nejdříve byl vytvořen dataset obsahující data z bakterie *Clostridium beijerinckii*. Účelem tohoto experimentu je zjistit, zda se mění parametry pro různá čtení ze stejného organismu a slouží pro ověření, zda použité parametry rozlišují signály z jednoho organismu při použití jednoho druhu sekvenační chemie a průtokové komůrky. Jedná se o čtyři vzorky označeny CB01-CB04. Každý z těchto vzorků obsahuje 1000 náhodně vybraných signálů a informaci o čtení příslušející danému signálu. Tento organismus byl nasekvenován zařízením MinION a byl použit Ligation sequencing kit 009.

Každý signál byl popsán statistickými parametry uvedenými v předchozí kapitole. Jedná se o střední hodnotu, rozptyl, směrodatnou odchylku, variační koeficient, koeficient šikmosti, koeficient špičatosti a Hjorthovy deskriptory, mezi něž patří aktivita, komplexita a mobilita. Data byla vykreslena pomocí krabicových grafů za použití funkce boxplot.

Výsledky statistické analýzy

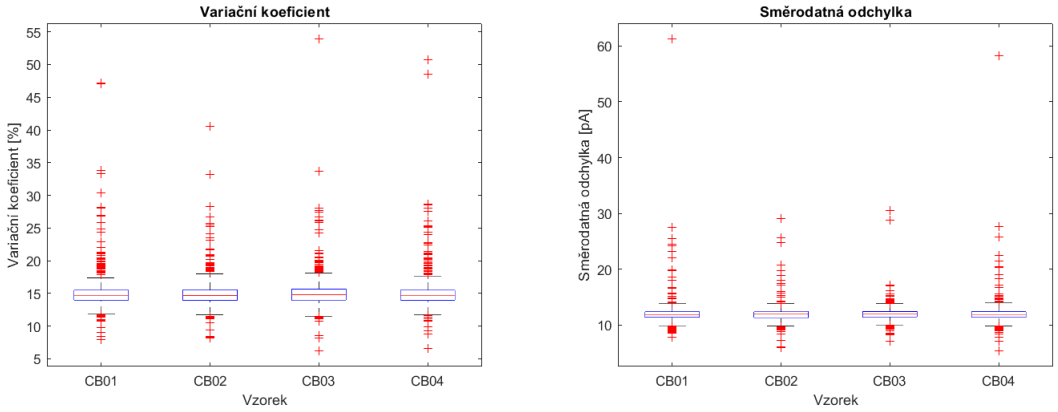


Obr. 2.3: Střední hodnota a rozptyl signálů prvního datasetu

.

Jak můžete vidět na obrázku 2.3, tak parametry střední hodnoty se pro jednotlivé vzorky nijak významně neliší. Obsahuje velké množství odlehlých hodnot v rozmezí 250-480. U hodnot rozptylu můžeme pozorovat že vzorky CB01, CB02 a CB04 se taktéž výrazně neliší, jedinou odlišnost můžeme pozorovat u datového souboru CB03. Odlišnost však není natolik výrazná, aby data od sebe dokázala spolehlivě odlišit.

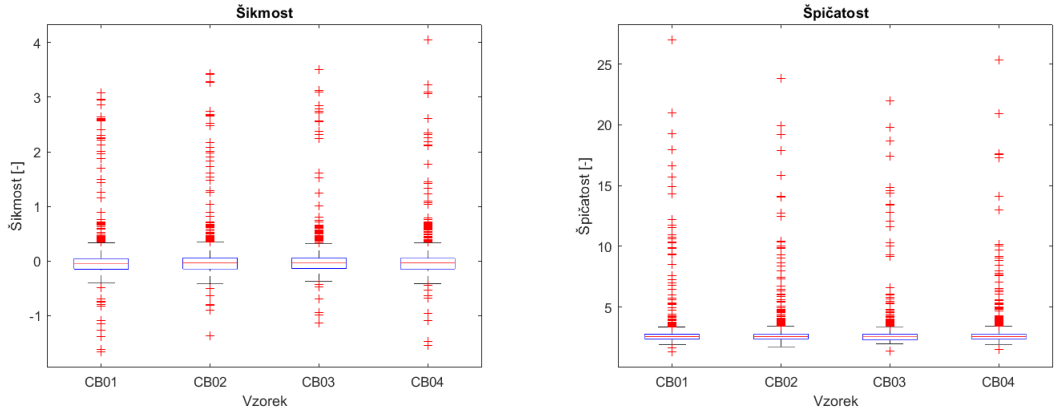
Obrázek 2.4 znázorňuje statistické parametry, jako je variační koeficient a směrodatná odchylka. Na obou obrázcích je patrné velké množství odlehlých hodnot a také to, že se mediány a kvartily u všech dat překrývají. Jedinou odlišnost můžeme



Obr. 2.4: Variační koeficient a směrodatná odchylka signálů prvního datasetu

.

pozorovat u vzorků CB02 a CB03 u směrodatné odchylky oproti zbylým dvěma datovým souborům. Liší se hodnota mediánů. Jako i u předchozí analýzy je patrné, že vzorky od sebe nelze spolehlivě odlišit.



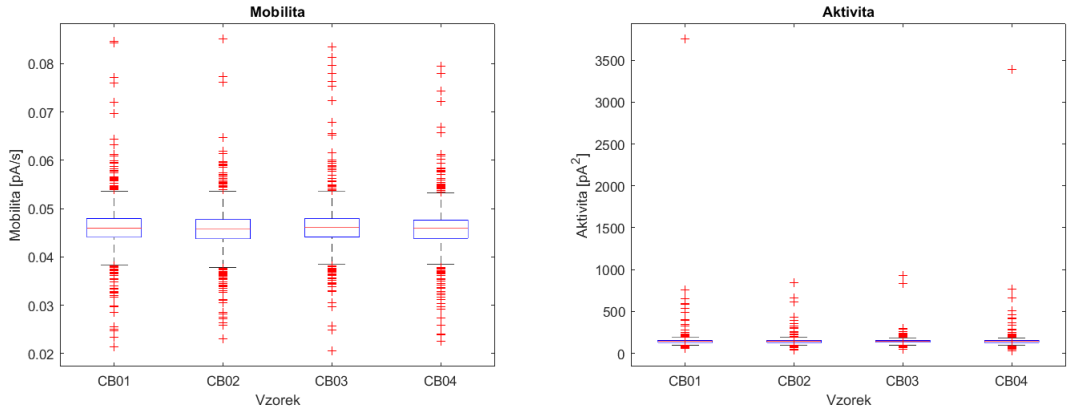
Obr. 2.5: Šikmost a špičatost signálů prvního datasetu

.

Parametr šikmosti nevykazuje přílišné rozdílnosti, překrývají se jak 'fousky', kvartily i mediány. Špičatost vzorků CB01, CB02 a CB04 si je velmi podobná, jediný menší rozdíl můžeme pozorovat u CB03, kde je kvartil nepatrně širší než zbylá data.

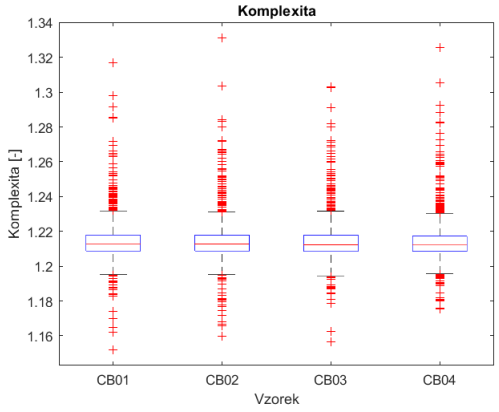
U parametru mobility vypadají výsledky na první pohled téměř identicky. U parametru aktivity si můžeme všimnout malého rozdílu u vzorku CB03.

Jako všechny předchozí parametry, tak ani komplexita nevykazuje významné rozdíly mezi daty.



Obr. 2.6: Mobilita a aktivita signálů prvního datasetu

.



Obr. 2.7: Komplexita signálů prvního datasetu

.

Abychom se však ujistili, že náš závěr je objektivní, byly použity statistické testy. Nejdříve bylo potřeba zjistit, jestli rozložení dat je normální či nikoli. To bylo vyhodnoceno pomocí Shapiro-Wilkova testu. Výsledné hodnoty jsou pro všechny parametry 0, což znamená, že data nejsou normálně distribuována.

Protože data nemají normální rozložení, byl použit neparametrický Kruskal-Wallisův test. Výsledky jsou uvedeny v tabulce 2.1.

Tab. 2.1: p-hodnoty

Statistický parametr	p-hodnota
Střední hodnota	0.279992014603674
Směrodatná odchylka	0.268752382251644
Variační koeficient	0.234913302392045
Koeficient šikmosti	0.382306633677903
Koeficient špičatosti	0.028315205283533
Rozptyl	0.268752382251644
Mobilita	0.207132652350674
Aktivita	0.268752382251644
Komplexita	0.732510483764367

Následně byl proveden Tukeyho test, který je schopen ukázat, které jednotlivé dvojice se od sebe liší a které nikoli. Většina parametrů není vhodná pro rozlišení žádné z dvojic, výjimkou je pouze parametr špičatosti, který je schopen jednu dvojici rozlišit.

Pokud tedy tuto analýzu stručně shrneme, výsledky dosahují předpokladů, že se hodnoty pro jeden organismus nebudou výrazně lišit. Můžeme si všimnout že kvartily a mediány se u všech vypočítaných statistických parametrů překrývají, tudíž není možné skupiny dat od sebe odlišit. Dále si můžeme všimnout i velkého množství odlehlých hodnot, které jsou však pro data CB01-CB04 téměř totožné. To je významné zjištění z hlediska možné následné klasifikace různých organismů a její shlukové analýzy. Dále z toho vyplývá, že můžou být vybrány náhodné signály z různých čtení pro jejich klasifikaci, a že data nejsou vázána na jednu konkrétní část genomu. To může mít význam i pro klasifikaci organismů přímo ze surového signálu bez nutnosti basecallingu.

