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

NUMERICKÉ METODY PRO KLASIFIKACI METAGENOMICKÝCH DAT

NUMERICAL METHODS FOR CLASSIFICATION OF METAGENOMIC DATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. Tereza Vaněčková
AUTHOR

VEDOUCÍ PRÁCE — Ing. Helena Škutková, Ph.D.
SUPERVISOR

BRNO 2016

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Biomedicínské inženýrství a bioinformatika**
Ústav biomedicínského inženýrství

Studentka: Bc. Tereza Vaněčková ***ID:*** 147522
Ročník: 2 ***Akademický rok:*** 2015/16

NÁZEV TÉMATU:

Numerické metody pro klasifikaci metagenomických dat

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s pojmem metagenomika, výpočetními metodami využívanými pro zpracování metagenomu a veřejnými databázemi metagenomických dat. 2) Vypracujte literární rešerši metod pro klasifikaci organismů na základě taxonomicky specifických četností nukleotidových slov v DNA sekvenci. 3) Vypracujte fylogenetickou analýzu charakteristických četností nukleotidových slov u metagenomů dostupných z veřejných databází. 4) Realizujte alespoň tři odlišné metody klasifikace DNA sekvencí na základě četností nukleotidových slov. 5) Otestujte zvolené metody pro účely metagenomické klasifikace a identifikace organismů na veřejně dostupných záznamech metagenomů. 6) Vypracujte statistické srovnání použitých metod a diskuzi výsledků.

DOPORUČENÁ LITERATURA:

[1] LACZNY, C. C., N. PINEL, N. VLASSIS and P. WILMES. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. Scientific Reports. 2014-3-31, vol. 4, pp. 1-12.

[2] VINGA, SUSANA a JONAS ALMEIDA. Alignment-free sequence comparison—a review. Bioinformatics, March 1, 2003 2003, 19(4), 513-523.

Termín zadání: 8.2.2016 ***Termín odevzdání:*** 20.5.2016

Vedoucí práce: Ing. Helena Škutková, Ph.D.
Konzultant diplomové práce:

prof. Ing. Ivo Provazník, Ph.D., *předseda oborové rady*

ABSTRAKT

Tato práce se zabývá metagenomikou a výpočetními metodami využívanými pro zpracování metagenomu. Literární rešerše metod nevyžadujících zarovnání ukázala, že metody založené na studiu taxonomicky specifických četností nukleotidových slov se jeví jako vhodný a dostatečně účinný nástroj pro zpracování metagenomických čtení sekvenačních technologií nové generace. Pro vyhodnocení potenciálu těchto metod byly testovány vybrané příznaky založené na studiu četností nukleotidových slov na sadě simulovaných metagenomických čtení. Analýza byla provedena pro různou délku slov a vyhodnocena s ohledem na úspěšnost klasifikace pomocí hierarchického shlukování v originálním datovém prostoru a K-means shlukování v redukovaném datovém prostoru.

KLÍČOVÁ SLOVA

Metagenomika, technologie sekvenování, nukleotidová slova, *k*-mery, hierarchické shlukování, PCA, K-means shlukování, metody nevyžadující zarovnání

ABSTRACT

This thesis deals with metagenomics and numerical methods for classification of metagenomic data. Review of alignment-free methods based on nucleotide word frequency is provided as they appear to be effective for processing of metagenomic sequence reads produced by next-generation sequencing technologies. To evaluate these methods, selected features based on *k*-mer analysis were tested on simulated dataset of metagenomic sequence reads. Then the data in original data space were enrolled for hierarchical clustering and PCA processed data were clustered by K-means algorithm. Analysis was performed for different lengths of nucleotide words and evaluated in terms of classification accuracy.

KEYWORDS

Metagenomics, sequencing technologies, nucleotide words, *k*-mers, hierarchical clustering, PCA, K-means clustering, alignment-free methods

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.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma Numerické metody pro klasifikaci metagenomických dat jsem vypracovala samostatně pod vedením vedoucí 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 autorka uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové 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 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

Na tomto místě bych ráda poděkovala vedoucí své diplomové práce slečně Ing. Heleně Škutkové, Ph.D. za odbornou pomoc, motivaci, čas strávený při konzultacích a cenné rady při zpracování diplomové práce.

Dále děkuji mé rodině a nejbližším za podporu během celého studia.

V Brně dne ....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Metagenomika představuje poměrně nový vědní obor, který se zabývá studiem genetického materiálu izolovaného přímo z prostředí. Umožňuje studium velkého množství mikroorganismů bez předchozí kultivace.

S rozvojem nové generace sekvenovacích technologií vzrostl i počet metagenomických projektů s cílem pochopit ekologickou roli, metabolismus a evoluci mikrobiálních společenství. Je zde prostor pro vývoj nových nástrojů pro analýzu a porovnávání metagenomických datasetů se zaměřením na rychlou a uživatelsky přívětivou implementaci těchto přístupů.

Cílem této práce je studium numerických metod využívaných pro klasifikaci metagenomických dat. Práce je zaměřena na studium taxonomicky specifických nukleotidových slov v metagenomických čteních. V odborné anglické literatuře se řadí do skupiny tzv. *alignment-free* metod, tedy metod nevyžadujících zarovnání. Mezi jejich výhody patří zejména nižší výpočetní náročnost.

V první části práce jsou uvedeny obecné přístupy metagenomiky, metody sekvenování se zaměřením na technologie nové generace a databáze metagenomických dat. Navazuje přehled výpočetních metod pro zpracování metagenomu, dělený na základě přístupu sekvenování – shotgun a amplikon. Třetí část je věnována numerickým metodám vyhodnocení sekvenčních příznaků založených na studiu četnosti nukleotidových slov. Ve čtvrté části jsou popsány algoritmy pro klasifikaci dat, redukci dimenzionality, a také metodika hodnocení úspěšnosti klasifikace.

Praktická část této práce je založena na analýze simulovaných metagenomických dat. V páté části jsou diskutovány charakteristiky simulovaných metagenomických datasetů, přehled testovaných metod a výsledky testování dílčího nastavení algoritmů. Představeny jsou rovněž poznatky a doporučení, které z analýzy vyplývají. Ukázka aplikace na reálných metagenomických datech je pak hodnocena v šesté kapitole.

1 METAGENOMICKÁ DATA

Tradiční mikrobiální celogenomové sekvenování závisí na schopnosti kultivace mikroorganismů, což může být obtížné, jelikož většinu mikroorganismů nelze kultivovat za standardních podmínek. Přístupy nezávislé na kultivaci využívající nové generace sekvenování DNA poskytly vědcům a lékařům nový pohled na studium biodiverzity mikrobiomu.

Až do nedávné doby se pro studium mikrobiálních společenstev využívalo nepřesné metody fingerprintingu nebo molekulárního klonování s využitím Sangerova sekvenování. Sangerovo sekvenování může poskytnout přesné informace o složení komunity a generovat dostatečně velké soubory dat pro srovnání komunity v celém společenství. Nevýhodou jsou však vysoké náklady a časová náročnost. S příchodem nové generace sekvenování se však charakteristika mikrobiálních komunit stala proveditelná a efektivní z hlediska nákladů a časové náročnosti. [1]

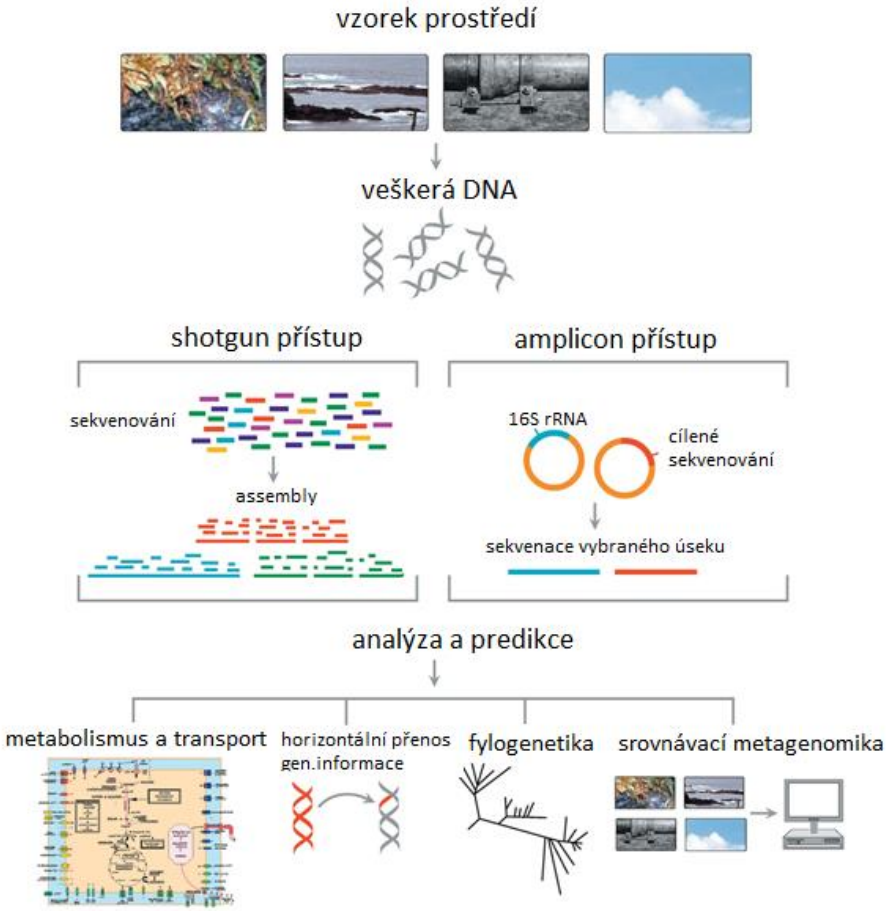
Metagenomika umožňuje hlubší pochopení ekologické role, metabolismu a evoluční historie mikrobů v daném ekosystému analýzou vzorku DNA bez předchozí kultivace. Přispěla k významným objevům v oblasti studia mikrobiálních komunit člověka, např. projekt *Human Microbiome Project* spojený s výzkumem vzorků střeva či ústní dutiny. [2] Jedním z mnoha objevů je například studium diverzity mikrobiomu spojeného s poraněními a osídlením ran určitým druhem mikrobiálních komunit. [3] Dalšími aplikacemi jsou také výzkumy zaměřené na životní prostředí, zabývající se studiem mořských bakteriálních komunit, či bakterií vyskytujících se v půdě a vzduchu. V neposlední řadě lze metagenomiku využít v potravinářském průmyslu pro analýzu bezpečnosti potravin, v zemědělství pro studium půdních vzorků, nebo v oblasti čištění odpadních vod a obnovitelných zdrojů energie. [4]

V této kapitole budou představeny dva přístupy sekvenování metagenomu – shotgun a amplikon. Dále budou uvedeny charakteristiky a principy sekvenovacích technologií se zaměřením na technologie sekvenování nové generace, které se v současnosti nejvíce využívají pro sekvenaci mikrobiomu. V závěrečné části je prezentován přehled databází surových dat z platforem nové generace sekvenování.

1.1 Přístupy sekvenování

Pro zkoumání a charakteristiku metagenomu se používají dva přístupy - „shotgun“ a „amplikon“, schematicky znázorněné na Obr. 1.1. Každý z nich má své výhody a nedostatky a jejich použití závisí na konkrétní aplikaci. Na základě volby přístupu se také odvíjí i volba sekvenovacích technologií a následných bioinformatických nástrojů pro zpracování a analýzu dat (dále v kapitole 2).

Metody založené na amplikonovém sekvenování jsou vhodné pro charakteristiku a porovnání celkového taxonomického a funkčního složení komunit bakterií a plísní. Shotgun sekvenovací metody jsou efektivní pro charakteristiku komunit mikroorganismů, a to zkoumáním funkčního potenciálu organismů, zkoumáním genomové DNA (metagenomika), nebo procesů probíhajících v buňkách zkoumáním RNA transkriptů (metatranskriptomika). [1]



Obr. 1.1: Schematické znázornění přístupů sekvenování (převzato a upraveno z [5])

1.1.1 Amplikonové sekvenování

Metody, které využívají amplikonové sekvenování populace, mají uplatnění zejména při studiu bakteriální části lidského mikrobiomu. Nejrozšířenějším genomickým regionem pro studium bakteriálních kmenů je gen kódující RNA malé podjednotky (*small subunit,* SSU), obvykle známý jako „16S rRNA“. Tento gen je považován za ideální díky tomu, že je přítomen u všech bakterií. Obsahuje konzervované úseky, které jsou univerzální téměř u všech bakterií a zároveň i variabilní regiony, které se značně liší mezi různými taxony. Proto je možné označit tento gen primery a použít ho k identifikaci taxonů. Primery, které se připojují k 16S regionu, jsou používány k PCR amplifikaci různých fragmentů genu nalezeného u různých organismů vzorku z prostředí. Tímto způsobem je produkována populace 16S amplikonů, která odráží složení komunity organismů v daném vzorku. Následná analýza knihovny sekvencí může být provedena za účelem odhalení korelací mezi určitými faktory a konkrétními taxony nebo změny v celkové struktuře komunity. [1; 6]

Amplikonové sekvenování vyžaduje menší množství vyizolované DNA (obvykle okolo 10 ng) a je efektivní z hlediska nákladů a časové náročnosti. Využití je vhodné zejména pro charakteristiku a porovnávání komunit. Může však poskytovat zkreslené výsledky, zejména kvůli použití specifických primerů a vícenásobných cyklů amplifikace. [1]

1.1.2 Shotgun sekvenování

Metoda shotgun (v překladu „brokovnice“) umožňuje profilování komunity založené na fragmentech celého genomu různých organismů (včetně virů, archeí a mikroeukaryot) v ní obsažených. Shotgun sekvenování metagenomu umožňuje zkoumat taxonomické složení a může poskytnout i informaci o funkci. [1; 7] Tento přístup je v současnosti jediným způsobem, jak zkoumat profil celé mikrobiální komunity. Umožňuje stanovení relativní četnosti různých organismů, neboť není prováděna amplifikace DNA, která může zavádět zkreslení.