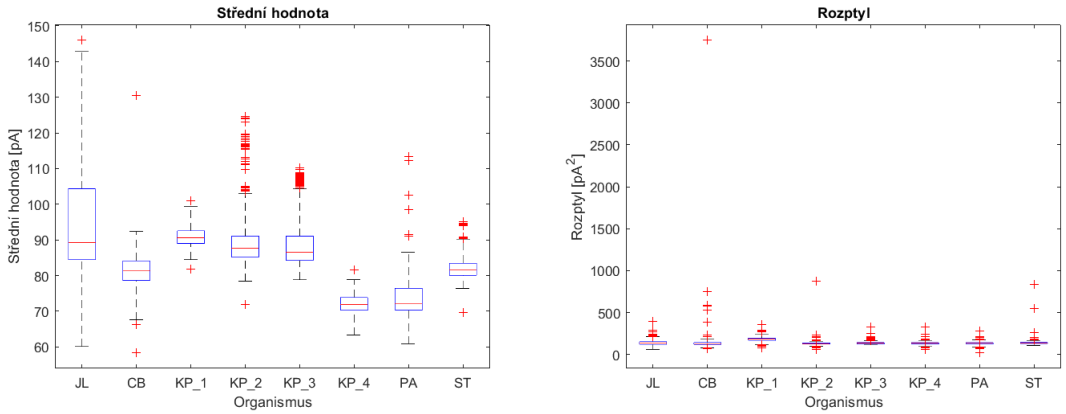
2.4 Dataset různých organismů sekvenovaných různými přístroji ONT

Tento dataset obsahuje 8 souborů formátu .mat jejichž přehled je uveden v tabulce 2.2. Obsahuje celkem pět organismů získaných ze dvou druhů sekvenačních kitů a dvou typů průtokových komůrek, každý o velikosti 800 signálů. Z databáze různých sekvenací na UBMI byly vybrány a následně vyextrahovány ze souboru FAST5 surové signály. Poté bylo nezbytné vybrané FAST5 soubory nabasecallovat a poté data sekvenované pomocí Rapid Barcoding kitu demultiplexovat. Demultiplexace je proces odstranění identifikátoru, což je v podstatě umělá sekvence přidána k DNA. Basecalling a demultiplexace byla provedena pomocí gridového počítání na výpočetních strojích Metacentra v programu Guppy (verze 5.0.15-gpu), provozovaného přímo společností ONT. Hlavním důvodem, proč bylo potřeba provést tento krok je, aby v následné analýze byla porovnávána stejná čtení a proto, že basecalling je úspěšný přibližně v 50% případů a některá čtení se nepodaří nabasecallovat v dostatečné kvalitě. Pokud by data nebyla nabasecallovaná a dataset by byl vytvořen bez basecallingu, mohlo by dojít ke ztrátě některých čtení. Bylo potřeba zadat jaký přístroj a sekvenační kit byl při sekvenaci použit a poté byl výpočet spuštěn. Výsledek basecallingu byl rozdělen do dvou složek, fail a pass. Byla tedy vybrána data ze souboru pass, který obsahuje čtení, ke kterým se podařilo úspěšně přiřadit znakovou sekvenci. Poté byla ze souborů FAST5 a FASTQ vyextrahována data, která měla stejné ID čtení, a tím bylo zajištěno, že vytvořený dataset bude obsahovat stejná data jako basecallované soubory FASTQ se kterými se v této bakalářské práci bude pracovat v následujících kapitolách. Dalším nutným krokem bylo, jako u minulého datasetu, předzpracování dat. Signály tedy byly převedeny na pA hodnoty a filtrovány mediánovou filtrací.

Tab. 2.2: Přehled vytvořeného datasetu

Název organismu	Název datasetu	Sekvenační zařízení	Kit
Clostridium beijerinckii	CB	MinION 006	LSK 009
Klebseilla pneumoniae	KP_1	MinION 006	RBK 004
Klebseilla pneumoniae	KP_2	Flonge 001	RBK 004
Klebseilla pneumoniae	KP_3	Flonge 001	LSK 009
Klebseilla pneumoniae	KP_4	MinION 006	LSK 009
Pantoea agglomerans	PA	MinION 006	LSK 009
Schlegelella thermodepolymerans	ST	MinION 006	LSK 009
Janthinobacterium lividum	JL	MinION 006	RBK 004

Výsledky statistické analýzy



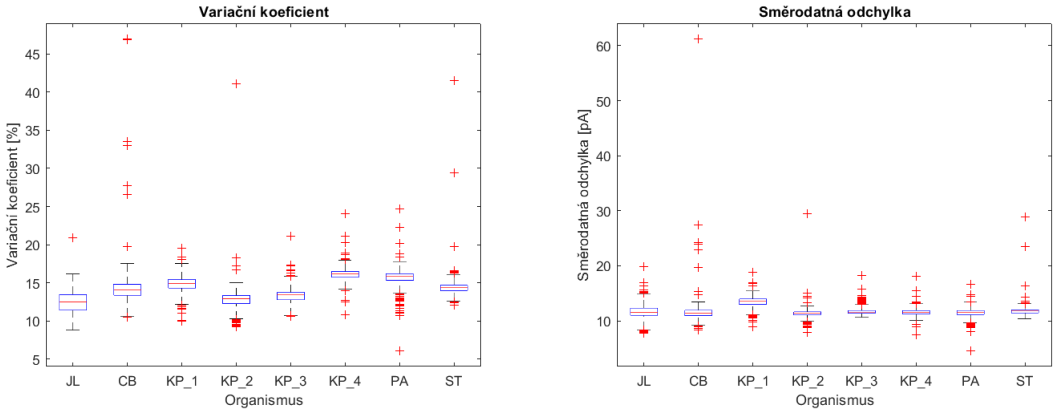
Obr. 2.8: Střední hodnota a rozptyl signálů druhého datasetu

.

Na obrázku 2.8 jsou vyobrazeny vypočtené statistické parametry střední hodnoty a rozptylu. V případě střední hodnoty je patrné i z obrázku, kdy se kvartily ani mediány výrazně nepřekrývají, že organismy půjde spolehlivě odlišit. Pouze u organismu KP_2 a KP_3 je možné si všimnout jisté podobnosti. Jednak mají oba organismy velké množství odlehlých hodnot, ale i překrývající se kvartily. To může být způsobeno především tím, že se jedná o stejný organismus sekvenovaný pomocí průtokové komůrky Flonge, která je obecně považována za levnější a méně přesnou oproti MinION. Zajímavé je, že ostatní parametry stejného organismu KP_1 a KP_4 se od KP_2 a KP_3 liší, i když se jedná o stejný organismus. Jediný rozdíl je, jak již bylo zmíněno, že k sekvenaci byly použity různé kity a průtokové komůrky.

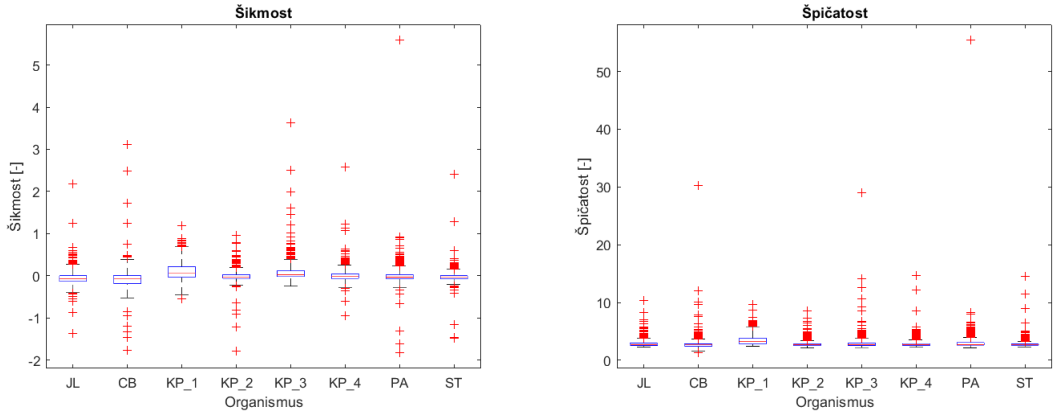
Parametr rozptylu se od prvního pohledu k rozlišení organismů od sebe úplně nehodí. Organismy se JL, CB, KP_1 a KP_2 se sice nepřekrývají, zbylé čtyři však ano.

U variačního koeficientu se nepřekrývají ani kvartily, ani mediány, proto se tedy variační koeficient jeví jako vhodný parametr k rozlišení organismů. Směrodatná odchylka se také jeví jako vhodný parametr k rozlišení. Nutno podotknout, že kvartily KP_2 a KP_4 se překrývají, což je ve své podstatě správný výsledek.



Obr. 2.9: Variační koeficient a směrodatná odchylka signálů druhého datasetu

.

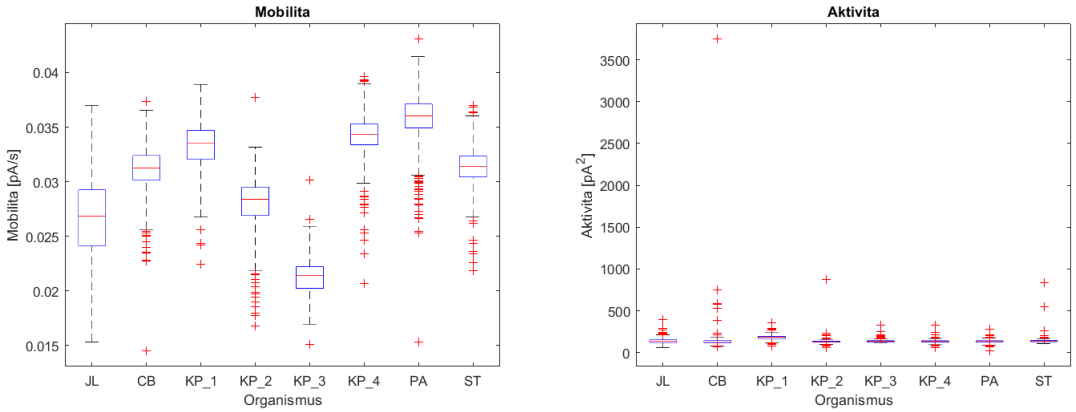


Obr. 2.10: Šikmost a špičatost signálů druhého datasetu

.

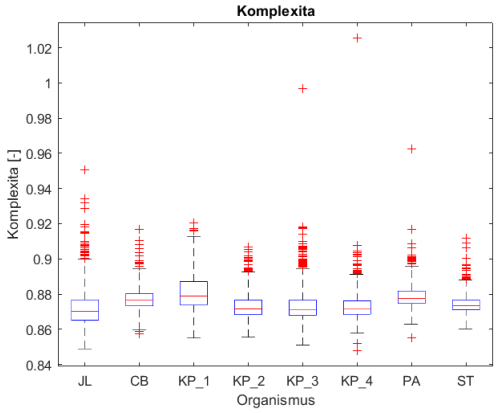
Z obrázku 2.10 je patrné, že oba statistické parametry nejsou úplně vhodné pro klasifikaci organismů. Kvartily a mediány se v obou případech značně překrývají. Hodnoty šikmosti oscilují kolem nuly a obsahují velké množství odlehlých hodnot. U parametru špičatosti se výrazně liší pouze KP_1.

Vizuálně je patrné, že parametr mobility je pro klasifikaci organismů o něco vhodnější než parametr aktivity. Kvartily ani mediány se nepřekrývají ani u jednoho z parametrů. U mobility má organismus JL výrazně větší fousky, než ostatní organismy.



Obr. 2.11: Mobilita a aktivita signálů druhého datasetu

.

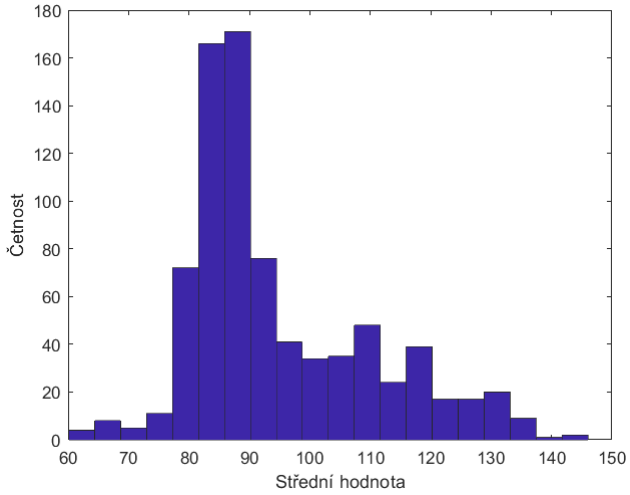


Obr. 2.12: Komplexita signálů druhého datasetu

.

U parametru komplexity se hodnoty sice překrývají, ale vzhledem k tomu že se většinou překrývají data z organismu KP_2, KP_3 a KP_4, tak by tento parametr mohl být pro klasifikaci celkem vhodný. Vyskytuje se zde velké množství odlehlých hodnot v rozmezí 0,9 až 0,93.

Pokud to tedy shrneme, tak dojdeme k závěru, že nejvhodnější parametr pro rozlišení se jeví variační koeficient a mobilita. Dalším parametrem vhodným k rozlišení organismů by mohla být komplexita, a to především z důvodu, že správně rozlišila organismy KP_2, KP_3 a KP_4 od ostatních. Naopak jako ne úplně vhodné se zdá být použití parametrů šikmosti a špičatosti. Vzhledem k tomu, že vizuální hodnocení může být velmi neobjektivní, bylo potřeba si tyto závěry potvrdit statistickými testy.