Shotgun sekvenování je používáno za účelem pochopení funkce rozlišných mikrobiálních komunit. Obvykle je využíváno referenční databáze jako například KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [8] a COG (*Cluster of Orthologous Groups of proteins*) [9]. Nedostatkem shotgun sekvenování je, že analýza často zahrnuje srovnání mezi různými částmi různých genomů. Klasifikace může být nespolehlivá také proto, že je k dispozici pouze omezené množství referenčních sekvencí. Další nevýhodou je jeho časová a finanční náročnost z důvodu nutnosti sekvenace velkého množství vyizolovaného materiálu DNA.

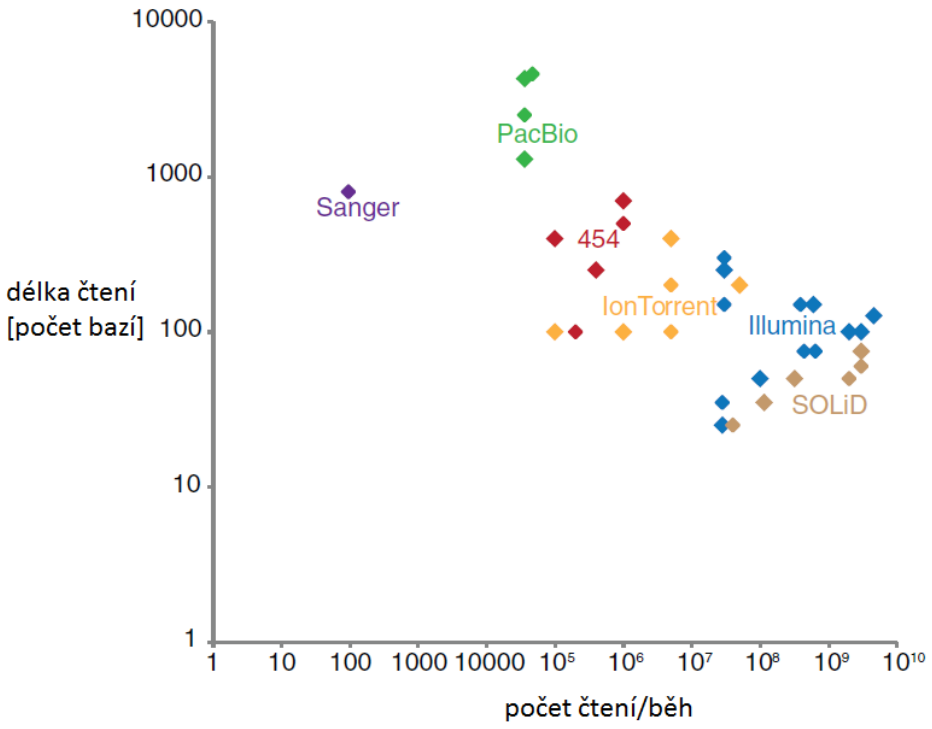
1.2 Technologie sekvenování

Sekvenování DNA slouží ke stanovení primární struktury, tedy pořadí nukleotidových bází v molekule DNA. V současnosti existuje řada metod sekvenování DNA. Nejstaršími klasickými metodami jsou chemická (Maxam-Gilbertova) a enzymová (Sangerova) metoda, které vznikly nezávisle na sobě již v roce 1977. Jsou obvykle označovány jako technologie první generace. Automatizovaná Sangerova metoda byla pro svou jednoduchost a spolehlivost považována za zlatý standard až do konce minulého století, avšak i v dnešní době má využití pro malé projekty. [1; 7]

Pro dosažení vyšší efektivity sekvenování začaly vznikat počátkem 21. století zcela nové metody, tzv. metody sekvenování nové generace (angl.. „*next-generation sequencing"*, NGS). Tyto nové technologie zahrnují různé strategie přípravy templátu, sekvenování a zobrazování, zarovnání genomu a metod kompletování. Vyznačují se masivním paralelním sekvenováním a vynikají vysokým výkonem, rychlostí a nízkými náklady na sekvenování. [7]

V posledních letech se na trhu objevují technologie založené na sekvenování jediné molekuly (angl. „*single-molecule sequencing*", SMS), označované též jako metody třetí generace. Tyto metody se od NGS liší tím, že nevyžadují amplifikaci vzorku DNA, který má být osekvenován. Jako metody třetí generace se označují komerčně dostupné systémy Helicos a systém vytvořený firmou Pacific Biosciences (SMRT). Další komerční novinkou je nanopórové sekvenování od firmy Nanopore Technologies a systém Starlight od Life Technologies (VisiGen). [7] Jejich výhodou je zejména produkce dlouhých čtení (až 20 kb a delší pro PacBio) a také rychlá délka běhu v rámci několika hodin. Nové platformy se neustále vyvíjejí. Jejich hlavním cílem je vytvořit technologie, které umožní produkovat delší čtení a více čtení za běh přístroje, tedy takové, které by zaplnily pravou horní část grafu (Obr. 1.2). [1; 10]

V této kapitole budou představeny vybrané sekvenovací technologie nové generace, které se výrazně uplatnily v oblasti studia metagenomu. Jedná se o technologie 454 pyrosekvenování, Genome Analyzer, systém SOLid a Iont Torrent. Přístup 454 pyrosekvenování je efektivní volbou pro aplikace, které vyžadují větší délku čtení. Ze třech výše zmíněných NGS technologií, Illumina HiSeq generuje největší množství dat za nejnižší cenu, SOLiD Systém má nejvyšší přesnost a Roche 454 poskytuje největší délku čtení. [10]



Obr. 1.2: Sekvenovací technologie znázorněné v prostoru charakterizovaném délkou čtení na ose X a počtem čtení za běh přístroje na ose Y. Jednotlivé skupiny platforem jsou vyznačeny stejnou barvou (převzato z [1])

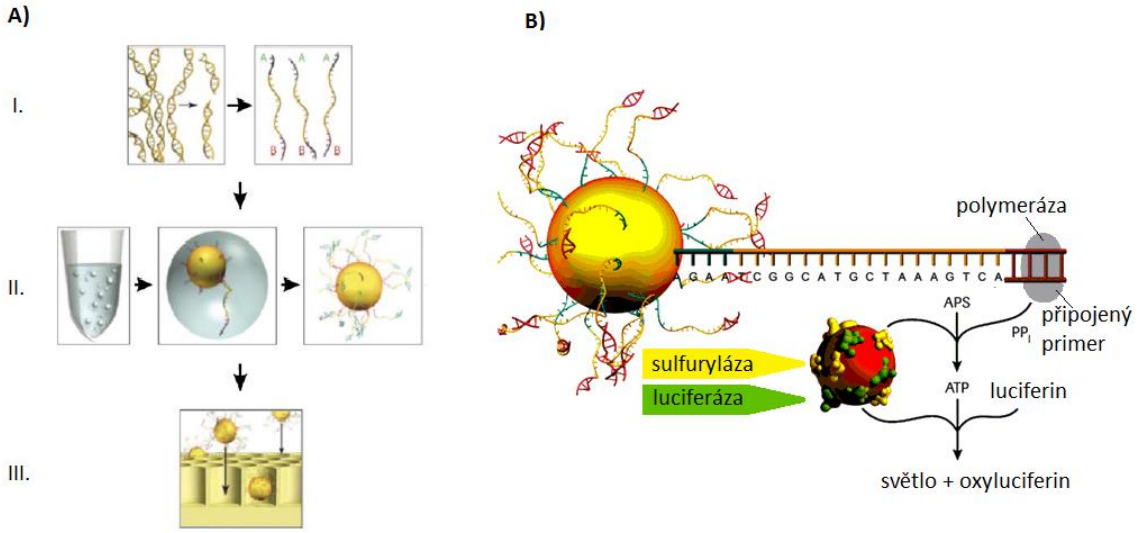
1.2.1 454 Life Sciences (Roche)

V roce 2005 představila firma Roche přístroj založený na principu pyrosekvenace a stal se tak prvním systémem NGS. [1]

Postup 454 pyrosekvenovací metody začíná přípravou vzorku DNA (knihovny). Delší vlákna jsou rozštěpena na fragmenty o velikosti 100-800 bp. Poté následuje připojení adaptorů A a B na 3‘ a 5‘ konce úseků a denaturace dvouvláknové DNA na jednovláknovou. Adaptor B obsahuje biotinovou značku, pomocí které jsou fragmenty přichyceny na magnetickou kuličku obalenou streptavidinem. Selektovány jsou pouze ty úseky DNA, které obsahují oba adaptory. Zbytek vzorku je odmyt. Takto vzniklé jednořetězcové fragmenty DNA jsou připraveny k hybridizaci ke speciálním DNA kuličkám, které mají na svém povrchu komplementární sekvenci DNA sloužící jako primer pro následnou amplifikaci. Technologie 454 využívá klonální amplifikace pomocí emulzní polymerázové řetězové reakce (emPCR). Výsledkem je mnoho kopií (kolem 10 milionů) templátu na každé kuličce. Poté jsou kuličky umístěny do jamek pikotitrační desky s optickým vláknem a promývány emulzí tak, aby do každé jamky zapadla jediná kulička. Do jamek jsou následně přidány enzymy nezbytné pro sekvenovací proces (DNA

polymeráza, ATP sulfuryláza a luciferáza). Přes pikotitrační destičku pak protéká roztok obsahující vždy jeden z nukleotidů (adenin, guanin, cytosin, thymin). Pokud je daný nukleotid komplementární k templátu, je DNA polymerázou začleněn do nově syntetizovaného řetězce a dojde k uvolnění pyrofosfátu. Pyrofosfát je díky enzymu ATP sulfuryláza přeměněn na ATP, jež umožní konverzi luciferinu na oxyluciferin katalyzované enzymem luciferázou. Při této reakci dochází k emisi viditelného světla. Pro zachycení světelného záblesku slouží citlivý CCD čip. [1; 11]

V současnosti nejpokročilejší platforma ve skupině 454 technologií (GS FLX+ Systém uvedený na trh v roce 2011) je schopna vyprodukovat až 1 milion čtení za běh s délkou čtení až 1000 bp. Díky poměrně velké délce čtení lze předpokládat i vyšší přesnost. Nedostatkem 454 pyrosekvenování však je, že často chybuje při čtení homopolymerů (úseky nukleotidů, kde jsou všechny báze identické). Kromě toho, při porovnání s ostatními NGS technologiemi je cena za bázi poněkud vyšší. Podpora této platformy končí v roce 2016. [1]



Obr. 1.3: (A) Postup sekvenování metodou 454: (I) konstrukce knihovny, (II) ligace specifických adaptorů na fragmenty DNA a napojení DNA na kuličky, (III) umístění kuliček do jamek pikotitrační desky. (B) Průběh pyrosekvenační reakce (převzato a upraveno z [12; 13]).

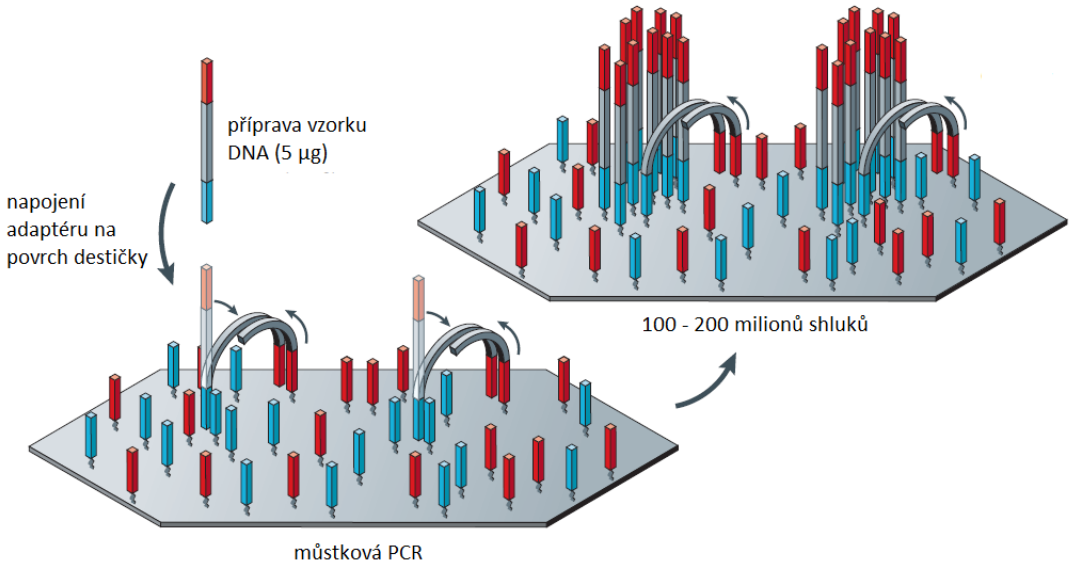
1.2.2 Genome Analyzer (Illumina)

Firma Illumina (dříve Solexa) uvedla na trh v roce 2006 sekvenátor Genom Analyzer. Přístroj je založen na stejném principu jako 454 Life Sciences, tedy sekvenaci pomocí syntézy. [1]