Obr. 2.13: Histogram střední hodnoty signálu

.

Abychom zvolili správný test, je třeba zjistit rozložení dat. K tomu mohou sloužit histogramy. Na obrázku 2.13 je jasně vidět, že data nevykazují normální rozložení. Pouze vizuální hodnocení nemusí být vždy úplně objektivní, byly tedy vypočítány testy normality pomocí funkce normalitytest.m[34]. Konkrétně byl použit Shapiro-Wilkův test, výsledky tohoto testu jsou uvedeny v tabulce2.3. Bylo zjištěno, že většina statistických parametrů nevykazuje normální rozložení. Jedinou výjímkou je organismus KP_4 u parametru střední hodnoty. Protože je v datech velké množství odlehlých hodnot a také proto, že data převážně nevykazují normální rozložení, byl ke statistické analýze použit neparametrický Kruskal-Wallisův test. V tabulce 2.4 jsou uvedeny výsledky testu pro jednotlivé statistické parametry. Na základě výsledků Kruskall-Wallisova testu se zdají nejvhodnějšími parametry pro rozlišení střední hodnota signálu, směrodatná odchylka, variační koeficient, rozptyl, mobilita a aktivita jejichž hodnoty jsou 0. Naopak nejméně vhodný se jeví parametr šikmosti

Tab. 2.3: Výsledek Shapiro-Wilkova testu

	JL	CB	KP_1	KP_2	KP_3	KP_4	PA	ST
Střední hodnota	0	0	0.0009	0	0	0.3190	0	0
Rozptyl	0	0	0	0	0	0	0	0
Variační koeficient	0	0	0	0	0	0	0	0
Směrodatná odchylka	0	0	0	0	0	0	0	0
Šikmost	0	0	0	0	0	0	0	0
Špičatost	0	0	0	0	0	0	0	0
Mobilita	0.0004	0	0	0	0	0	0	0
Aktivita	0	0	0	0	0	0	0	0
Komlexita	0	0	0	0	0	0	0	0

Tab. 2.4: Výsledek Kruskal-Wallisova testu pro druhý dataset

Statistický parametr	p-hodnota
`Střední hodnota`	0
`Směrodatná odchylka`	0
`Variační koeficient`	0
`Koeficient šikmosti`	2.4889e-198
`Koeficient špičatosti`	4.1757e-168
`Rozptyl`	0
`Mobilita`	0
`Aktivita`	0
`Komplexita`	0

Tab. 2.5: Výsledek post-hoc testu střední hodnoty, rozptylu a variačního koeficientu

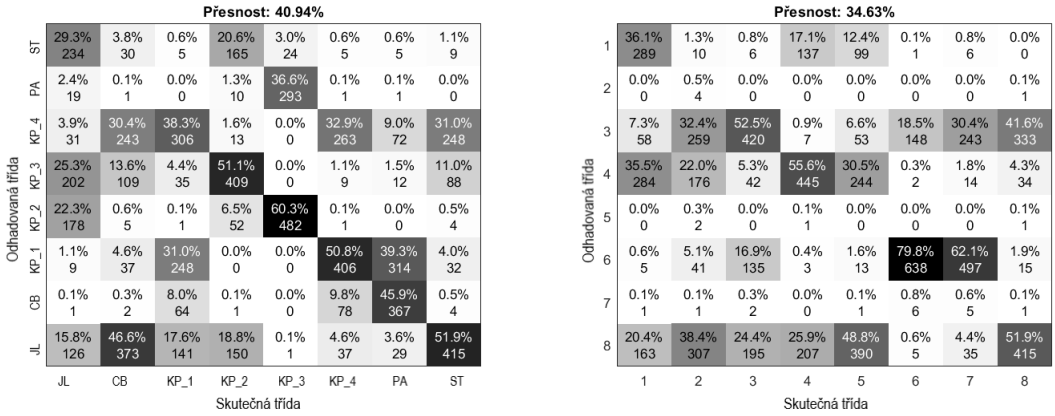
Parametr	Odlišitelné organismy	Neodlišitelné organismy
Střední hodnota	JL vs. CB/KP_1/KP_3/KP_4/PA/ST	JL vs. KP_2
	CB vs. JL/KP_1/KP_2/KP_3/KP_4/PA	CB vs. ST
	KP_1 vs. JL/CB/KP_2/KP_3/KP_4/PA/ST	-
	KP_2 vs. CB/KP_1/KP_4/PA/ST	KP_2 vs. JL/KP_3
	KP_3 vs. JL/CB/KP_1/KP_4/PA/ST	KP_3 vs. KP_2
	KP_4 vs. JL/CB/KP_1/KP_2/KP_3/ST	KP_4 vs. PA
	PA vs. JL/CB/KP_1/KP_2/KP_3/KP_4/ST	PA vs. KP_4
	ST vs. JL/KP_1/KP_2/KP_3/KP_4/PA	ST vs. CB
Rozptyl	JL vs. CB/KP_1/KP_3/KP_2/ST	JL vs. KP_3/KP_4/PA
	CB vs. JL/KP_1/KP_2/KP_3/KP_4/ST	CB vs. PA
	KP_1 vs. JL/CB/KP_2/KP_3/KP_4/PA/ST	-
	KP_2 vs. JL/CB/KP_1/KP_3/KP_4/PA/ST	-
	KP_3 vs. CB/KP_1/KP_2/PA/ST	KP_3 vs. JL/KP_4
	KP_4 vs. CB/KP_1/KP_2/ST	KP_4 vs. JL/KP_3/PA
	PA vs. KP_1/KP_2/KP_3/ST	PA vs. JL/CB/KP_4
	ST vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA	-
Variační koeficient	JL vs. CB/KP_1/KP_3/KP_4/PA/ST	JL vs. KP_2
	CB vs. JL/KP_1/KP_2/KP_3/KP_4/PA/ST	-
	KP_1 vs. JL/CB/KP_2/KP_3/KP_4/PA/ST	-
	KP_2 vs. CB/KP_1/KP_3/KP_4/PA/ST	KP_2 vs. JL
	KP_3 vs. JL/CB/KP_1/KP_2/KP_3KP_4/PA/ST	-
	KP_4 vs. JL/CB/KP_1/KP_2/KP_3/PA/ST	-
	PA vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA/ST	-
	ST vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA	-

Tab. 2.6: Výsledek post-hoc testu směrodatné odchylky, šikmosti a špičatosti

Parametr	Odlišitelné organismy	Neodlišitelné organismy
Směrodatná odchylka	JL vs. CB/KP_1/KP_3/KP_2/ST	JL vs. KP_3/KP_4/PA
	CB vs. JL/KP_1/KP_2/KP_3/KP_4/ST	CB vs. PA
	KP_1 vs. JL/CB/KP_2/KP_3/KP_4/PA/ST	-
	KP_2 vs. JL/CB/KP_1/KP_3/KP_4/PA/ST	-
	KP_3 vs. CB/KP_1/KP_2/PA/ST	KP_3 vs. JL/KP_4
	KP_4 vs. CB/KP_1/KP_2/ST	KP_4 vs. JL/KP_3/PA
	PA vs. KP_1/KP_2/KP_3/ST	PA vs. JL/CB/KP_4
	ST vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA	-
Koeficient šikmosti	JL vs. KP_1/KP_2/KP_3/KP_4/PA/ST	JL vs. CB
	CB vs. KP_1/KP_2/KP_3/KP_4/PA/ST	CB vs. JL
	KP_1 vs. JL/CB/KP_2/KP_4/PA/ST	KP_1 vs. KP_3
	KP_2 vs. JL/CB/KP_1/KP_3/ST	KP_2 vs. KP_4/PA
	KP_3 vs. JL/CB/KP_2/KP_3/KP_4/PA/ST	KP_3 vs. KP_1
	KP_4 vs. JL/CB/KP_1/KP_3/ST	KP_4 vs. KP_2/PA
	PA vs. JL/CB/KP_1/KP_3/PA	PA vs. KP_2/KP_4/ST
	ST vs. JL/CB/KP_1/KP_2/KP_3/KP_4	-
Koeficient špičatosti	JL vs. CB/KP_1/PA	JL vs. KP_2/KP_3/KP_4/ST
	CB vs. JL/KP_1/KP_2/KP_3/KP_4/PA/ST	-
	KP_1 vs. JL/CB/KP_2/KP_3/KP_4/PA/ST	-
	KP_2 vs. CB/KP_1/PA	KP_2 vs. JL/KP_3/KP_4/ST
	KP_3 vs. CB/KP_1/PA	KP_3 vs. JL/KP_2/KP_4/ST
	KP_4 vs. CB/KP_1/PA	KP_4 vs. JL/KP_2/KP_3/ST
	PA vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA/ST	-
	ST vs. CB/KP_1/PA	ST vs. JL/KP_2/KP_3/KP_4

Tab. 2.7: Výsledek post-hoc testu vybraných parametrů: aktivita a komplexita

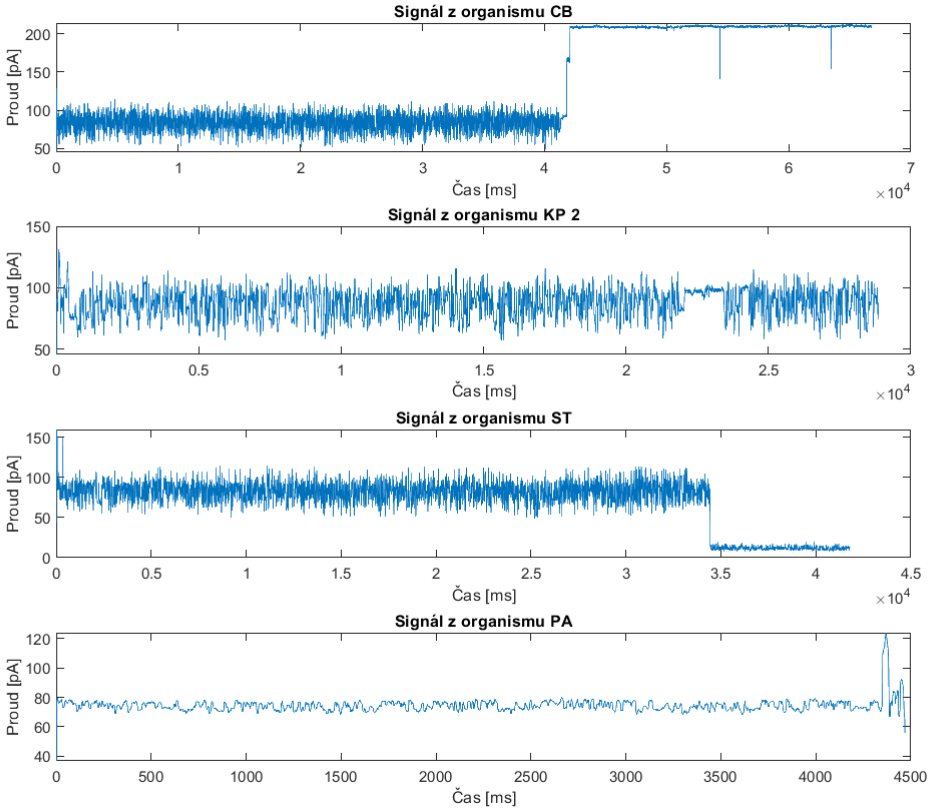
Parametr	Odlišitelné organismy	Neodlišitelné organismy
Mobilita	JL vs. CB/KP_1/KP_3/KP_4/PA/ST	JL vs. KP_2
	CB vs. JL/KP_1/KP_2/KP_3/KP_4/PA	CB vs. ST
	KP_1 vs. JL/CB/KP_2/KP_3/KP_4/PA/ST	-
	KP_2 vs. CB/KP_1/KP_3/KP_4/PA/ST	KP_2 vs. JL
	KP_3 vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA/ST	-
	KP_4 vs. JL/CB/KP_1/KP_2/KP_3/PA/ST	-
	PA vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA/ST	-
	ST vs. JL/KP_1/KP_2/KP_3/KP_4/PA	ST vs. CB
Aktivita	JL vs. CB/KP_1/KP_3/KP_2/ST	JL vs. KP_3/KP_4/PA
	CB vs. JL/KP_1/KP_2/KP_3/KP_4/ST	CB vs. PA
	KP_1 vs. JL/CB/KP_2/KP_3/KP_4/PA/ST	-
	KP_2 vs. JL/CB/KP_1/KP_3/KP_4/PA/ST	-
	KP_3 vs. CB/KP_1/KP_2/PA/ST	KP_3 vs. JL/KP_4
	KP_4 vs. CB/KP_1/KP_2/ST	KP_4 vs. JL/KP_3/PA
	PA vs. KP_1/KP_2/KP_3/ST	PA vs. JL/CB/KP_4
	ST vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA	-
Komplexita	JL vs. CB/KP_1/PA/ST	JL vs. KP_2/KP_3/KP_4
	CB vs. JL/KP_1/KP_2/KP_3/KP_4/PA/ST	-
	KP_1 vs. JL/CB/KP_2/KP_3/KP_4/ST	KP_1 vs. PA
	KP_2 vs. CB/KP_1/PA/ST	KP_2 vs. JL/KP_3/KP_4
	KP_3 vs. CB/KP_1/PA/ST	KP_3 vs. JL/KP_2/KP_4
	KP_4 vs. CB/KP_1/PA/ST	KP_4 vs. JL/KP_2/KP_3
	PA vs. JL/CB/KP_2/KP_3/KP_4/ST	PA vs. KP_1
	ST vs. JL/CB/KP_1/KP_2/KP_3/KP_4/PA	-



Obr. 2.14: Konfuzní matice. Vlevo pro mobilitu a vpravo pro variační koeficient.

Abychom zjistili, které konkrétní dvojice organismů lze od sebe odlišit, byla provedena ještě post-hoc analýza. Výpočty byly provedeny pomocí Tukeyho testu a výsledky zapsány do tabulek. Pomocí těchto výpočtů byly vybrány parametry, které jsou nejvíce vhodné pro shlukování. Pro vybrané parametry, které měly největší úspěšnost, byla vypočítána matice záměn.

Z výsledků post-hoc analýzy je patrné, že signály mezi sebou dokážou nejlépe rozlišit parametry variační koeficient a mobilita. Je však také potřeba vzít v úvahu, že KP_1 až KP_4 jsou jeden organismus, tudíž by měly od sebe být těžko rozlišitelné. To do jisté míry nejlépe vyhodnotil parametr komplexity a koeficient šikmosti. Nabízí se však otázka, jak moc velký vliv má to, že organismus KP_1 až KP_4 byly sekvenovány každý jiným sekvenačním kitem a jiným sekvenačním zařízením, a jestli každá sekvenační chemie a zařízení neprodukují specifické signály. To by znamenalo, že při dekódování signálu na DNA sekvence je potřeba brát v úvahu, s jakým zařízením a kitem se pracovalo.



Obr. 2.15: Signály - odlehlé hodnoty

.

Signály, u kterých vyšly vypočtené hodnoty výrazně odlišnějších od ostatních, byly dohledány a vybrané z nich i zobrazeny. Jako příklad byly vybrány signály viz obrázek 2.15, protože se jako odlehlé hodnoty vyskytovali opakovaně u více vypočtených statistických parametrů. Můžeme si všimnout, že signály z organismu CB a ST mají společné to, že oba obsahují skok, který pravděpodobně způsobil, že parametry z něj vypočtené se lišily od většiny ostatních. Signál KP_2 obsahuje na svém začátku výraznější hodnoty pravděpodobně proto, že se musel ustálit. U signálu z organismu PA můžeme vidět fragment na konci, který pravděpodobně ovlivnil výpočty. Odlehlé hodnoty tedy většinou způsobují signály buď s nějakými skoky, nebo fragmenty a rušením. Pokud se, ale pozorně na signály podíváme, můžeme si povšimnout i další zajímavosti. Je možné už na první pohled vidět, že signály jsou druhově specifické.

3 Shluková analýza

Shluková analýza je statistická metoda strojového učení pro zpracování dat. Funguje na principu organizace položek do skupin, nebo shluků na základě toho, jak úzce jsou propojeny. Pokud jsou si velmi podobné, tak jsou zařazeny do stejného shluku a naopak. Shluková analýza může být hierarchická, anebo nehierarchická. Hierarchické metody jsou pak rozděleny na divizní a aglomerativní. U aglomerativních metod je každý objekt zařazen do jednoho shluku, a poté na základě podobnosti jsou jednotlivé shluky spojeny. Tento proces se opakuje, dokud nejsou všechny body sloučeny v jeden hierarchicky konstruovaný shluk nazývaný dendrogram [35]. Nehierarchická shluková analýza rozdělí objekty do několika shluků stejného řádu. Příkladem je algoritmus k-means.[36] Protože hlavním cílem této práce je klasifikovat data ze sekvenátoru ONT, o kterých nevíme nějaké bližší informace krom toho z jakého organismu pochází, jsou metody učení bez učitele vhodnější. U učení bez učitele před spuštěním modelu nevíte, kolik shluků v datech existuje. Na rozdíl od mnoha jiných statistických metod se shluková analýza obvykle používá, když neexistuje žádný předpoklad o pravděpodobných vztazích v datech. Poskytuje informace o tom, kde v datech existují asociace a vzory, ale ne o tom, jaké mohou být, nebo co znamenají.

Prvními kroky je sběr dat a vhodná úprava dat, tyto postupy byly popsány v předchozích kapitolách této práce. Poté je potřeba vypočítat matici vzdáleností. I přesto, že existuje několik variant, jak vzdálenost měřit, nejběžnějším způsobem je euklidovská vzdálenost. Různé způsoby počítání matice vzdáleností jsou představeny v další podkapitole.

V této studii byly použity dvě výše zmíněné shlukovací techniky. Počet shluků byl nastavován na 8 a 2 a podle toho, o jakou analýzu se zrovna jednalo. Poté bylo provedeno shlukování. Závěrem byla přesnost metod vyhodnocena pomocí konfuzní matice.

3.1 Vzdálenostní metriky

Informaci o tom, jak si jsou podobny dva objekty, můžeme zjistit pomocí vzdálenostních metrik. Možností, jak ji měřit, existuje nepřeberné množství. Mezi doporučenou a zároveň i nejvíce používanou je považována euklidovská vzdálenost. Mezi další používané vzdálenosti patří Minkowského, Kosinova, nebo Manhattanská vzdálenost.

Euklidovská vzdálenost Má větší význam u příznaků s větším rozdílem mezi objekty [37].

$$d_{euc}(\bar{x}, \bar{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.1}$$

Manhattanská vzdálenost Manhattanskou vzdálenost odstraní větší význam příznaků s větším rozdílem mezi objekty o délce n. Vzdálenost mezi vektory x a y lze vyjádřit jako [37]:

$$d_{man}(\bar{x}, \bar{y}) = \sum_{i=1}^{n}|(x_i - y_i)| \tag{3.2}$$

Kosinový koeficient Podle kosinové vzdálenosti je vzdálenost dvou, vektorů délky n, úhel, který svírají. Je necitlivý na násobení vektoru příznaků konstantou. Lze vyjádřit takto [37]:

$$c_{xy} = \frac{\sum_{i=1}^{n}(x_i y_i)}{\sum_{i=1}^{n}(x_i^2)\sum_{i=1}^{n}(y_i^2)} \tag{3.3}$$

Pearsonův korelační koeficient Jeho výhodou může být necitlivost na posunutí vektoru příznaků o konstantu, nebo na jeho násobení konstantou[38].

$$c_{xy} = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i)}{\sqrt{\left[\sum_{i=1}^{n} x^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right]\left[\sum_{i=1}^{n} y^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2\right]}} \tag{3.4}$$

Spearmanův korelační koeficient Jedná se o neparametrický korelační koeficient, který je robustní vůči odlehlým hodnotám a obecně odchylkám od normality, neboť stejně jako řada dalších neparametrických metod pracuje pouze s pořadími px a py pozorovaných hodnot [39]

$$s_{xy} = 1 - \frac{6\sum_{i=1}^{n}(p_i^x - p_i^y)^2}{n(n^2-1)} \tag{3.5}$$

3.2 Hierarchiecké shlukování

Existuje několik metod pro měření vzdáleností mezi shluky i mezi dvojicí objektů. Jednotlivé shlukovací algoritmy se liší způsobem stanovení vzdáleností mezi nově vzniklým shlukem (dvojici sloučených objektů) a ostatními objekty a jejich výběr je rozhodující pro vzhled dendrogramu.

Metoda nejbližšího souseda (single linkage) – vzdálenost shluků je odvozena ze vzdálenosti dvou nejbližších objektů z různých shluků.

Metoda nejvzdálenějšího souseda (complete linkage) – vzdálenost shluků je odvozená z nevzdálenějších objektů.

Metoda průměrné vazby (UPGMA) – spojení dle průměrné vzdálenosti mezi objekty shluků.

Wardova metoda – založena na principu analýzy rozptylu, shlukuje objekty tak, aby byl součet druhých mocnin vzdáleností objektů od centroidů jejich shluků minimální (minimalizace rozptylu).

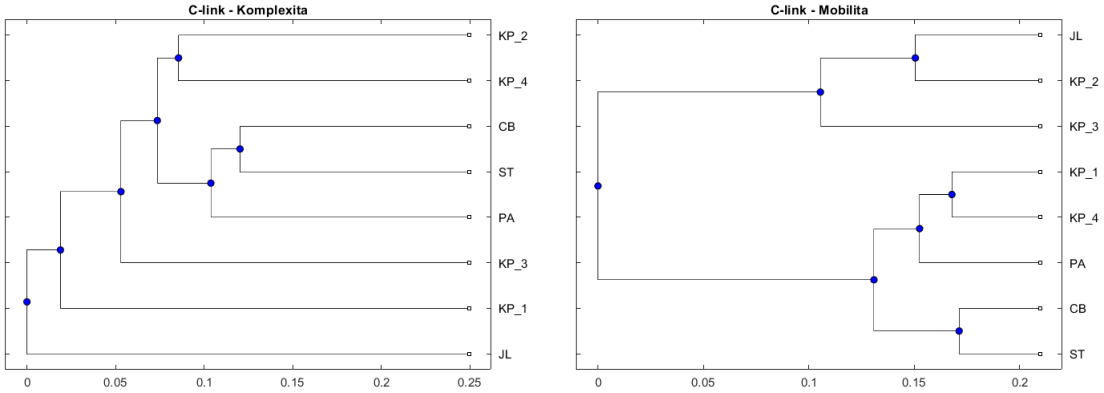
Hierarchické metody shlukování: Hlavní výhodou hierarchického shlukování je jeho jednoduchá interpretace výsledků dendrogramů, není nutné vědět počet shluků předem a jedná se o reprodukovatelnou metodu.

Mezi nevýhody patří to, že výsledek je silně závislý na zvolené metrice vzdálenosti mezi objekty a způsobu měření. Jedná se o výpočetně náročnou metodu nevhodnou pro velké soubory dat. Dále většinou není příliš efektivní, neumožňuje průběžnou změnu shluků a je citlivá na šum.