Příprava knihovny zahrnuje náhodné štěpení vzorku DNA na kratší úseky. Následně jsou konce fragmentů zarovnány, fosforylovány a jejich 3‘ konce adenylovány a napojeny

adaptory podobně jako u pyrosekvenování. Amplifikace vzorku probíhá tzv. můstkovou PCR. Po denaturaci jsou fragmeny upevněny adaptorem k povrchu reakční komůrky díky oligonukleotidům komplementárním k adaptorům a následně dochází k jejich ohnutí (odtud název „můstková“) a připojení druhého adaptoru k povrchu (Obr. 1.4). Při tomto procesu dochází k vytvoření shluků DNA na amplifikační destičce. Nakonec dojde k odstranění reverzních vláken. Sekvenační primery jsou hybridizovány k adaptorovým sekvencím a do reakční komůrky se shluky je vpravena směs polymerázy a čtyř rozdílně fluorescenčně značených nukleotidů. V každém cyklu je pokaždé začleněn pouze jeden nukleotid. Díky excitaci laserem lze pomocí CCD kamery zaznamenat pozici a typ nukleotidu. Terminační skupina na 3‘-konci nukleotidu i fluorescenční barva jsou poté odstraněny a cyklus je opakován až do přečtení celé sekvence. [1; 11]

V současnosti nabízí společnost Illumina několik přístrojů různé výkonnosti: MiSeq, MiSeqDx, NextGen 500, HiSeq2500 a HiSeq X Ten. MiSeq a Mi-SeqDx jsou sekvenátory vhodné spíše do malých laboratoří pro sekvenování malých genomů, amplikonové sekvenování a cílené sekvenování vybraných oblastí genů. NextGen 500 pak představuje vysokokapacitní stolní sekvenátor, který je vhodný na řadu DNA i RNA sekvenačních aplikací. HiSeq2500 a nejvýkonnější sekvenační platforma HiSeq X Ten jsou ultra vysokokapacitní sekvenátory, které jsou vhodné zejména pro velké sekvenační studie. [1; 11]



Obr. 1.4: Po připojení adaptoru k pevnému povrchu dojde k ohnutí vlákna a připojení fragmentu, čímž dochází k vytvoření „můstku" [7]

1.2.3 SOLiD (Applied Biosystems/Life Technologies)

V roce 2007 byla firmou Applied Biosystems (dnes Life Technologies) představena platforma SOLiD (z angl. „*sequencing by oligonukleotide ligation and detection*“). Jedná se o technologii založenou na sekvenování ligací. Příprava knihovny opět zahrnuje fragmentaci DNA, poté připojení krátkých adaptorů komplementárních k adaptorům imobilizovaným na povrchu magnetických kuliček. Dále následuje amplifikace a kovalentní navázání kuliček na sklíčko, které se vloží do kazety umožňující fluidní průtok. Systém SOLiD využívá sondy dlouhé osm nukleotidů. Každá ze sond má známou sekvenci prvních dvou bází a je označena jednou ze čtyř různých fluorescenčních barev. Každá barva tedy představuje 4 z 16 možných dinukleotidových sekvencí. [1; 11]

Technologie SOLiD se vyznačuje přesností pořadí nukleotidů sekvence, jelikož čte každý nukleotid dvakrát. Jedná se však o nejméně využívanou technologii NGS. [11]

1.2.4 Ion Torrent (Life Technologies)

Přístroj Ion Personal Genome Machine (PGM) se od ostatních systémů se liší způsobem detekce jednotlivých nukleotidů. Je založen na technologii schopné přímo převádět chemický signál do digitální podoby. Obsahuje destičku jamek s kuličkami, ke kterým jsou připojeny fragmenty DNA. Po přidání nukleotidu k rostoucímu DNA vláknu dojde k emisi vodíkových protonů, které mírně změní okolní pH. Mikrodetektory citlivé na změnu pH jsou připojeny k jamkám destičky, která se tak chová jako polovodičový čip a zaznamenává tyto změny. Pro přípravu knihovny se rovněž využívá emPCR. Kuličky s templátem jsou umístěny na čip tak, že v každé jamce je pouze jedna molekula DNA. V každém cyklu jsou na čip postupně přidávány jednotlivé druhy nukleotidů a dochází k syntéze DNA. Pokud nedojde ke komplementaritě nukleotidů, detektor zaznamená nulový signál. V případě začlenění dvou nukleotidů je signál dvojnásobný. [1; 11; 14]

Výhodou tohoto systému je zkrácení sekvenačního běhu (méně než 2 hodiny), vzhledem k tomu, že tato technologie nevyužívá fluorescenční detektory, kamery atp. Technologie je velice jednoduchá, rychlá a levná. Množství produkovaných dat je určeno hustotou jamek na čipu. [1; 11]

1.3 Databáze metagenomických dat

Sběr a uchovávání surových dat z platforem nové generace sekvenování a dalších informací o vzorcích je nezbytné pro umožnění interpretace dat, srovnávací analýzu a také k zajištění opakovatelnosti projektů. Tzv. „metadata“ zahrnují mimo jiné taky biochemická data (pH, teplotu, obsah solí), geografická data (souřadnice zeměpisné šířky a délky, nadmořskou výšku) a informace o zpracování vzorku (čas sběru vzorku, metoda DNA extrakce, typ sekvenovací platformy). Údaje ve specializovaných databázích metagenomických dat se mohou lišit v závislosti na typu vzorku. Pro umožnění vzájemné spolupráce těchto databází je nutné dodržování konzistence a standardizovaných formátů. [15; 16]

Genomes OnLine Database (GOLD) umožňuje komplexní přístup k informacím o metagenomických projektech po celém světě. [17]

Sequence Read Archive (SRA) je databáze surových dat a informací z platforem sekvenování, založena institucí *National Center for Biotechnology Information* (NCBI). Je součástí International *Nucleotide Sequence Database Collaboration* (INSDC) a pracuje na přístupových a přenosových protokolech svých sesterských databází *European Read Archive* (ERA) a *DNA Data Bank of Japan* (DDBJ). Preferovaný formát pro vstupní data je formát BAM, který je schopen uchovávat i informace o zarovnání k referenční sekvenci. SRA spolupracuje s *NCBI SRA Toolkit*, který je používán u všech tří členských databází INSDC za účelem poskytnutí komprese dat a konverze do jiných formátů, jako například FASTQ. [18]

Další webovou databází metagenomických dat je *MeganDB*. Tato databáze je speciálně navržena pro spolupráci s analytickým nástrojem MEGAN (MEta Genome ANalyzer), který mimo jiné umožňuje srovnání vlastních souborů s veřejnými databázemi. Databáze poskytuje RMA soubory pro nástroj MEGAN a také surové sekvence ve formátu FASTA. [19]

EBI Metagenomics je portál pro analýzu, archivaci a vyhledávání metagenomických a metatranskriptomických dat. Tato databáze umožňuje funkční a taxonomickou analýzu dat. Služba je navržena tak, aby zajistila formátování dat a metadat v souladu se směrnicemi a standardy *European Nucleotide Archive* a *Genomic Standards Consortium* (GSC). EBI Metagenomics umožňuje identifikaci rRNA sekvence pomocí nástroje rRNASelector a provádí taxonomickou analýzu na základě 16S rRNA s využitím nástroje Qiime. Pro funkční analýzu předem navržených protein kódujících úseků je využíván nástroj InterPro. [20]

2 VÝPOČETNÍ METODY PRO ZPRACOVÁNÍ METAGENOMU

Účelem bioinformatických nástrojů spojených se zpracováním metagenomu je kontrola kvality dat, statistická analýza a vizualizace dat. Tyto nástroje a výpočetní metody mají své uplatnění zejména ve srovnávací metagenomice a pro funkční a fylogenetickou analýzu.

S vývojem sekvenačních metod nové generace dochází také k vývoji stále pokročilejších programů pro zpracování sekvenačních dat. Metody sekvenování nové generace produkují velké množství kratších úseků DNA (100-800 bp pro 454 pyrosekvenování, 35 bp pro SOLiD, 35-100 bp pro Illumina, 30-35 pro Helicos), což přestavuje problém pro počítačové zpracování. Náročnost skládání je dána právě délkou a počtem fragmentů. Vznikají tedy nové, speciální bioinformatické nástroje pro práci s krátkými sekvenčními fragmenty a pro práci s velkým množstvím dat. [21; 22; 23]

Existuje mnoho přístupů pro analýzu metagenomických dat. V této kapitole budou prezentovány vybrané bioinformatické nástroje a algoritmy rozlišené na základě charakteru analyzovaných dat – shotgun a amplikon.

2.1 Amplikon přístup

Jak již bylo zmíněno, pro tento přístup je charakteristické využití taxonomicky specifických ribozomálních RNA genů, např. 16S rRNA. Analýzu, srovnání mikrobiálních komunit a vizualizaci získaných dat umožňují nástroje jako například QIIME [24], mothur [25], MEGAN [19] a MG-rast [26].

2.1.1 Shlukování OTU

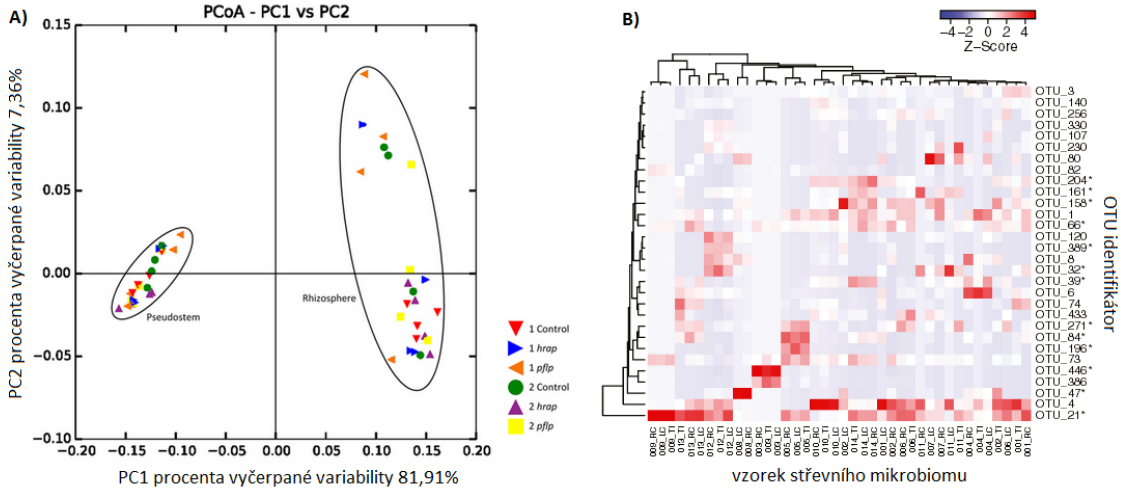
Základním procesem při analýze mikrobiálního společenství je shlukování sekvencí do operačních taxonomických jednotek (angl. *operational taxonomic units*, OTUs). [21] OTU zde představuje určitou taxonomickou úroveň (například druh nebo rod) na základě nastavení prahové hodnoty pro podobnost sekvencí. Avšak tento přístup, nazývaný též jako *de novo* shlukování OTU, je nepřesným a neúčinným řešením pro identifikaci druhů. Alternativním přístupem je shlukování OTU se známou referencí. Jedním z nejrozšířenějších algoritmů pro klasifikaci sekvencí je BLAST [27], který umožňuje srovnání s referenční databází, jako je například databáze SSU podjednotek (RDP [28]).

2.1.2 UniFrac

Efektivní vzdálenostní metrikou pro srovnání mikrobiálních komunit s využitím fylogenetické informace je UniFrac [29]. Jedná se o měřítko β-diverzity. Vzdálenost je počítána mezi páry vzorků, přičemž každý vzorek reprezentuje mikrobiální komunitu. Ze všech taxonů je vytvořen fylogenetický strom. Větev, která vede k oběma taxonům je označena jako *sdílená* a větev, která vede pouze k jednomu ze vzorků je označena jako *nesdílená*. Vzdálenost mezi dvěma vzorky se pak vypočítá jako poměr součtu nesdílených délek větví a součtu všech délek větví stromu, jedná se tedy o podíl celkové délky nesdílených větví. [29]

2.1.3 Vizualizace dat

Fylogenetické vztahy mezi sekvencemi DNA mohou být odvozeny s využitím existující referenční databáze s přidruženou fylogenetickou informací nebo odvozením fylogenetického stromu *de novo*. Následnou analýzou, např. analýzou hlavních koordinát (*principal coordinates analysis*, PCoA) lze vizualizovat vztah mezi mikrobiálními komunitami a odhalit mikrobiální diverzitu. Další možností je vizualizace s využitím tzv. teplotních map (angl. *heat maps*). [21; 22] Oba tyto přístupy jsou zobrazeny na Obr. 2.1.



Obr. 2.1: Možnosti vizualizace dat. a) PCoA biplot založený na vážených UniFrac vzdálenostech pro vizualizaci 16S rRNA genů gammaproteobakterií [30], b) teplotní mapa OTU jednotek (osa y), na ose x je prezentováno 30 vzorků střevního mikrobiomu člověka s nejvyšší četností [31].

2.2 Shotgun přístup

S rozvojem NGS technologií a snižování nákladů na sekvenaci vzrostl zájem výzkumníků o shotgun metagenomické studie. Od první aplikace shotgun metagenomiky (v roce 2006) došlo k velkému pokroku bioinformatických nástrojů a rozvoji databází dříve využívaných v genomice a přijetí metod používaných v ekologii. [21; 22] V této části budou popsány vybrané metody pro zpracování shotgun metagenomických dat.