3.2.1 Výsledky hierarchického shlukování

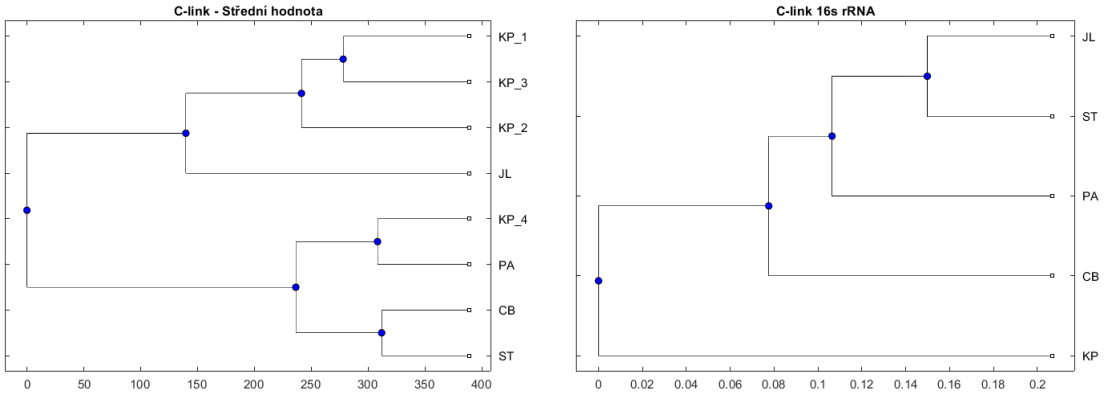
Bylo provedeno hierarchické shlukování vybraných parametrů. Tato metoda se však neprojevila jako úplně vhodná pro naši analýzu, protože jak bylo výše zmíněno, při větším množství dat je velmi nepřehledná, proto zde výsledky nejsou uvedeny. Jediné, co má smysl zde uvést, je fylogenetická analýza vybraných příznaků s basecallovanými a následně zarovnanými sekvencemi genu 16s rRNA, který díky tomu, že pomalu mutuje, je vhodný k fylogenetickým analýzám. Byla vypočtena p-distace a následně bylo všech 800 vzorků zařazeno do jednoho vektoru jako 800 příznaků daného organismu. Tím jsme získali porovnání, jak moc se liší výsledná analýza basecallovaných sekvencí s výsledky shlukování vybraných statistických parametrů. Další shlukování basecallovaných sekvencí v našem případě nemá smysl, protože sekvence by musely být zarovnány. V tomto konkrétním případě to není možné, protože čtení nebo konkrétně signály byly vybrány zcela náhodně. Obsahují tedy čtení i z nekódujících úseků genů, tudíž nejsou tzv. species specific a nelze je tak použít k fylogenetické analýze.

Další možností jak klasifikovat basecallované sekvence bez nutnosti zarovnání je pomocí výpočtu četností dinukleotidů [43]. Byly tedy vypočítány frekvence dinukleotidů pro každý organismus a následně byla vypočítána vzdálenost vektorů frekvencí pomocí euklidovské vzdálenosti. Euklidovská vzdálenost byla použita, protože se používá i u jiných analýz v této práci. Cílem bylo vytvořit fylogenetický strom z basecallovaných sekvencí surových signálů ONT. Pro tvorbu výsledného fylogenetického stromu byla použita metoda C-link.



Obr. 3.1: C-link komplexity a mobility

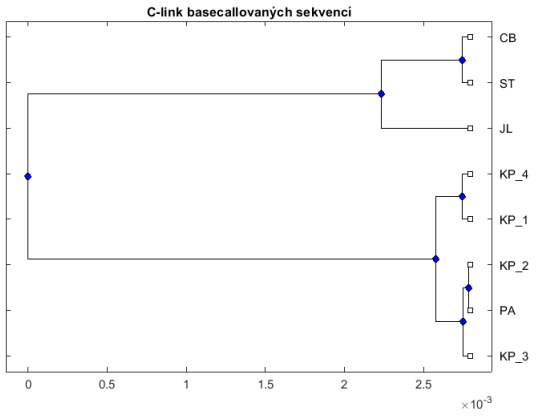
.



Obr. 3.2: C-link střední hodnoty a basecallovaných sekvencí

.

Můžeme vidět, že shlukování všech tří vybraných parametrů se od basecallovaných sekvencí z genu 16s rRNA značně liší. Tato analýza slouží k porovnání shlukování vybraných parametrů a basecallovaných sekvencí. Jak můžeme vidět různé přístupy vykazují zcela odlišné výsledky a každý parametr shlukuje vzorky jinak.



Obr. 3.3: C-link basecallovaných sekvencí - frekvence dinukleotidů

.

Pokud ale porovnáme výsledky s metodou využití četnosti dinukleotidů, můžeme pozorovat, že parametr mobility a parametr střední hodnoty správně vyhodnotili vzorky ST a CB. Jistá podobnost tedy mezi výsledky je, přesnost však není nijak velká a až na některé vzorky, se výsledky značně liší.

3.3 Nehierarchické shlukování

Algoritmus k-means (česky také nazýváno jako metoda k-průměrů) je jedna z nejintuitivnějších a nejběžnějších metod nehierarchického shlukování. Na začátek je potřeba nastavit počet shluků. Centroidy, kterými je každý shluk reprezentován svým geometrickým středem, jsou na začátku rozmístěny náhodně (můžou být nastaveny i ručně). Dalším postupem je posun centroidů na základě vypočtených vzdáleností, aby se zvýšila podobnost objektů v rámci jednotlivých shluků a zvýšila odlišnost objektů z různých shluků.

3.3.1 Výsledky nehierarchického shlukování

Tato studie se zabývá vlivem sekvenační chemie a sekvenátorů na výsledné shlukování organismů a také na klasifikaci organismů jako takovou.

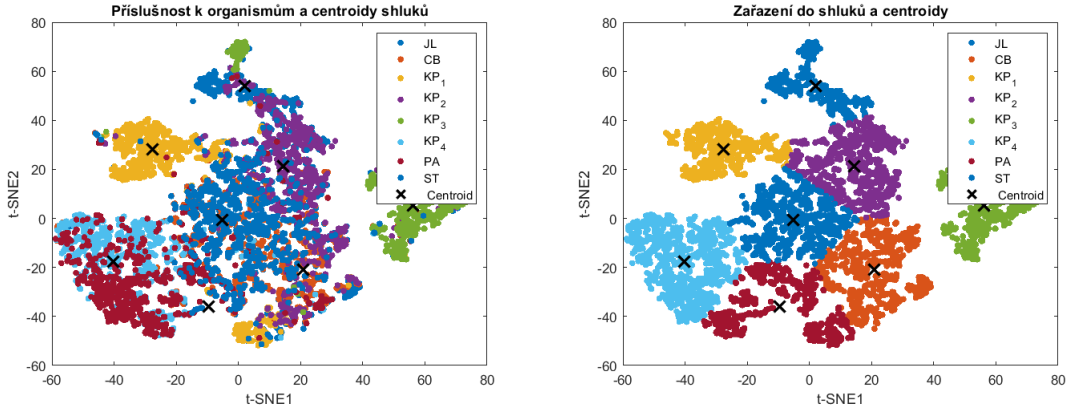
Výhodou tohoto přístupu je, že není vyžadována znalost kompletního genomu nebo genů. Velké množství krátkých čtení také není třeba zarovnávat, sestavovat do kontigů či porovnávat s referenčními databázemi, které často bývají nekompletní[40].

Použitou vzdálenostní metrikou je euklidovská vzdálenost jako v předchozích analýzách. Byly vybrány vhodné parametry schopné odlišit organismy na základě

předchozí statistické analýzy. Vybrané příznaky byly standardizovány pomocí z-score. Analýza samotná je provedena pomocí funkce k-means. Nejdříve bylo potřeba určit počet shluků, ty byly voleny podle toho, na základě čeho jsme chtěli data klasifikovat. Logické rozdělení je shlukování podle počtu vzorků, nehledě na použitou sekvenační chemii. Hodnota tedy byla nastavena na 8. Dalším postupem bylo shlukování na základě použité sekvenační chemie nebo sekvenačního přístroje. V tomto případě byla nastavena hodnota na 2. Následně jsou iterativně hledány centroidy shluků. Počáteční pozice centroidů jsou zvoleny náhodně. Následně byly jednotlivé vzorky zařazeny do shluků a přečísleny pomocí funkce renum_clust[42]. Pomocí konfuzní matice byla vypočtena výsledná přesnost metody.

Shlukování podle počtu vzorků

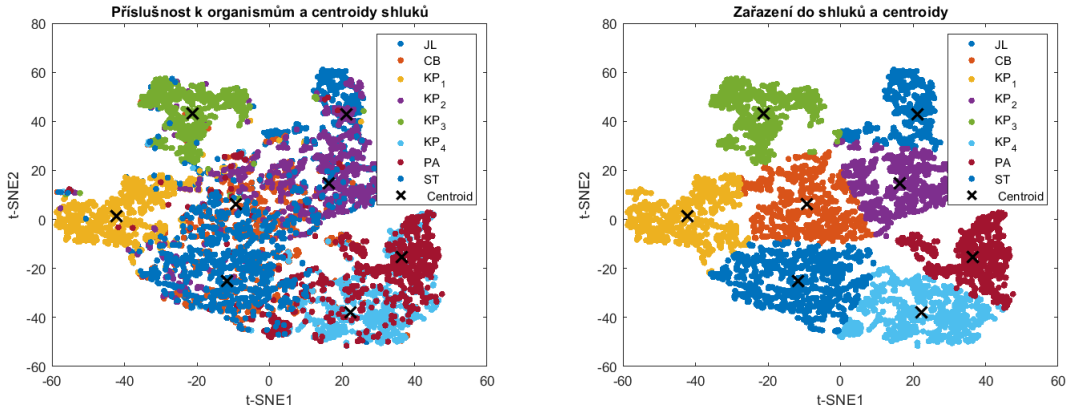
Jelikož se jedná o více než dvourozměrná data, která by bylo problém vizualizovat, byla provedena statistická analýza t-SNE, která ke každému datovému bodu dá umístění na dvourozměrné mapě[41].



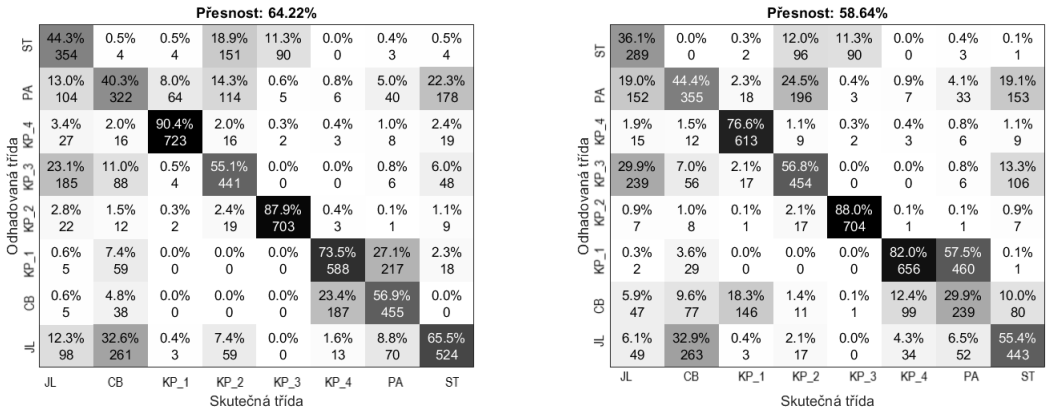
Obr. 3.4: Vizualizace rozložení dat a shlukování pro parametry odchylky, mobility, komplexity, varičního koeficientu, špičatosti, střední hodnoty, aktivity a rozptylu.

.

Na obrázku 3.4 a 3.5 můžeme pozorovat příslušnosti organismů ke shlukům a vedle jejich zařazení. Na první pohled je možné rozlišit shluky vzorků KP_3, KP_1. U ostatních vzorků sice lze pozorovat shlukování kolem jednoho centroidu, data se však překrývají, a to zhoršuje přesnost klasifikace.



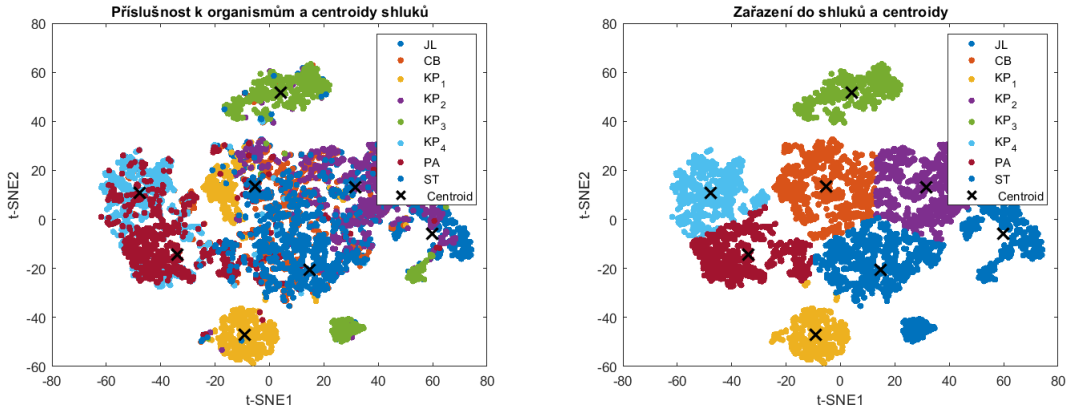
Obr. 3.5: Vizualizace rozložení dat a shlukování pro parametry odchylky, mobility, komplexity, varičního koeficientu, špičatosti, střední hodnoty, aktivity



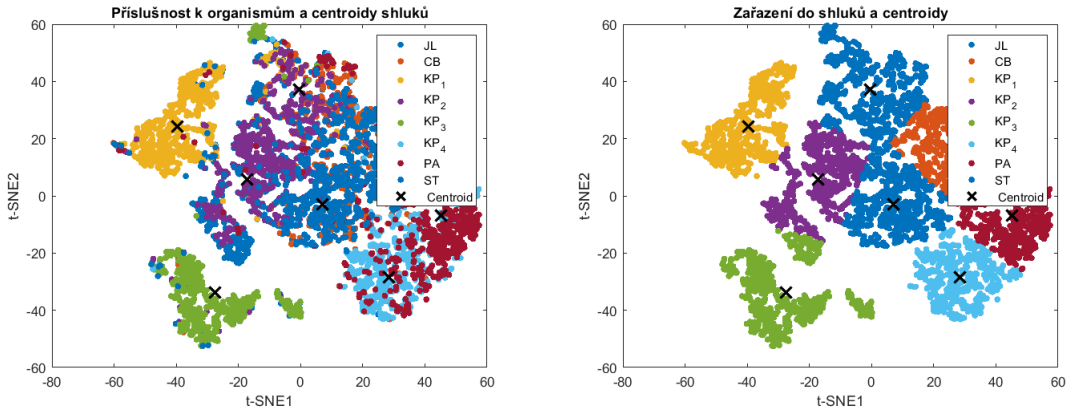
Obr. 3.6: Konfuzní matice, klasifikace do 8 tříd - vlevo pro obr. 3.4 a vpravo pro obr. 3.5

.

Výsledkem shlukování vybraných parametrů jsou konfuzní matice na obr 3.6. Můžeme vidět, že shlukování odchylky, mobility, komplexity, varičního koeficientu, špičatosti, střední hodnoty a aktivity dosahuje horších výsledků, než když klasifikujeme vzorky s výše zmíněnými parametry a navíc přidaným parametrem rozptylu. Přesnost je větší o necelých 6 %.



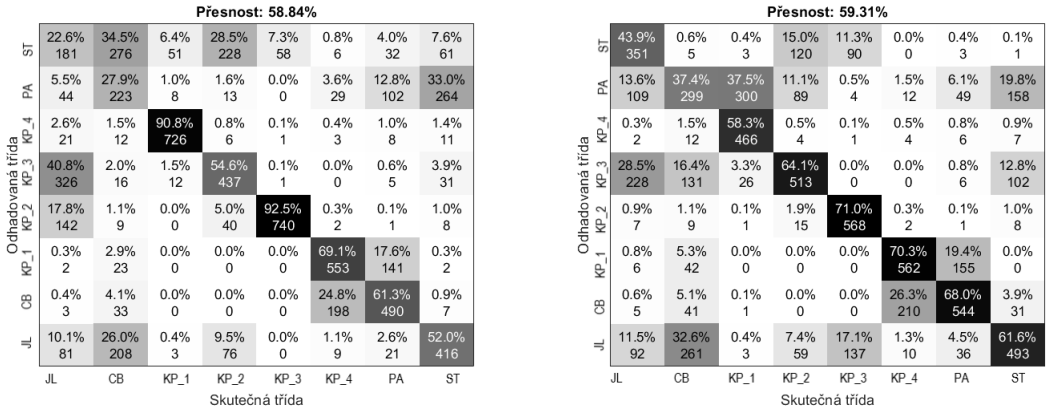
Obr. 3.7: Vizualizace rozložení dat a shlukování pro parametry odchylky, mobility, komplexity, varičního koeficientu, špičatosti, střední hodnoty.



Obr. 3.8: Vizualizace rozložení dat a shlukování pro parametry odchylky, mobility, komplexity, varičního koeficientu, špičatosti

.

Na obrázku 3.7 vidíme skutečné rozložení shluků a jejich zařazení. Vzorek KP_3 tvoří více shluků a data nejsou schopna tvořit disjunktní vzorky. I přes tuto skutečnost, přesnost klasifikace dosahuje 58.84 %. Nejlépe se podařilo rozlišit vzorek KP_1 a KP_3 kdy je přesnost klasifikace okolo 90 %. Přesnost metody na obr 3.8 sice v žádném vzorku nepřesahuje hodnotu 70 %, dokáže však s touto přesností rozlišit více vzorků, a ne pouze dva, jak je tomu v předchozím případě.

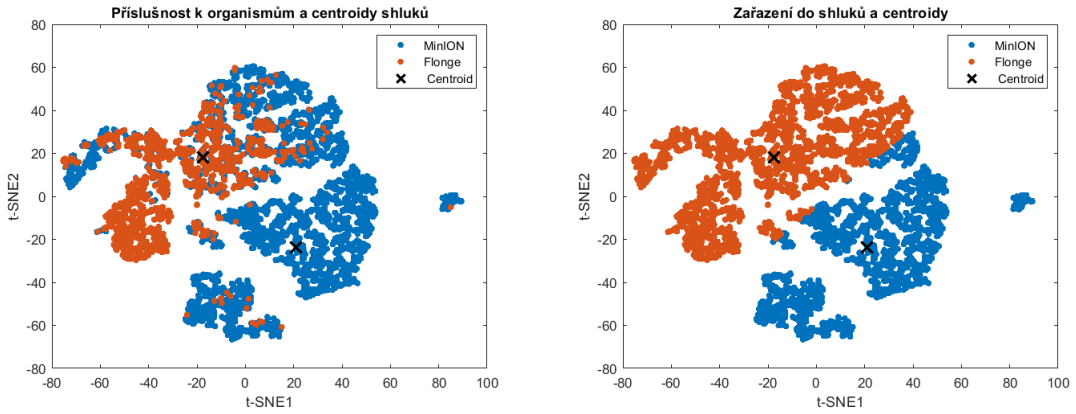


Obr. 3.9: Konfuzní matice, klasifikace do 8 tříd - vlevo pro obr 3.7 a vpravo pro obr3.8

.

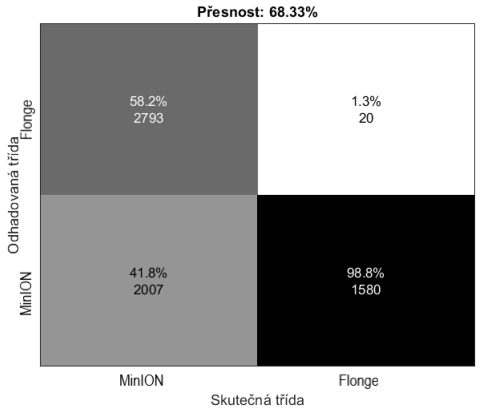
Klasifikace všech organismů pomocí vybraných výsledků nepřinesly příliš uspokojivé výsledky. Bylo zjištěno, že čím méně parametrů použijeme, tím menší je přesnost, naopak pokud použijeme většinu prametrů, přesnost se zvyšuje. Největší celková přesnost byla okolo 64 %, což není vyloženě špatný výsledek a některé vzorky např. KP_1 byly rozlišeny s přesností až 90 %, což znamená, že tato metoda má potenciál, jen je třeba buď shlukovat méně organismů, nebo lépe filtrovat vstupní data a eliminovat odlehlé hodnoty. V praxi je však potřeba dosáhnout mnohem lepších výsledků i v různých podmínkách a samozřejmě, čím větší množství organismů je schopna metoda rozlišit, tím lépe.

Shlukování podle počtu sekvenačních přístrojů



Obr. 3.10: Vizualizace rozložení dat a shlukování pro parametry střední hodnoty, aktivity a mobility

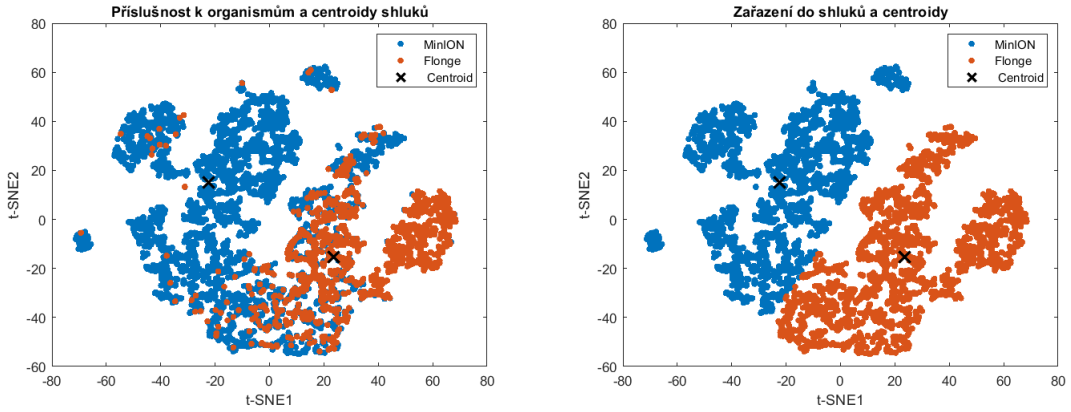
.



Obr. 3.11: Konfuzní matice, klasifikace do 2 tříd

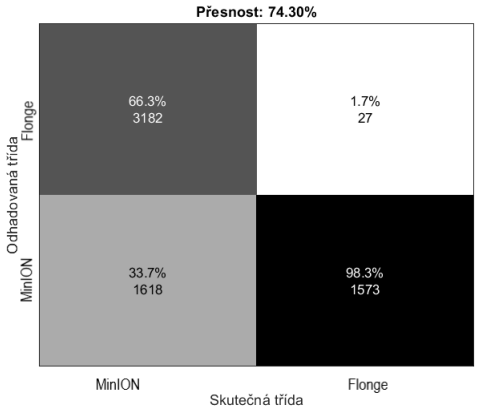
.