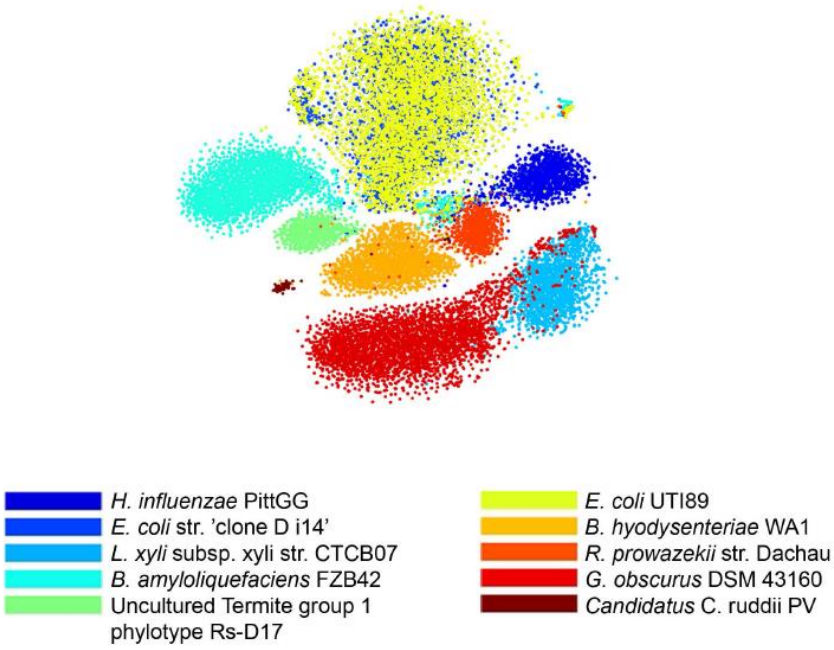
2.2.1 Taxonomická analýza

Pro odvození taxonomického původu metagenomických čtení se používají dva druhy metod, metody založené na podobnosti a metody založené na složení.

Metody založené na podobnosti porovnávají sekvence s referenčním setem genomických sekvencí. Jedním z nejvíce využívaných nástrojů je MEGAN, který používá algoritmus BLAST pro porovnávání metagenomických čtení s databází anotovaných sekvencí podle NCBI taxonomie. [19] Dalšími široce rozšířenými nástroji jsou MG-RAST [26] a WebCARMA [32].

Metody založené na složení obvykle zjistí specifické vlastnosti, jako například obsah CG nebo taxonomicky specifických četností oligonukleotidových slov (blíže v kapitole 3). [33] Pro klasifikaci metagenomických sekvencí mohou být využité různé techniky. Z hlediska přístupů strojového učení lze tyto techniky dále dělit na supervizované a nesupervizované. Supervizované metody učení používají během trénovací fáze referenční sadu charakteristik sekvencí pro každou z taxonomických tříd. Klasifikátor naučený na tomto souboru je poté používán k identifikaci fragmentů neznámého původu do taxonomických tříd. Příkladem je Bayesovský klasifikátor [34] a nástroj PhyloPythia [33]. *Nesupervizované učící techniky* nevyžadují pro klasifikaci referenční datové sady. Klasifikátor je vytrénován na základě jednotlivých charakteristik analyzovaného datasetu. Jedná se například o nástroj TETRA [35] založený na výpočtu korelačních koeficientů mezi tetranukleotidovými vzory a přístupy využívající Kohonenovy samoorganizující mapy (angl. *self-organizing maps*, SOMs) [36].

Jedním ze současných nástrojů pro vizualizaci metagenomických dat nezávislých na referenci je VizBin [37]. Tento přístup, popsaný v publikaci *Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction* [38], je založen na centrované logratio (clr) transformaci oligonukleotidových podpisů genomických fragmentů a následné nelineární redukci dimenzionality Barnes-Hutovým algoritmem (Barnes-Hut stochastic neighbor embedding, BH-SNE). Na Obr. 2.2 je uveden příklad Barnes-Hut SNE vizualizace clr-transformovaných tetranukleotidových podpisů.



Obr. 2.2: Vizualizace clr-transformovaných tetranukleotidových podpisů s využitím BH-SNE algoritmu [38]

2.2.2 Funkční analýza

Funkční charakteristika komunit a rekonstrukce metabolických drah je obvykle cílovým krokem shotgun metagenomiky. V závislosti na sekvenační strategii a pokrytí mohou být sekvence využívány přímo jako krátké fragmenty nebo mohou být sestavovány do větších celků, tzv. *kontigů*. Pro analýzu se využívají výkonné nástroje, které jsou obvykle dostupné jako webové aplikace (kdy se metagenomická data zpracovávají na webovém serveru), jako např. již zmíněné nástroje MG-RAST, webCARMA, dále CAMERA [39] nebo IMG/M [40]. Některé z nástrojů jsou dostupné také ke stažení na PC, jako MEGAN4 [41]. Tyto nástroje zahrnují platformy pro predikci genů, zařazení do funkčních kategorií (databáze COGs [9]) a eggNOG [42]), určení genové ontologie (The gene ontology database (GO) [43]) a metabolických drah (databáze KEGG [8]).

Jedním z hlavních výhod online zdrojů, jako MG-RAST, je možnost srovnání také s veřejně přístupnými metagenomickými daty. [21]

3 TAXONOMICKY SPECIFICKÉ ČETNOSTI NUKLEOTIDOVÝCH SLOV

Přístup klasifikace organismů na základě analýzy nukleotidových slov je, jak již bylo zmíněno v předchozí kapitole, jednou z nesupervizovaných metod taxonomické analýzy založené na studiu složení sekvencí. Dle rešerše literatury se jeví jako vhodný a účinný nástroj pro předzpracování velkých metagenomických datasetů a jejich rozdělení do shluků odpovídajících jednotlivým taxonům obsaženým v metagenomickém vzorku.

Řada studií prokázala přítomnost taxonomicky specifických nukleotidových podpisů (označovaných také jako nukleotidová slova, *k*-mery, či *k*-tice) v genomických sekvencích. [44; 45; 46] Takové příznaky mohou být reprezentovány jako vektory ve vysokodimenzionálním Euklidovském prostoru. Pro umožnění interpretace člověkem je nezbytná redukce dimenzí, obvykle na dvourozměrná data. V ideálním případě by měla taková metoda zachovávat taxonomickou strukturu dat. Karlin a kol. [47] dále prokázali, že frekvence *k*-merů jsou podobné v rámci různých oblastí jednoho genomu, ale liší se mezi genomy různých druhů. Nesou tedy fylogenetickou informaci. Z tohoto poznatku tedy vychází předpoklady, že je tato metoda vhodná pro analýzu metagenomických datasetů, které obsahují metagenomická čtení pocházející z různých částí mnoha rozdílných organismů.

Metody využívající nukleotidové podpisy, prozatím nebyly široce užívány pro kvantitativní analýzu metagenomických sekvencí. S rozvojem sekvenovacích technologií nové generace a tím i jejich dostupnosti však narostl potenciál využití těchto metod. Výhodou tohoto přístupu je, že není vyžadována znalost kompletního genomu nebo genů. Velké množství krátkých metagenomických čtení také není třeba zarovnávat (odtud také označení „*alignment-free*“ metody v anglické literatuře), sestavovat do kontigů, či porovnávat s referenčními databázemi, které často bývají nekompletní. [48]

V této kapitole budou představeny vybrané numerické reprezentace metagenomických čtení založených na analýze nukleotidových slov.

3.1 Četnosti nukleotidových slov

K-mer je slovo skládající se z *k* znaků abecedy nukleotidů A = {A, C, G, T}.

Množina W_k se skládá ze všech možných *k*-merů, tedy variací *k*-té třídy ze 4 prvků (znaků abecedy nukleotidů A) s opakováním (3.1). Tato množina má *n* elementů a jednotlivá slova jsou abecedně seřazena. [49]

$$W_k = \{w_{k,1}, w_{k,2}, \dots, w_{k,n}\}$$
$$n = 4^k \tag{3.1}$$

Standardním přístupem pro výpočet *k*-merů v sekvenci o délce *m* je využití plovoucího okna o délce *k*, které se posouvá vždy o jeden nukleotid od prvního nukleotidu po *m* - *k* + 1. V této metodě je povoleno překrývání jednotlivých *k*-merů. Sekvence může být reprezentována *n*-dimenzionálním vektorem c_k tvořeným četnostmi jednotlivých k-merů (3.2) [49]:

$$c_k = \left(c(w_{k,1}), c(w_{k,2}), \dots, c(w_{k,n})\right), \tag{3.2}$$

kde *n* je celkový počet všech možných *k*-merů. Vzhledem k tomu, že analyzované sekvence mohou být různě dlouhé, je vhodné provést korekci četností *k*-merů na počet slov v sekvenci (dělitel $m - k + 1$). **Vektor frekvencí *k*-merů** f_k je tedy získán jako relativní četnost každého *k*-meru dle vzorce (3.3) [49]

$$f_k = \left(f(w_{k,1}), f(w_{k,2}), \dots, f(w_{k,n})\right) \tag{3.3}$$
$$= \left(\frac{c(w_{k,1})}{m-k+1}, \frac{c(w_{k,2})}{m-k+1}, \dots, \frac{c(w_{k,n})}{m-k+1}\right).$$

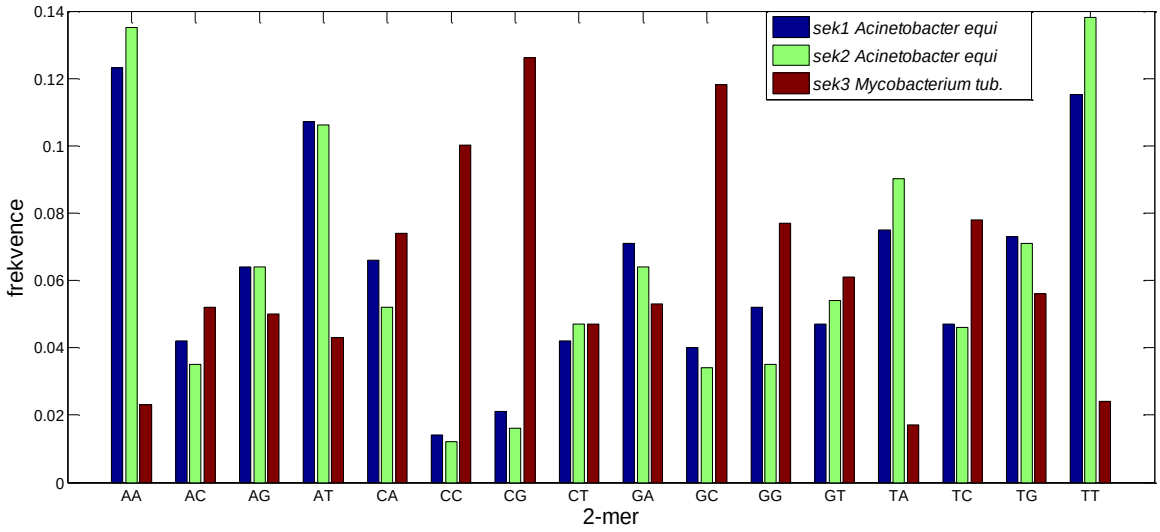
Například pro sekvenci *X* = *AAACACAG* o délce *m* = 8 by byl výpočet vektoru frekvencí dinukleotidů (*k*=2, též dimery) proveden:

$$W_2 = \{AA, AC, AG, AT, CA, CC, \dots\}$$
$$c_2^X = (2, 2, 1, 0, 2, 0, \dots)$$
$$f_2^X = \left(\frac{2}{7}, \frac{2}{7}, \frac{1}{7}, 0, \frac{2}{7}, 0, \dots\right).$$

Vektory c_2^X a f_2^X mají délku $n = 4^2 = 16$.

V následujícím příkladu (Obr. 3.1) je představena analýza taxonomicky specifických *k*-merů pro *k*=2. Zobrazeny jsou frekvence dimerů pro segmenty bakteriálních genomů o délce 1000 bp. Sekvence sek1 a sek2 reprezentují taxon *Acinetobacter equi*

a sek3 odpovídá taxonu *Mycobacterium tuberculosis*. Již z frekvencí dimerů je patrné, že mezi frekvenčními vektory sekvencí majících původ ze stejného bakteriálního genomu (sek1 a sek2) je relativně velká podobnost. Naproti tomu frekvence dimerů pro sek3, reprezentující odlišný taxon, se v několika oblastech výrazně liší, zde zejména ve frekvencích dimerů AA, AT, CC, CG, GC, TA a TT.



Obr. 3.1: Frekvence dinukleotidů pro 3 genomické fragmenty bakteriálních genomů

3.2 Reprezentace pozicemi nukleotidových slov (Yang)

Ve studii založené na výpočtu četnosti *k*-merů Yanga a kol. [50] je využívána modifikovaná metoda. Vzhledem k tomu, že analyzované sekvence mohou být různě dlouhé, uvažuje pouze četnosti *k*-merů.

Nejprve jsou četnosti *k*-merů seřazeny vzestupně (3.4)

$$S_k = \left(c\left(w_{k,1}\right), c\left(w_{k,2}\right), \ldots, c\left(w_{k,n}\right)\right). \tag{3.4}$$

Pokud si jsou četnosti *k*-merů rovny, jsou seřazeny abecedně dle názvu *k*-meru.