Na obrázcích můžeme vidět zřetelné rozlišení shluků do dvou skupin. Byla použita matice příznaků skládajících se ze střední hodnoty, aktivity a mobility. Tyto příznaky dosahují druhých nejlepších výsledků. Jakmile se počet příznaků zvýší, přesnost klasifikace se snižuje.



Obr. 3.12: Vizualizace rozložení dat a shlukování pro parametr střední hodnoty, variačního koeficientu a mobility

.

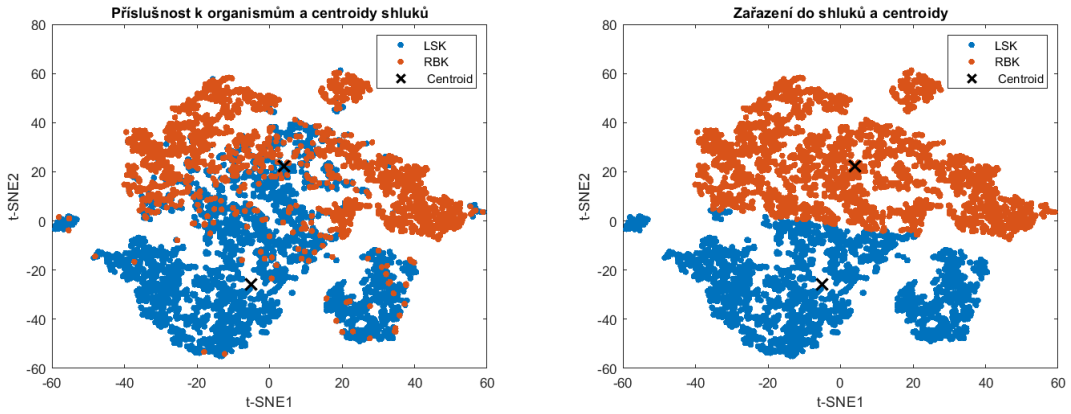


Obr. 3.13: Konfuzní matice, klasifikace do 2 tříd

.

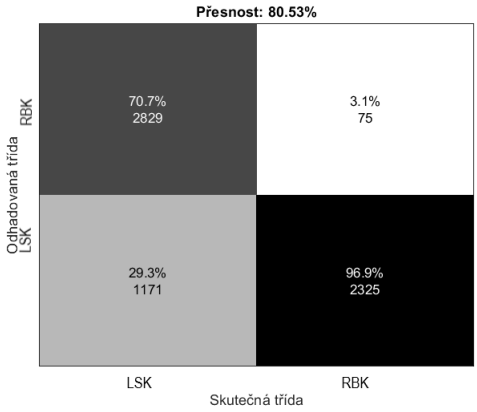
Nejlepších výsledků dosahovala kombinace střední hodnoty, variačního koeficientu a mobility. Přesnost je tedy okolo 74 %. Závěrem tedy lze říci, že klasifikace na základě sekvenačních přístrojů je přesná a byl potvrzen vliv sekvenačních přístrojů na výslednou sekvenaci a signály. To bylo potvrzeno i v předchozí analýze, kdy vzorky KP_1, KP_2, KP_3 a KP_4, ač jsou z jednoho organismu, jsou vyhodnoceny jako různé organismy.

Shlukování podle počtu sekvenačních kitů



Obr. 3.14: Vizualizace rozložení dat a shlukování pro parametry střední hodnoty, variačního koeficientu, mobility, aktivity, rozptylu, komplexity a odchylky

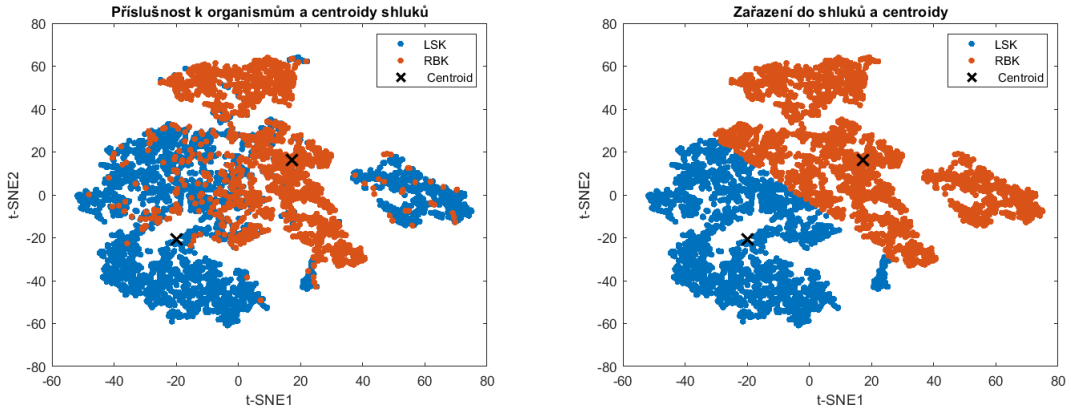
.



Obr. 3.15: Konfuzní matice, klasifikace do 8 tříd

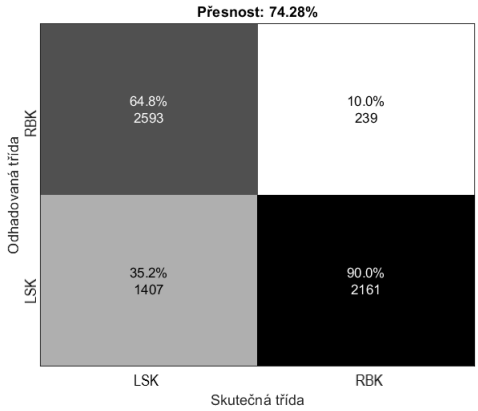
.

Byly použity příznaky střední hodnoty, variability, mobility, aktivity, rozptylu, komplexity a odchylky. Přesnost se pohybuje okolo 80 %. U příznaků střední hodnoty, variability, mobility, aktivity, rozptylu, komplexity je přesnost klasifikace okolo 74 %.



Obr. 3.16: Vizualizace rozložení dat a shlukování pro parametry střední hodnoty, variačního koeficientu, mobility, aktivity, rozptylu, komplexity

.



Obr. 3.17: Konfuzní matice, klasifikace do 2 tříd

.

Závěrem lze říci, že větších přesností bylo dosaženo, pokud bylo použito více parametrů. Pokud bylo použito méně parametrů, byla přesnost výrazně nižší. Z toho vyplývá, že pokud použijeme více statistických parametrů, je možné pomocí kmeans klasifikovat vzorky podle použité sekvenační chemie až s přesností 80 %. Naopak pokud použijeme parametrů méně, je možné pomocí stejného postupu klasifikovat s podobnou, nebo o něco menší přesností, jakým sekvenačním zařízením byly vzorky sekvenovány. Tím je tedy potvrzeno, že volba sekvenačního kitu nebo sekvenačního zařízení hraje velkou roli. Znamená to, že každý nástroj ONT generuje mírně odlišné signály a tudíž je potřeba tento fakt brát v úvahu během basecallování signálů.

Závěr

Cílem teoretické části této semestrální práce bylo stručně představit historii sekvenování, seznámit čtenáře s třetí generací sekvenátorů, principy sekvenování a různými sekvenačními nástroji od společnosti ONT. V praktické části práce byly vytvořeny dva datasety. Na těchto datech byl z každého signálu proveden výpočet vybraných parametrů z časové oblasti. Pro každou datovou sadu byl vypočítán Shapiro-Willkův test, kdy bylo zjištěno, že data nevykazují normální rozložení. Poté byl proveden Kruskal-Wallisův test, který vyhodnotil a následně eliminoval nevhodné parametry pro další shlukovou analýzu. Dále byl proveden Tukeyho test. Výsledkem jsou dvojice organismů a v případě organismu *Klebseilla pneumoniae* i sekvenační kity a chemie, které se od sebe dají odlišit.

Na prvním datasetu *Clostridium beijerinckii* bylo ověřeno, že vypočítané statistické parametry se na různých částech signálu a různých čteních stejného sekvenačního běhu příliš nemění a nelze je od sebe spolehlivě rozlišit. Poté byl stejný postup zopakován na zvoleném datasetu obsahující 5 organismů a 2 různé druhy sekvenačních kitů a přístrojů. Rozdíl je patrný již z krabicových grafů. Některé parametry byly schopny organismy a přístroje rozlišit lépe než jiné např: parametr střední hodnoty. Některé výsledky přinesly překvapivé zjištění, že i v případě sekvenování jednoho organismu můžeme získat úplně jiné výsledky. To je nejvíce patrné na parametru mobility, kde i různé vzorky organismu *Klebseilla pneumoniae* vykazují výrazné odlišnosti. Pravděpodobným důvodem je tedy, jak již bylo výše zmíněno, použití různých nástrojů ONT.

Dalším krokem bylo shlukování na základě vybraných parametrů. Po provedení shlukové analýzy na vybraných parametrech bylo provedeno hierarchické shlukování basecallovaných sekvencí. Další shlukování basecallovaných sekvencí se ukázalo jako zbytečné a nic nevypovídající, protože signály jsou vybrány náhodně. Z toho důvodu se v datasetech s velkou pravděpodobností mohou objevovat čtení z nekódujících úseků genomu, které nejsou druhově specifické. Proto tedy shlukování bylo provedeno na genu 16s rRNA, abychom mohli porovnat fylogenetické stromy.

Dále bylo provedeno shlukování statistických parametrů pomocí algoritmu k-means. Byly měněny počty shluků podle toho, zda jsme porovnávali rozlišení mezi jednotlivými vzorky, nebo sekvenačními nástroji.

Analýza ukázala, že signály jsou druhově specifické. Dále také, že různé sekvenační nástroje produkují specifické signály a nehledě na to, o jaký organismus se jedná, je možné je klasifikovat na základě vybraných příznaků s celkem velkou přesností.

Literatura

[1] HEATHER, James M. a Benjamin CHAIN. The sequence of sequencers: The history of sequencing DNA. Genomics. January 2016, 2016(107), 1-8. Dostupné z: doi:10.1016/j.ygeno.2015.11.003

[2] MRAZ, M., K. MALINOVA, J. MAYER a S. POSPISILOVA. MicroRNA isolation and stability in stored RNA samples [online]. 390. Biochemical and Biophysical Research Communications, 2009, 1-4 [cit. 2021-10-29]. ISSN 0006-291X. https://doi.org/10.1016/j.bbrc.2009.09.061. Dostupné z: https://www.sciencedirect.com/science/article/pii/S0006291X09018634

[3] 3 HEATHER, James M. a Benjamin CHAIN. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016, 107(1), 1-8. ISSN 08887543. DOI:10.1016/j.ygeno.2015.11.003

[4] Oxford Nanopore Technologies. [Online] In: [cit. 28.10.2021]. Dostupné z: https://nanoporetech.com

[5] DUMSCHOTT, Kathryn, Maximilian H-W SCHMIDT, Harmeet Singh CHAWLA, Rod SNOWDON a Björn USADEL. Oxford Nanopore sequencing: new opportunities for plant genomics? Journal of Experimental Botany. 2020, 2020(18), 5313-5322.

[6] SLATKO, Barton E., Andrew F. GARDNER a Frederick M. AUSUBEL. Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology. Blackwell Publishing Inc., 2018, roč. 122, č. 1, s. e59. ISSN 19343647. DOI: 10.1002/cpmb.59

[7] LANNOY, Carlos de, Dick de RIDER a Judith RISSE. The long reads ahead: de novo genome assembly using the MinION. F1000Research. 2017, 1083(6). Dostupné z: doi:https://doi.org/10.12688/f1000research.12012.2)

[8] Types of nanopores. Oxford Nanopore Technologies [online]. [cit. 2021-10-29]. Dostupné z: https://nanoporetech.com/how-it-works/types-of-nanopores

[9] MAGLIA, Giovanni, Andrew J. HERON, David STODDART a Deanpen JAPRUNG. Analysis of single nucleic acid molecules with protein nanopores. Methods in enzymology. 2010, 2010(475), 591-623. Dostupné z: doi:10.1016/S0076-6879(10)75022-9

[10] JAIN, Miten. Bioinformatic Analysis of Nanopore Data. BRANTON, Daniel a David DREAMER. Nanopore sequencing: An introduction. 1. Word Scientific, 2019, s. 147-158.

[11] MANRAO, Elizabeth A., Ian DERRINGTON, Andrew H. LASZLO a Kyle W. LANGFORD. Reading DNA at a Single-Nucleotide Resolution with a Mutant MspA nanopore and phi29 DNA Polymerase. Nature Biotechnology. 2012, 2012(30), 349-353. Dostupné z: doi:10.1038/nbt.2171

[12] JAIN, Miten, Hugh E. OLSEN, Benedict PATEN a Mark AKESON. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biology. 2016, 2016(239), 17. Dostupné z: doi:10.1186/s13059-016-1103-0

[13] MATEI, David, L. J. DURSI, Yao DELIA, Paul C. BOUTROS a Jared T. SIMPSON. Nanocall: An Open Source Basecaller for Oxford Nanopore Sequencing Data. BioRxiv [online]. Oxford, England, 2016, 33(1), 49-55. Dostupné z: doi:https://doi.org/10.1101/046086

[14] ZHANG, Yao-zhong, Arda AKDEMIR, Georg TREMMEL, Seiya Imoto IMOTO, Satoru Miyano MIYANO, Tetsuo Shibuya SHIBUYA a Rui YAMAGUCHI. Nanopore basecalling from a perspective of instance segmentation. BMC Bioinformatics. 2020, 2020(3), 21. Dostupné z: doi:https://doi.org/10.1186/s12859-020-3459-0

[15] XUAN, Lv, Chen ZHIGUANG, Lu YUTONG a Yang YUEDONG. An End-to-end Oxford Nanopore Basecaller Using Convolution-augmented Transformer. BioRvix. 2020, 2020. Dostupné z: doi:https://doi.org/10.1101/2020.11.09.374165

[16] RANG, Franka J., Wigard P. KLOOSTERMAN a Jeroen DE RIDDER. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. 13 .7.2018n. l., 2018(1), 19. Dostupné z: doi:https://doi.org/10.1186/s13059-018-1462-9

[17] WICK, Ryan R., Louise M. JUDD a Kathryn E. HOLT. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biology. 2019, 2019(1), 20. Dostupné z: doi:https://doi.org/10.1186/s13059-019-1727-y

[18] JINGWEN, Zeng, Cai HONGMIN, Peng HONG, Wang HAIYAN, Zhang YUE a Aktsu TATSUA. Causalcall: Nanopore Basecalling Using a Temporal Convolutional Network. Frontiers in Genetics. 2020, 2020(10), 1332. ISSN 1664-8021. Dostupné z: doi:10.3389/fgene.2019.01332

[19] OXFORD NANOPORE TECHNOLOGIES. Guppy Protocol. 2020. [On- line]. Dostupné z: https://community.nanoporetech.com/protocols/Guppy- protocol/v/gpb_2003_v1_revt_14dec2018

[20] BAO, Yuwei, Jack WADDEN, John R ERB-DOWNWARD, Piyush RANJAN, Robert P DICKSON, David BLAAUW a Joshua D WELCH. *Real-Time, Direct Classification of Nanopore Signals with SquiggleNet. bioRxiv [online]. 2021, 2021.01.15.426907.*

[21] Oxford Nanopore Technologies - Promethion. [Online] In: [cit. 28.10.2021]. Dostupné z: https://nanoporetech.com/products/promethion

[22] LU, Hengyun, Francesca GIORDANO a Zemin NING. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics, Proteomics & Bioinformatics. 2016(14,5), 265–279. Dostupné z: doi:10.1016/j.gpb.2016.05.004

[23] Oxford Nanopore Launches Flongle for Rapid, Smaller DNA/RNA Sequencing Tests in Any Environment. PR Newswire [online]. Oxford, England: Oxford Nanopore Technologies, 2019 [cit. 2021-10-29]. Dostupné z: https://www.prnewswire.com/news-releases/oxford-nanopore-launches-flongle-for-rapid-smaller-dnarna-sequencing-tests-in-any-environment-300811090.html

[24] Oxford Nanopore Technologies - Voltrax. [Online] In: [cit. 28.10.2021]. Dostupné z: https://nanoporetech.com/products/voltrax

[25] Home/Products/Devices. Oxford Nanopore Technologies [online]. 2021 [cit. 2021-10-29]. Dostupné z: https://store.nanoporetech.com/eu/devices.html

[26] SEDLÁŘ, K. Methods for comparative analysis of metagenomic data. Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Biomedical Engineering, 2018. 124 p. Doctoral thesis. Doctoral thesis supervisor: prof. Ing. Ivo Provazník, Ph.D.

[27] PAYNE, Alexandr, Nadine HOLMES, Vardhman RAKYAN a Matthew LOOSE, BIROL, Inanc, ed. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics [online]. 2019, 35(13), 2193–2198. [cit. 2021-11-06]. Dostupné z: doi:10.1093/bioinformatics/bty841

[28] ZHANG, Hongen. Overview of Sequence Data Formats. Methods in molecular biology: Clifton, N.J. 2016, 2016(1418), 3-17. Dostupné z: doi:10.1007/978-1-4939-3578-9_1

[29] BALESTRASSI, P. P., A. P. OAIVA, A. C. Zambroni DE SOUZA, Elmira POPOVA a J. B. TURRIONI. A multivariate descriptor method for change-point detection in nonlinear time series. Journal of Applied Statistics. 2011, 28(2), 327-342. ISSN 0266-4763. Dostupné z: doi:10.1080/02664760903406496

[30] VOURKAS, M., S. MICHELOYANNIS a G. PAPADOURAKIS. Use of ANN and Hjorth parameters in mental-task discrimination: 2000 First International Conference Advances in Medical Signal and Information Processing (IEE Conf. Publ. No. 476). Bristol, UK: IET, 4-6.92000n. l. ISBN 0-85296-728-4. ISSN 0537-9989.

[31] PAVLÍK, Tomáš a Ladislav DUŠEK. Biostatistika. Vyd. 1. Brno: Akademické nakladatelství CERM, 2012, 131 s. ISBN 978-80-7204-782-6.

[32] MATLAB, 2020. 9.8.0.1359463 (R2020a), Natick, Massachusetts: The MathWorks Inc.

[33] Understanding Boxplots [online]. Towards data science, 2018 [cit. 2021-11-07]. Dostupné z: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

[34] Öner, M., & Deveci Kocakoç, Ý. (2017). JMASM 49: A Compilation of Some Popular Goodness of Fit Tests for Normal Distribution: Their Algorithms and MATLAB Codes (MATLAB). Journal of Modern Applied Statistical Methods, 16(2), 30. Copyright (c) (2016) Öner, M., Deveci Kocakoc, I.

[35] OLSON, Cf. Parallel algorithms for hierarchical clustering. Parallel computing [online]. Vol. 21, 1995, pp. 1313–1325. ISSN 01678191. Available from: doi:10.1016/0167- 8191(95)00017-I

[36] GORI, F., D. MAVROEDIS, M. JETTEN a E. MARCHIORI. Genomic signatures for metagenomic data analysis: Exploiting the reverse complementarity of tetranucleotides. In: Systems Biology (ISB), 2011 IEEE International Conference on Information Visualisation. Zhuhai: IEEE Publishing, 2011, s. 149-154. ISBN 9781457716614.

[37] ČEPEK, M. Shluková analýza: přednáška předmětu Základy vytěžování dat [online]. Katedra kybernetiky a katedra počítačů, FEL, ČVUT v Praze, [cit. 2015-12-05]. Dostupné z: <http://data.cedupoint.cz/oppa_e-learning/1_STM/15.pdf>.

[38] PAVLÍK, T. Asociace ve čtyřpolní tabulce a základy korelační analýzy [online]. Institut biostatistiky a analýz Masarykovy univerzity, 2011 [cit. 2015-12-05]. Dostupné z: <http://www.iba.muni.cz/esf/res/file/bimat-prednasky/biostatistika-pro-matematickou-biologii/BpMB-11.pdf>.

[39] JANOUŠOVÁ, E. E-learningová učebnice matematické biologie: Vícerozměrné metody pro analýzu a klasifikaci dat. Institut biostatistiky

a analýz Masarykovy univerzity [online]. [cit. 2016-01-01]. Dostupné z: <http://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat–vicerozmerne-metody-pro-analyzu-dat>.

[40] JIANG, B., K. SONG, J. REN, M. DENG, F. SUN a X. ZHANG. Comparison of metagenomic samples using sequence signatures. BMC Genomics [online]. London: BioMed Central, 2012, (13) [cit. 2015-12-31]. DOI: 10.1186/1471-2164-13-730.

[41] VAN DER MAATEN, L.J.P.; HINTON, G.E. (Nov 2008). "Visualizing Data Using t-SNE", Journal of Machine Learning Research. 9: 2579–2605.

[42] VANĚČKOVÁ, T. Numerické metody pro klasifikaci metagenomických dat. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2016. 59 s. Vedoucí diplomové práce: Ing. Helena Škutková, Ph.D.

[43] WEI YOU, KUN WANG, HUIXIAO LI, YANG JIA, XIAOQIN WU, YANING DU, Classification of DNA Sequences Basing on the Dinucleotide Compositions, Department of Mechanical and Electrical Engineering, North China Institute of Science and Technology 2009

Seznam symbolů a zkratek

ONT Oxford Nanopore Technologies

PHRED score Quality score

DNA deoxyribonukleová kyselina

PCR polymerázové řetězové reakce

SMRT Single molecule real time sequencing

ASCII American Standard Code for Information Interchange

RNA ribonukleová kyselina

ddATP Dideoxyribonukleotid

ddCTP Dideoxycytidintrifosfát

ddGTP 2',3'-Dideoxyguanosine 5'-triphosphate

ddTTP 2',3'-dideoxythymidin-5'-triphosphate

USB Universal Serial Bus

Seznam příloh

A Obsah elektronické přílohy

- `/` ..... kořenový adresář přiloženého archivu
  - `dataset_analyza` ..... soubory potřebné pro analýzu datasetu
    - `aktivita.mat`
    - `CB_MIN_LSK`
    - `HjorthParameters_cb.mat`
    - `JL_MIN_RBK`
    - `KP_FL_LSK`
    - `KP_FL_RBK`
    - `KP_MIN_LSK`
    - `KP_MIN_RBK`
    - `Main_dataset_final`
    - `odchylka.mat`
    - `PA_MIN_LSK`
    - `rozptyl.mat`
    - `RR_MIN_LSK`
    - `sikmost.mat`
    - `spicatost.mat`
    - `ST_MIN_LSK`
    - `stredni_h.mat`
    - `var_coef.mat`
  - `shlukovani` ..... shluková analýza kmeans
    - `aktivita`
    - `komplexita`
    - `mobilita`
    - `odchylka`
    - `renum_clust.mat`
    - `rozptyl`
    - `sikmost`
    - `skript_nehier_shlukovani.mat`
    - `spicatost`
    - `stredni_h`
    - `variacni_koeficient`
  - `tvorba_datasetu` ..... tvorba datasetu
    - `load_data.mat`
    - `make_passed_reads`