Vektor reprezentující sekvenci O_k je pak tvořen **pozicemi seřazených hodnot četností *k*-merů** (3.5):

$$O_k = \left(o\left(w_{k,1}\right), o\left(w_{k,2}\right), \ldots, o\left(w_{k,n}\right)\right). \tag{3.5}$$

V této studii je pro výpočet vzdálenosti dvou sekvencí využíváno Euklidovské vzdálenosti pořadí nukleotidových slov. Při výpočtu kombinace (konsensu) pro slova různé délky *k*, je vhodné tuto vzdálenost normalizovat celkovým počtem slov, tedy podělit hodnotou 4^k.

3.3 Reprezentace intervalovými vzdálenostmi *k*-merů (Ding)

Ding a kol. [51] navrhli metodu reprezentace sekvencí označovanou jako **normalizovaná průměrná intervalová vzdálenost *k*-merů**.

Tato metoda, stejně jako předchozí, povoluje překrývání *k*-merů v sekvenci. Mějme množinu všech možných *k*-merů W_k = { w_1, $w_{2, \ldots,}$ w_n }. Pozice *k*-merů jsou ukládány v pozičních vektorech $P_{w1,w2,\ldots,wn}$. Četnosti *k*-merů c_k jsou tedy rovny délce korespondujících pozičních vektorů. Jedná se však o redundantní vyjádření. Normalizovanou průměrnou intervalovou vzdálenost *k*-merů lze z tohoto vyjádření získat následovně (4.2) [51]:

$$E(w_1, w_2, \ldots, w_n) = \begin{cases} (P_{w_1,w_2,\ldots w_n}(c_k) - P_{w_1,w_2,\ldots w_n}(1))/c_k m, & c_k \neq 0, \\ 0, & c_k = 0, \end{cases} \quad (3.6)$$

kde c_k je četnost *k*-meru a *m* délka sekvence.

$E(w_1, w_2, \ldots, w_n)$ tedy nezáleží pouze na četnostech *k*-merů, ale také na délce sekvence a první a poslední pozici *k*-meru. Dle autorů kombinace poziční informace a četnosti *k*-merů umožňuje získat ze sekvence více informace. To si demonstrujme na následujícím příkladu. Mějme dvě sekvence *sek1* a *sek2* o délce *m*=8. Zatímco četnost nukleotidového slova *AA* je pro obě sekvence rovna 3, normalizovaná průměrná intervalová vzdálenost *E(AA)* se liší a proto tato reprezentace obsahuje více potenciální fylogenetické informace.

sek1 = AACGGAAA c_{AA}=3 *E*(*AA*) = (7-1)/(3*8) = 0,25

sek2 = CCGGAAAA c_{AA}=3 *E*(*AA*) = (7-5)/(3*8) = 0,083

3.4 Reprezentace relativními vzdálenostmi *k*-merů (Tang)

Kombinace informace o pozici a četnosti *k*-merů byla aplikována také ve studii Tang a kol. [52]

Normalizovaná průměrná relativní vzdálenosti *k*-merů může být vyjádřena opět s využitím informace uložené v pozičních vektorech následovně (4.3) [52]:

$$D(w_1, w_2, \ldots, w_n) = \begin{cases} \dfrac{\sum_{i=1}^{c_k}(P_{w_1,w_2,\ldots w_n}(i) - P_{w_1,w_2,\ldots w_n}(1))}{c_k(m-k+1)}, & c_k \neq 0, \\ 0, & c_k = 0, \end{cases} \quad (3.7)$$

S využitím sekvencí *sek1* a *sek2* z předchozího příkladu si opět výpočet znázorněme:

sek1 = AACGGAAA c_{AA}=3 $D(AA)$ = (7-1)+(6-1)+(1-1)/3*(8-2+1) = 0,52

sek2 = CCGGAAAA c_{AA}=3 $D(AA)$ = (7-5)+(6-5)+(5-5)/ 3*(8-2+1) = 0,143

3.5 Reprezentace symetrizovanými *k*-mery (Gori)

Reprezentace sekvencí tzv. **symetrizovanými *k*-mery** může být výhodná, vzhledem ke skutečnosti, že metagenomická data mohou být sekvenována z obou vláken DNA řetězce. Bere tedy v úvahu četnosti *k*-merů f_i a sečte je s jejich reverzními komplementy f_i^C. V případě výskytu tzv. palindromatického *k*-meru se sčítání neprovádí. Definujme tedy vektor reprezentující sekvenci jako $\rho^S := (a_{1, \dots,} a_n)$, přičemž $a_i = f_i + f_i^C$ pokud $w_i \neq w_i^C$, jinak $a_i = f_i$. Vhodné je také provést korekci četností symetrizovaných k-merů na počet slov (dělitel $m - k + 1$) pro získání vektoru frekvencí symetrizovaných *k*-merů. [53]

Výhodou této metody je, že jednotlivé sekvence lze reprezentovat vektorem nukleotidových podpisů o nižším počtu dimenzí. Délka vektoru reprezentující sekvenci je pak např. pro délku slova *k*=2 rovna 10, vzhledem k tomu, že ze 16 možných slov existují právě 4 palindromatické *k*-mery a zbývajících 12 je symetrizovaných. V případě *k*=5 pak získáváme pro symetrizované *k*-mery vektor o 512 dimenzích, dojde tedy k redukci délky vektoru na polovinu oproti předchozím reprezentacím (celkový počet slov pro *k*=5 je 1024).

Analýza symetrizovaných *k*-merů byla využívána v řadě současných metagenomických studií. [38; 54; 53] Dle Lazny *a kol.* [38] je vhodné tyto symetrické *k*-mery transformovat centrovanou logratio transformací, která se provede dle následujícího vzorce (4.4):

$$\rho_{clr} = clr(\rho^S) = \left(\ln\frac{a_1}{g_\rho}, \dots, \ln\frac{a_n}{g_\rho}\right), \tag{3.8}$$

kde g_ρ je geometrický průměr vektoru ρ^S.

3.6 Další metody

Další modifikovaná metoda byla prezentována ve studii Qi a kol. [55]. Tato metoda založená na výpočtu dinukleotidů bere v úvahu jak standardní frekvenční četnosti nukleotidových slov, tedy přilehlých nukleotidů, tak i nukleotidů nesousedících. Na

příklad pro sekvenci X=*ATCGATC* jsou přilehlé dinukleotidy AT, TC, CG, GA s frekvenční četností $\frac{2}{6}, \frac{2}{6}, \frac{1}{6}, \frac{1}{6}$ a nesousedící dinukleotidy (tedy takové dinukleotidy, které jsou oddělené jedním nukleotidem) AC, TG, CA, GT, AC mají frekvenční četnosti $\frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}$.

Podle Gori a kol. [53] jsou vhodné také různé modifikace symetrizovaných vektorů, například takové, kdy při slučování součtu *k*-merů a jejich reverzních komplementů se počítá pouze s minimální, či naopak maximální frekvencí výskytu. Ve studii popisují také reprezentace s využitím pouze palindromatických *k*-merů pro *k*=4.

Námětem pro studium je zejména zhodnocení vlivu délky *k*-merů pro různé metody a také studium jejich vhodně zvolených kombinací.

4 SHLUKOVÁ ANALÝZA

Shluková analýza je vícerozměrná statistická metoda, která se používá ke klasifikaci objektů. Slouží ke třídění jednotek do jednotlivých shluků tak, aby si objekty patřící do stejné skupiny byly podobnější než objekty z ostatních skupin. Shlukovou analýzou lze snížit počet dimenzí objektů takovým způsobem, že řadu uvažovaných proměnných zastoupí proměnná vyjadřující příslušnost objektu k definované skupině. Cílem shlukování je zejména popsat strukturu dat a nalézt shluky podobných objektů. [56]

Shlukování lze rozlišovat na hierarchické a nehierarchické. *Hierarchická shluková analýza* vytváří systém shluků a podshluků, přičemž každý shluk může obsahovat podshluky nižšího řádu a sám může být součástí shluku vyššího řádu. Výsledek lze graficky znázornit tzv. dendrogramem. *Nehierarchická shluková analýza* rozdělí objekty do několika shluků stejného řádu. Příkladem je algoritmus *K-means*. [56]

Vzhledem k tomu, že při analýze nukleotidových slov mohou vznikat vysokodimenzionální vektory reprezentující sekvence, je vhodné se také zaměřit na metody pro redukci dimenzionality. Z těchto metod bude v této kapitole diskutována *analýza hlavních komponent* a algoritmus *t-distributed stochastic neighbor embedding*.

Na závěr kapitoly bude objasněna metodika hodnocení úspěšnosti klasifikace.

4.1 Vzdálenostní metriky

Informaci o podobnosti dvou sekvencí DNA lze zjistit na základě výpočtu vzdálenosti vektorů frekvencí *k*-merů. Euklidovská vzdálenost je jednou z nejběžnějších vzdálenostních metrik ve studiích zabývajících se analýzou *k*-merů. [50; 51] Uplatnění má také Manhattanská vzdálenost (*City block*) a kosinová vzdálenost. [55]

Euklidovskou vzdálenost d_k^E mezi sekvencí X a Y lze vypočítat dle vzorce (4.1) [57]:

$$d_k^E\,(X,Y) = \sqrt{\sum_{i=1}^{n} (f_{k,i}^X - f_{k,i}^Y)^2} \tag{4.1}$$

Manhattanskou vzdálenost d_k^{Man} mezi dvěma frekvenčními vektory f^X a f^Y lze vyjádřit jako (4.2) [57]:

$$d_k^{Man}\,(X,Y) = \sum_{i=1}^{n} \left| f_{k,i}^X - f_{k,i}^Y \right|. \tag{4.2}$$

Podle *kosinové vzdálenosti* d_k^{cos} je vzdálenost dvou vektorů úhel, který svírají. Konkrétně lze tuto metriku vyjádřit dle vztahu (4.3) [57]:

$$d_k^{cos}(X,Y) = 1 - \left(\frac{\sum_{i=1}^{n}\left(f_{k,i}^{X} - f_{k,i}^{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left((f_{k,i}^{X})^2\right) * \left((f_{k,i}^{Y})^2\right)}}\right). \tag{4.3}$$

4.2 Korelační koeficienty

Vedle vzdálenostních metrik lze vzájemný vztah dvou sekvencí vyjádřit s využitím korelací. [49]

Pearsonův korelační koeficient r měří statistickou závislost dvou vektorů. Korelační koeficient je velmi ovlivněn odlehlými hodnota. Podmínkou je dodržení dvourozměrného normálního rozdělení. Počítá se pomocí směrodatných odchylek obou proměnných a jejich kovariance (míra vzájemné vazby mezi veličinami). Výpočet lze vyjádřit vzorcem (4.4) [58]:

$$r(X,Y) = \frac{\sum_{i=1}^{n}\left(f_{k,i}^{X} - \overline{f_{k,\iota}^{X}}\right)\left(f_{k,i}^{Y} - \overline{f_{k,\iota}^{Y}}\right)}{\sqrt{\sum_{i=1}^{n}\left(f_{k,i}^{X} - \overline{f_{k,\iota}^{X}}\right)^2 \sum_{i=1}^{n}\left(f_{k,i}^{Y} - \overline{f_{k,\iota}^{Y}}\right)^2}}. \tag{4.4}$$

Spearmanův korelační koeficient r_{sp} , označovaný též jako Spearmanův koeficient pořadové korelace, zachycuje statistickou závislost mezi dvěma veličinami. Pro malé rozsahy dat je jeho výpočet méně náročný než u Pearsonova korelačního koeficientu. Mějme dva frekvenční vektory f_k^X a f_k^Y. Uspořádáme *n* hodnot těchto vektorů podle velikosti a přiřadíme jim pořadová čísla $p_{k,i}^X$ a $p_{k,i}^Y$. Hodnota koeficientu je pak rovna (4.5) [58]:

$$r_{sp}(X,Y) = 1 - \frac{6\sum_{i=1}^{n}\left(p_{k,i}^{X} - p_{k,i}^{Y}\right)^2}{n(n^2-1)}. \tag{4.5}$$

4.3 Hierarchické shlukování

V analýze *k*-merů má zastoupení zejména hierarchická shluková analýza, která se opírá o shlukování do tzv. operačních taxonomických jednotek (OTU) [48; 50; 51]. S využitím dendrogramu lze tak vyobrazit vzájemný vztah analyzovaných sekvencí.

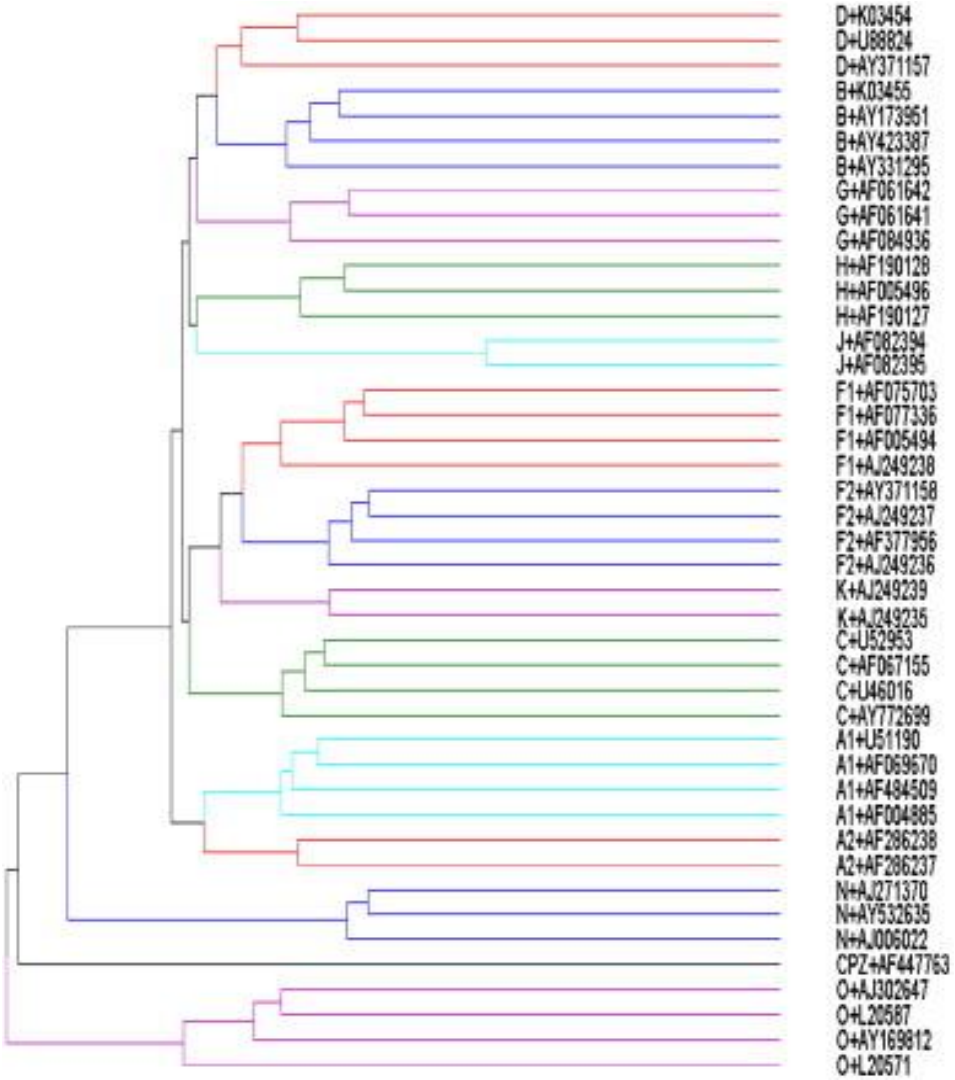
Obecný posup hierarchického shlukování je následující [56]:

1. Základním krokem shlukování je výpočet podobností mezi všemi dvojicemi objektů, tedy vytvoření tzv. **asociační matice**. Existuje mnoho způsobů výpočtu podobnosti objektů, popsané v části 4.1, nicméně jednou z nejrozšířenějších je Euklidovská vzdálenost. Dále mohou být použity metriky vycházející z korelačních koeficientů popsaných v části 4.2.
2. Dalším krokem je shlukování (párování) objektů. Algoritmy pro shlukování využívají informace o vzdálenosti/podobnosti k určení blízkosti objektů. Dva nejpodobnější objekty jsou spojovány do shluku. Poté se přepočítá asociační matice tak, že spojené objekty dále vystupují jako jediný objekt. Takový iterační proces končí spojením všech objektů do jediného shluku. Existují různé shlukovací algoritmy pro výpočet vzdálenosti/podobnosti spojených objektů vůči ostatním objektům:
    - **Metoda nejbližšího souseda** (*single linkage*) – vzdálenost je určena jako nejmenší vzdálenost mezi objekty shluku,
    - **Metoda nejvzdálenějšího souseda** (*complete linkage*) – vzdálenost je určena jako největší vzdálenost mezi objekty shluku,
    - **Centroidová metoda** – používá euklidovskou vzdálenost mezi centroidy dvou shluků,
    - **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); a další.
3. Na závěr je určen počet výstupních větví. Tím jsou data rozdělena do jednotlivých shluků.

Tzv. *míru věrohodnosti*, tedy kritérium pro volbu „nejlepšího dendrogramu“ je *kofenetický korelační koeficient c.* Jedná se o Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností založenou na dendrogramu. Pokud je hodnota *c* menší než přibližně 0,8, všechny objekty patří do jediného shluku. Obecně platí, že čím vyšší je kofenetický korelační koeficient, tím nižší je ztráta informací, vznikající v procesu slučování objektů do shluků. [57]

Příkladem hierarchického shlukování je metoda realizovaná v publikaci *A novel statistical measure for sequence comparison on the basis of k-word counts* [50]. V této studii je vektor reprezentující sekvenci tvořen pozicemi seřazených hodnot četností *k*-merů, jakož bylo popsáno v kapitole 3.2. Pro výpočet vzdálenosti mezi objekty je

využívána euklidovská metrika a metoda UPGMA jako shlukovací algoritmus. Prezentována je analýza pro *k*-mery, kde $k = \langle 2;7 \rangle$. Studie ukazuje lepší výsledky pro k=5 a vyšší. Na Obr. 4.1 je zobrazena analýza 43 subtypů HIV-1 sekvencí s využitím výpočtu pro k=5. Z tohoto dendrogramu je patrné, že výsledky jsou poměrně konzistentní. Všechny subtypy jsou jasně seskupeny a odpovídají zažité taxonomii. Dendrogram také znázorňuje vzájemný vztah jednotlivých subtypů.



Obr. 4.1: Dendrogram znázorňující vzájemný vztah subtypů HIV-1 sekvencí [50]

4.4 Nehierarchické (K-means) shlukování

Nejběžnější metodou nehierarchického shlukování je algoritmus **K-means,** označován také českým ekvivalentem jako metoda k-průměrů. Metoda zařazuje objekty do shluků na principu ANOVA, je tedy analogií Wardovy metody v hierarchickém shlukování. [59]

Počet shluků je nezbytné definovat předem, přičemž výběr nejvhodnějšího počtu shluků se provádí buď expertně nebo pomocí matematických metod výběru optimálního počtu shluků. Obdobně jako je tomu u hierarchického shlukování, prvním krokem je výběr vhodné metriky vzdáleností či podobností všech objektů pro vytvoření asociační matice. Následně jsou objekty do shluků řazeny tak, aby byla suma čtverců vzdáleností objektů k centroidům jejich shluků minimální. [59]

Kroky tohoto iteračního algoritmu jsou následující [56]:

1. Podle zvoleného počtu shluků *k* náhodně vybere středy shluků (centroidy).
2. Vypočítá vzdálenosti všech objektů ke každému z centroidů.
3. Přiřadí každý objekt do shluku s nejbližším centroidem.
4. Nové polohy centroidů se vypočítají jako průměr všech objektů v každém ze shluků.
5. Opakuje kroky 2 až 4 dokud nedochází ke změnám přiřazení objektů do shluků.

4.5 Redukce dimenzionality

Sekvenční příznaky založené na výpočtu četností nukleotidových slov lze popsat jejich pozicí ve vícerozměrném prostoru. V našem případě se může jednat až o tisíce dimenzí (pro analyzovanou délku nukleotidových slov větší než 5). Více než 3D prostor je pro člověka vizuálně neuchopitelný a hledání vztahů ve více než 3 dimenzích je problematické.

Z přístupů redukce dimenzionality vektorů reprezentujících sekvence bude představena analýza hlavních komponent a algoritmus *t-distributed stochastic neighbor embedding*, který analýzu hlavních komponent rovněž využívá jako předzpracování vstupních dat.

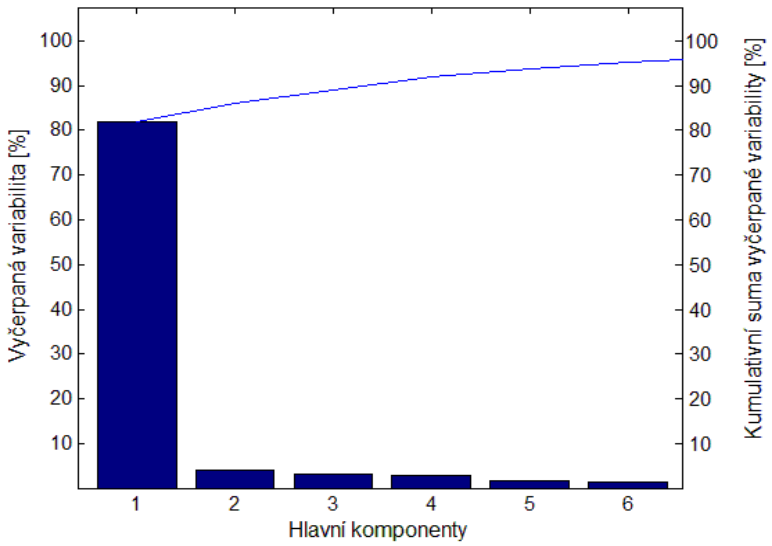
4.5.1 Analýza hlavních komponent

Označuje se jako PCA, z anglického *Principal Component Analysis*. Jedná se o statistickou metodu pro redukci dimenzionality. Pomáhá nalézt v *n*-dimenzionálním prostoru nejvhodnější pohled na data poskytující maximum informací o analyzovaných

objektech. Mezi dimenzemi existují obvykle korelační vztahy, jednotlivé dimenze se tedy navzájem vysvětlují a pro popis kompletní informace v datech není za potřebí využívat všech redundantních proměnných. [60]

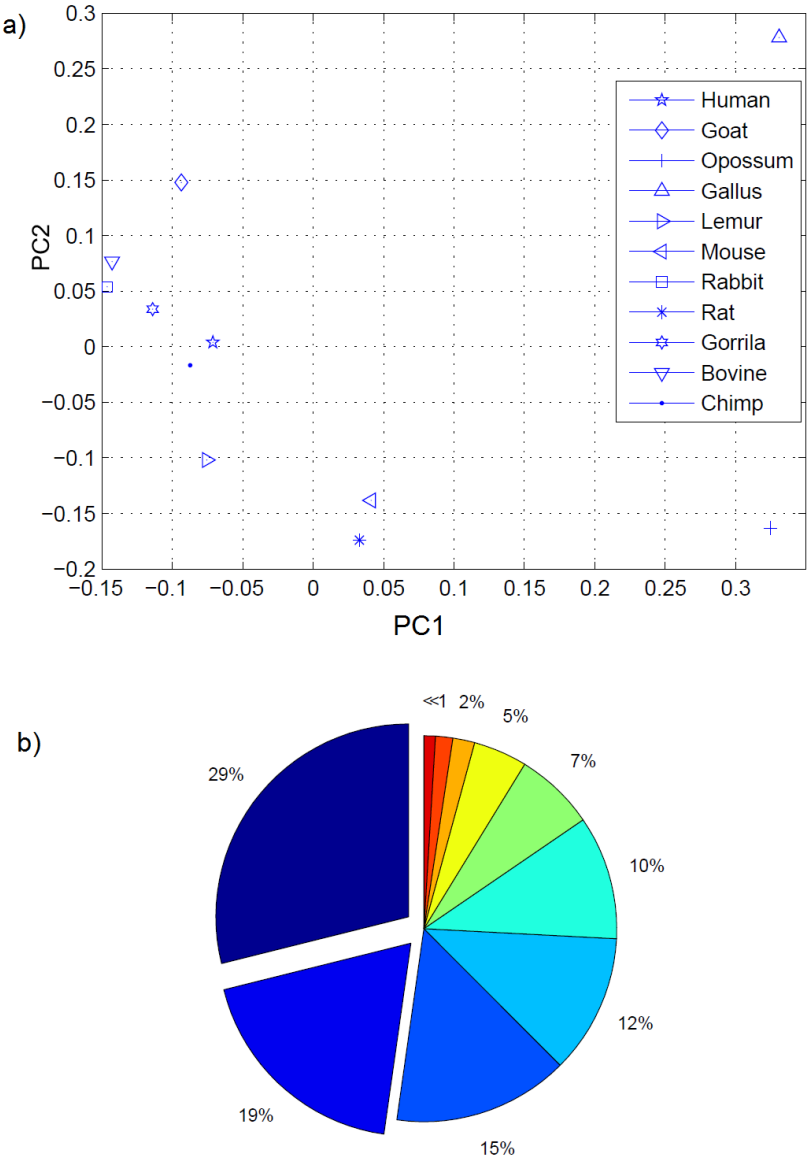
Taková vícerozměrná analýza výpočetně vychází z asociací proměnných popisujících objekty a snaží se na základě jejich korelací nebo kovariancí stanovit dimenze zahrnující větší podíl variability než připadá na původní proměnné. V našem případu je vhodné využití kovarianční matice, vzhledem k analýze proměnných o stejných jednotkách a podobném významu (frekvence a četnosti *k*-merů). Vlastní realizace je pak provedena výpočtem vlastních čísel (*eigenvalues*) a vlastních vektorů (*eigenvectors*) kovarianční asociační matice. Vlastní čísla matice jsou spjaty s variabilitou vyčerpanou vytvářenými faktorovými osami. Vlastní vektory pak definují směr nových faktorových os v prostoru původních proměnných. Výstupní proměnné, označované jako hlavní komponenty (PC), jsou vypočítány jako součin původních proměnných s příslušnými vlastními vektory. [60]

Cílem takové analýzy je, jak již bylo řečeno, výběr menšího počtu dimenzí pro další analýzu. Existuje řada pravidel pro výběr optimálního počtu dimenzí. Jednou z nich je tzv. *scree plot*, označovaný v literatuře také jako graf úpatí [61]. Jedná se o sloupcový diagram procent variability vyčerpané jednotlivými hlavními komponenty. Lze ho použít jako grafický nástroj hledající zlom ve vztahu počtu os (hlavních komponent) a vyčerpané variability, viz Obr. 4.2. Obvyklé je pracovat s proměnnými, které vyčerpávají 90 % variability, přičemž zbývajících 10 % je považováno za šum, či neužitečné složky. Tento přístup se však liší dle aplikace. Ideální pro výstup PCA je co nejmenší počet výstupních hlavních komponent za současné největší vyčerpané variability. [60]



Obr. 4.2: Scree plot, ukázka

Metoda PCA byla použita například ve studii výzkumníků Qi a kol. pro redukci vektoru sekvenčních příznaků o 336 dimenzích. [55] Na Obr. 4.3 je zobrazeno 11 vektorů ve 2D prostoru reprezentovaném dvěma hlavními komponentami s největší vyčerpanou variabilitou.



Obr. 4.3: a) PCA projekce 336 dimenzionálního vektoru 11 druhů ve 2D prostoru, b) podíl hlavních komponent na vyčerpané variabilitě (převzato z [55])

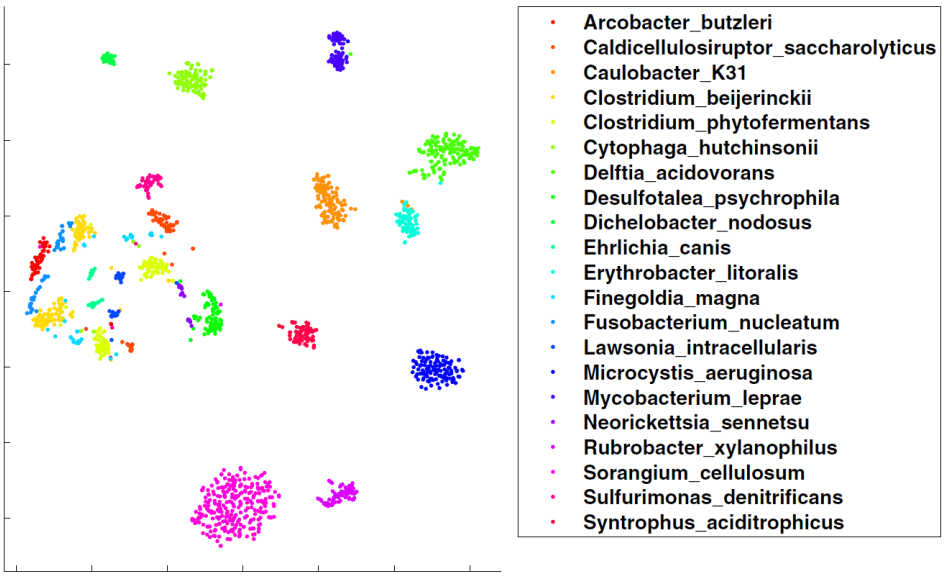
4.5.2 *T-distributed Stochastic Neighbor Embedding*

Algoritmus *t-distributed stochastic neighbor embedding* (t-SNE) je metodou strojového učení pro redukci dimenzionality vyvinutý autory van der Maaten a Hinton. Jedná se o nelineární techniku redukce dimenzí, která je vhodná pro zpracování vysokodimenzionálních dat v řadě aplikací. Algoritmus umožňuje vizualizaci podobností objektů vyjádřenou dvěmi, případně třemi dimenzemi. [62]

Tato část má za účel představit tento současný algoritmus. Vzhledem k tomu, že t-SNE bude v této práci použit pouze k účelům srovnání, nebude zde uveden detailní matematický popis. Kód k tomuto algoritmu a jeho modifikacím je dostupný na webových stránkách autora [63].

Vhodným předzpracováním dat pro tento algoritmus je redukce dimenzionality s využitím PCA. Samotný algoritmus se skládá ze dvou hlavních částí. Nejprve je vytvořeno Studentovo t-rozdělení pravděpodobností mezi dvojicemi vysokodimenzionálních objektů tak, aby odráželo vzájemnou podobnost objektů. Dále t-SNE vytvoří obdobné rozdělení pravděpodobnosti nad body v nízkodimenzionálním (2D/3D) prostoru a minimalizuje Kullback-Leiblerovu divergenci mezi těmito dvěmi rozděleními s ohledem na polohu bodů v prostoru. Žádoucí je, aby rozdělení pravděpodobnosti v nízkodimenzionálním prostoru reflektovalo rozdělení pravděpodobnosti v originálním prostoru.

V souvislosti se studiem metagenomu byl tento algoritmus prezentován v publikaci autorů Gisbrecht a kol. [54], jehož ukázka je na Obr. 4.4. Výše zmíněná Barnes-Hutova aproximace tohoto algoritmu je využívána ve studii Laczny a kol. [38]



Obr. 4.4: t-SNE vizualizace metagenomických dat (převzato z [54])

4.6 Metodika hodnocení úspěšnosti klasifikace

V této části uveďme evaluační techniky, které budou sloužit k porovnání jednotlivých metod shlukování a klasifikace.

Pojmem klasifikace rozumíme rozdělení sekvenčních příznaků do tříd (shluků) na základě jejich podobnosti. Přímé znázornění počtu objektů zařazených do odhadované třídy spolu s jejich skutečnou třídou lze vizualizovat s využitím **konfuzní matice** (též matice zmatení, z angl. *confusion matrix*), viz Tab. 4.1. [64]

Tab. 4.1: Konfuzní matice

		Skutečná třída	
		1	2
Odhadovaná třída	1	**a**	**b**
	2	**c**	**d**

Využívanou metrikou pro hodnocení úspěšnosti klasifikace (rozpoznávání) je tzv. **přesnost** (angl. *accuracy*, AC), kterou lze vyjádřit z výše uvedené konfuzní matice podle následujícího vzorce (4.6) [64]

$$AC = \frac{a+d}{a+b+c+d}, \tag{4.6}$$

tedy jako součet prvků na diagonále (správně klasifikované objekty) ku celkovému počtu objektů. Udává se zpravidla v procentech nebo jako desetinné číslo.

5 ANALÝZA SIMULOVANÝCH DAT

Analýza charakteristických četností nukleotidových slov byla provedena na simulovaném metagenomickém datasetu. Důvodem volby simulovaných metagenomických dat pro analýzu je zejména poskytnutí přesné reference o jednotlivých taxonech, které tak slouží pro ověření správnosti výsledků analýzy.

V této kapitole bude popsána praktická realizace analýzy dat. V první části je uveden popis skriptu pro simulaci metagenomu a charakteristika simulovaných dat. Dále je uveden přehled analyzovaných sekvenčních příznaků. Ve třetí části je znázorněno a diskutováno dílčí nastavení algoritmu hierarchické shlukování. Analyzována je také závislost délky slova na přesnosti klasifikace a limitace této metody. Třetí část této kapitoly představuje aplikaci analýzy hlavních komponent a následné K-means shlukování. Zde je mimo jiné diskutována také problematika volby počtu hlavních komponent. Dále je popsán přístup výpočtu konsensu hierarchického a K-means shlukování. Šesté části je věnován rozbor problematiky vizualizace dat. V závěrečných částech této kapitoly jsou pak aplikované metody srovnávány s výsledky jednoho ze současných algoritmů pro vizualizaci dat t-SNE. Představeny jsou také limitace četnostních metod.

Analýza byla implementována v programovacím prostředí MATLAB.

5.1 Charakteristika simulovaných dat

Vstupem simulátoru jsou kompletní bakteriální genomy vybraných organismů, výstupem pak jednotlivé genomické fragmenty (simulovaná „čtení“). Je možné nastavit počet fragmentů pro každý z genomů. Stejně tak lze zvolit délku fragmentu, která je však defaultně nastavená na konstantní. Volitelným nastavením je zadání minimální a maximální délky čtení. Délka fragmentu se pak generuje ze zadaného rozsahu. Počátky fragmentů jsou generovány náhodně z celé délky genomu (Gaussovské rozložení). Volitelným nastavením simulátoru je také lognormální distribuce čtení z jednotlivých genomů. To lze podle Laczny a kol. [38] aplikovat pro simulaci skutečnosti, že různé taxony mají různou četnost.

Simulovaná datová sada SET01 obsahuje genomické fragmenty o délce 1000 bp ze šesti kompletních bakteriálních genomů, viz Tab. 5.1. Délka 1000 bp je podle Laczny a kol. považována také jako minimální doporučená pro aplikaci četnostních metod [38]. Z hlediska sekvenačních technologií lze tato data považovat za sekvenační čtení nejpokročilejší platformy ze skupiny 454 sekvenovacích technologií, GS FLX+

sekvenátoru. Pro každý z těchto genomů bylo vygenerováno 100 fragmentů. Charakteristické pro tento dataset je, že organismy jsou taxonomicky vzdálené na úrovni bakteriálních kmenů.

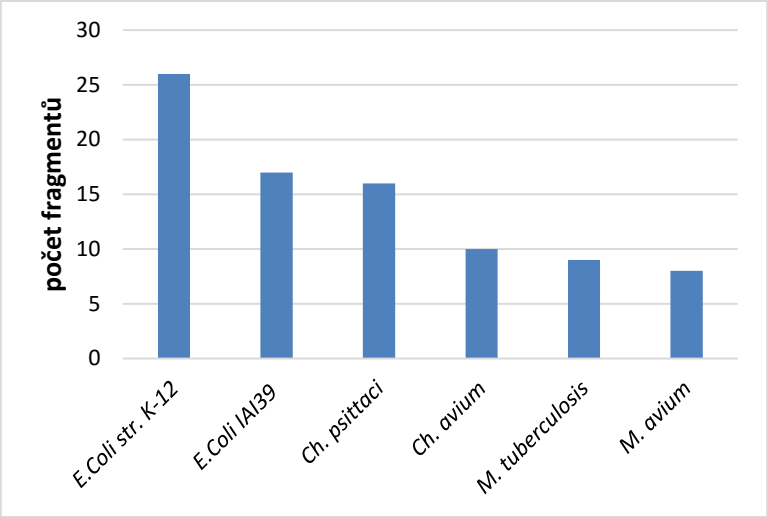
Tab. 5.1: SET01 Seznam kompletních bakteriálních genomů

	Název organismu	**GenBank ref.č.**
1	*Escherichia coli* CFT073	NC_004431.1
2	*Pseudomonas aeruginosa* strain PA1RG	CP012679.1
3	*Geodermatophilus obscurus* DSM 43160	NC_013757.1
4	*Acinetobacter equi* strain 114	CP012808.1
5	*Actinobacillus succinogenes* 130Z	NC_009655.1
6	*Chlamydophila felis* Fe/C-56 DNA	NC_007899.1

Pro doplnění analýzy byl vytvořen další dataset (SET02). Tab. 5.2 představuje vybrané kompletní bakteriální genomy zahrnuté v tomto datasetu. Byl sestaven tak, aby odrážel charakteristiky mikrobiálních komunit, a sice, že mohou zahrnovat taxony, které jsou si velmi blízké a různé taxony mohou mít různou četnost (Obr. 5.1). Data, obsahují fragmenty o délce 1000 bp z různých částí genomů dvojic příbuzných organismů, jako jsou dva kmeny bakterie *E. coli* a dva zástupci rodů *Chlamydia* a *Mycobacterium.*

Tab. 5.2: SET02 Seznam kompletních bakteriálních genomů

	Název organismu	**GenBank ref.č.**
1	*Escherichia coli* str. K-12	NC_000913.3
2	*Escherichia coli* IAI39	NC_011750.1
3	*Chlamydia psittaci* 01DC12	HF545614.1
4	*Chlamydia avium* 10DC88	NZ_CP006571.1
5	*Mycobacterium tuberculosis* H37Rv	NC_000962.3
6	*Mycobacterium avium subsp. paratuberculosis* str. k10	NC_002944.2



Obr. 5.1: Počet fragmentů genomů

5.2 Přehled analyzovaných sekvenčních příznaků

Prvním krokem analýzy je výpočet charakteristických četností nukleotidových slov (*k*-merů). Dle matematických vztahů uvedených v kapitole 3 byl pro každý z fragmentů vypočítán vektor příznaků reprezentujících frekvenci.

V této kapitole budou dále analyzovány a diskutovány metody, jejichž přehled je prezentován v následující tabulce (Tab. 5.3).

Tab. 5.3: Vybrané metody výpočtu charakteristických příznaků sekvencí

Označení metody	**Charakteristické příznaky**	**Autor, rok**
Frekvence *k*-merů	Vektor frekvencí *k*-merů	-
Yang	Pořadí seřazených *k*-merů	Yang a kol., 2013
Ding	Normalizovaná průměrná intervalová vzdálenost *k*-merů	Ding a kol., 2013
Tang	Normalizovaná průměrná relativní vzdálenost *k*-merů	Tang a kol., 2014
Frekvence symetrizovaných *k*-merů	Vektor clr-transformovaných frekvencí symetrizovaných *k*-merů	Gori, 2011; Laczny a kol., 2015

Vhodné je také připomenout, jaká je délka těchto vektorů reprezentujících sekvenci. Dle vztahu (3.1) je počet všech možných slov $n=4^k$. Délka vektoru podle daného výpočtu je platná pro příznaky frekvence *k*-merů, Yang, Ding a Tang. Délka vektoru frekvencí symetrizovaných *k*-merů je odlišná, a to vzhledem k tomu, že sčítáme *k*-mery a jejich reverzní komplementy a v některých případech (pro *k*=2,4,6) se vyskytují také palindromatické *k*-mery. V následující analýze budou obvykle srovnávány metody pro délku slova 2 až 7, uveďme si tedy pro názornost délku vektoru reprezentujících sekvenci (Tab. 5.4) :

Tab. 5.4: Přehled délky vektoru reprezentujícího sekvenci v závislosti na délce *k*-meru

Typ reprezentace /délka *k*-meru	***k*=2**	***k*=3**	***k*=4**	***k*=5**	***k*=6**	***k*=7**
Frekvence *k*-merů (délka vektoru)	16	64	256	1 024	4 096	16 384
Frekvence symetrizovaných *k*-merů (délka vektoru)	10	32	136	512	2 080	8 192

5.3 Hierarchické shlukování

Pro klasifikaci objektů (vektorů příznaků reprezentujících jednotlivé segmenty) do tříd odpovídajících jednotlivým taxonům obsaženým v metagenomickém datasetu byl použit algoritmus hierarchické shlukování.

5.3.1 Volba vzdálenostní metriky a metody shlukování

Jak již bylo představeno v kapitole 4.3, prvním krokem algoritmu je výpočet míry podobnosti mezi objekty, přičemž k tomuto účelu mohou být použity různé vzdálenostní metriky. Následně se provede shlukování objektů, kde je opět možnost volby z několika metod.

Optimální vzdálenostní metriku a metodu shlukování lze podle [57] určit na základě výpočtu hodnot korelačního koeficientu. Pro datovou sadu SET01 bylo provedeno testování pro délku nukleotidového slova *k*=5. Vybrané hodnoty kofenetického korelačního koeficientu jsou uvedeny v tabulce (Tab. 5.5).

Tab. 5.5: Hodnoty kofenetického korelačního koeficientu pro vybrané metody

		Metoda shlukování		
		m. nejbližšího souseda	**m. nejvzdálenějšího souseda**	**Wardova metoda**
metrika	**euklidovská**	0,2496	0,7716	0,7690
	Manhattanská	0,4952	0,814	0,8098
	kosinová	0,5809	0,7827	0,7945
	Spearmanova korelace	0,6415	**0,8309**	0,7605

Na základě kofenetického korelačního koeficientu lze tedy usuzovat, že nejvhodnější metrikou pro hierarchické shlukování je Spearmanova korelace spolu s metodou shlukování nejvzdálenějšího souseda. Avšak praktická realizace a testování na datasetu 600 simulovaných metagenomických fragmentů (SET01) ukázala, že toto doporučení je v rozporu se skutečnou vhodností dílčího nastavení hierarchického shlukování. Testovány byly různé vzdálenostní metriky a míry korelace, přičemž lze konstatovat, že jejich volba má pouze zanedbatelný vliv na úspěšnost klasifikace. Pro výpočet matice podobností mezi objekty (asociační matice) tedy bude pro další analýzu využívána **euklidovská metrika**, která má četné uplatnění v řadě studií založených na analýze *k*-merů a je v mnoha případech považována také jako doporučená [50; 51]. Stěžejní je však volba optimální metody shlukování, která determinuje vzhled výsledného dendrogramu, zařazení do shluků a odtud i úspěšnost (přesnost) klasifikace.

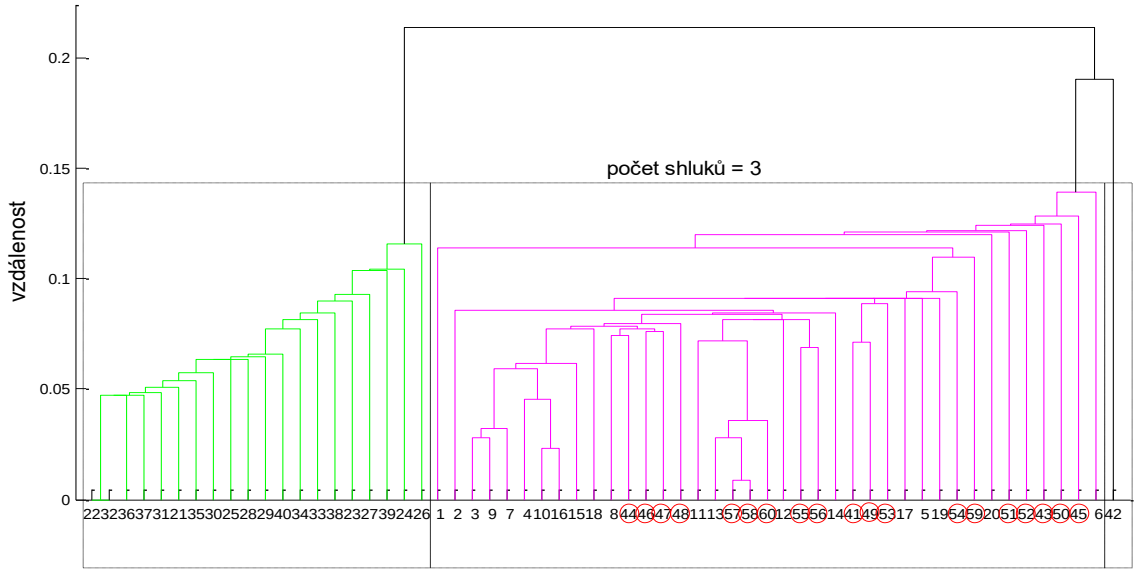
Výše uvedené tvrzení zdůvodněme na následujícím příkladu. Pro názornost byl zvolen dataset skládající se z 60 genomických fragmentů o délce 1000 bp se zastoupením 3 taxonů, viz Tab. 5.6. Analyzované příznaky jsou v tomto případě frekvenční vektory 5-merů.

Tab. 5.6: Označení genomických fragmentů

Taxony	**Číselné označení fragmentů**
Streptococcus pneumoniae	1 - 20
Gordonia bronchialis	21 - 40
Acinetobacter equi	41 - 60

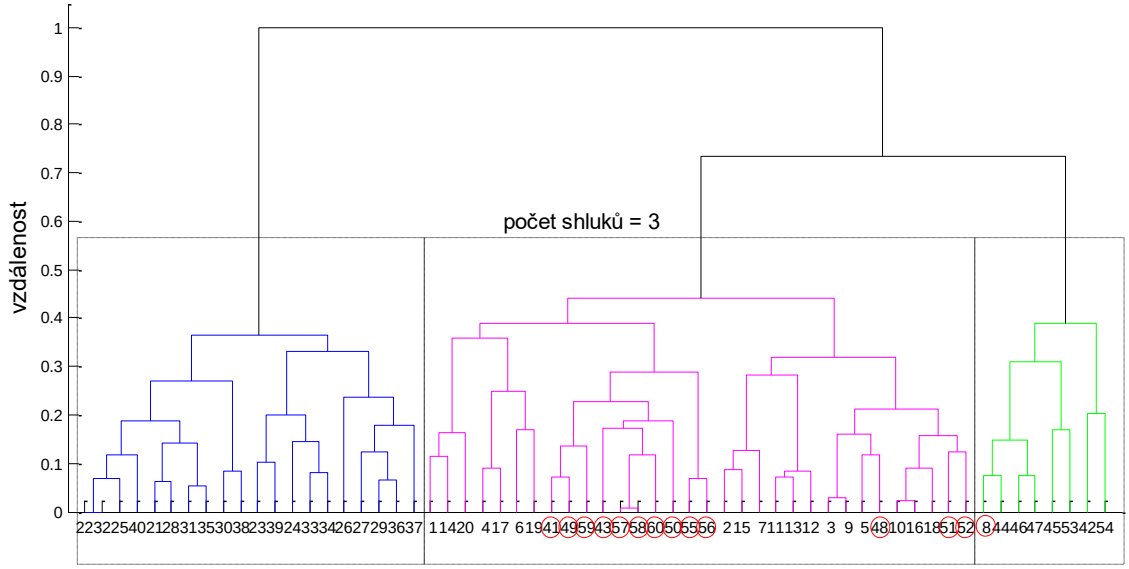
Na následujících obrázcích jsou zobrazeny výsledné dendrogramy pro euklidovskou vzdálenost mezi objekty s využitím metod shlukování nejbližšího souseda, nejvzdálenějšího souseda a Wardovy metody. Přičemž požadováno je rozdělení genomických fragmentů do 3 shluků dle jejich taxonomického původu. Nesprávně zařazené objekty jsou pak označeny červeným kruhem.

Jak je z grafu (Obr. 5.2) patrné, při použití **metody nejbližšího souseda** dojde k tvorbě dlouhých zřetězených shluků. Dále může dojít k tvorbě velmi malých shluků pro objekty na krajích řetězce, obsahující např. v tomto případě pouze jeden objekt (č. 42), což se pro naši aplikaci jeví jako nevhodné.



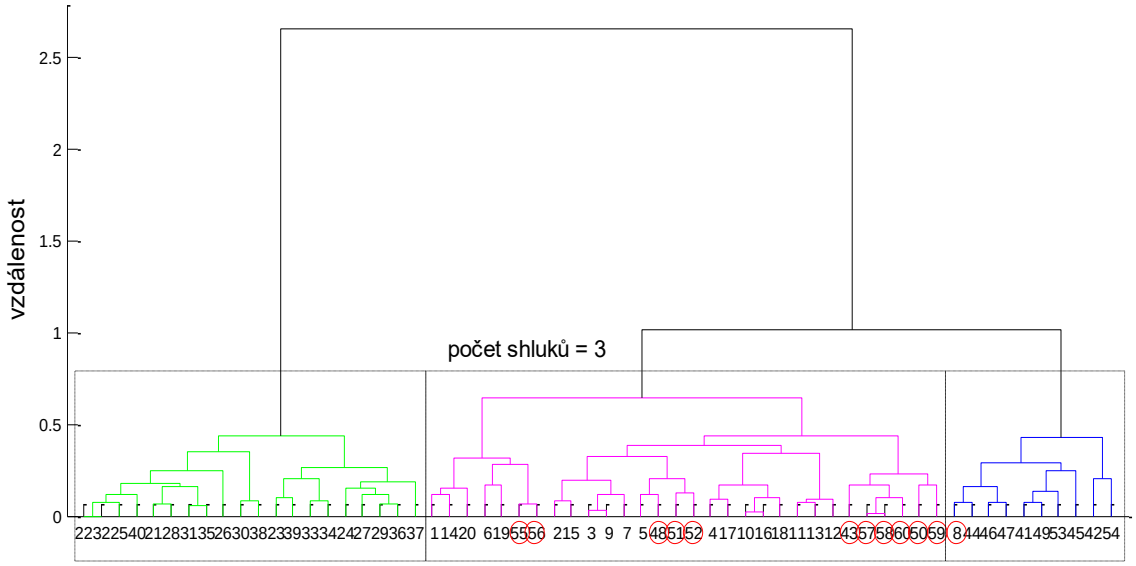
Obr. 5.2: Dendrogram - euklidovská vzdálenost, metoda nejbližšího souseda

Metoda nejvzdálenějšího souseda naopak zabraňuje vzniku zřetězených shluků a produkuje shluky, které jsou mezi sebou dobře odděleny, jak je zobrazeno na Obr. 5.3. Při testování na datasetu SET01 má však mnohdy tendenci tvořit menší shluky, obdobně jako je tomu i v případu metody nejbližšího souseda.



Obr. 5.3: Dendrogram - euklidovská vzdálenost, metoda nejvzdálenějšího souseda (úspěšnost klasifikace 76 %)

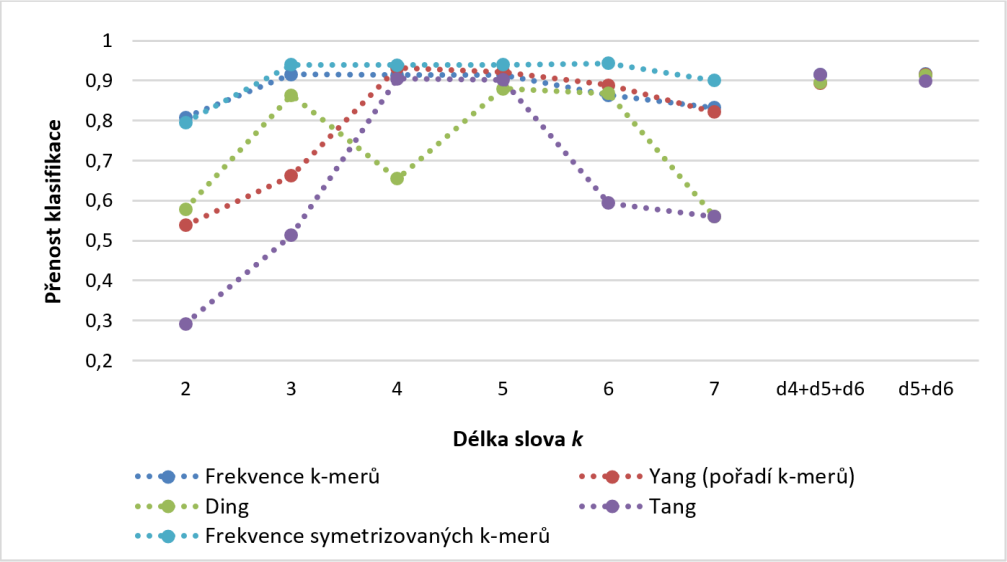
Jako nejvhodnější se jeví **Wardova metoda** (Obr. 5.4), která tvoří kompaktní, dobře oddělené shluky sférického charakteru. Takové shluky mají obvykle přibližně stejnou velikost a výše uvedené nedostatky lze s jejím použitím úspěšně eliminovat.



Obr. 5.4: Dendrogram - euklidovská vzdálenost, metoda Wardova (úspěšnost klasifikace 80 %)

5.3.2 Analýza přesnosti klasifikace pro různou délku *k*-merů

Dále byla testována úspěšnost klasifikace s ohledem na délku analyzovaných nukleotidových slov a také jejich kombinací. V následujícím grafu je zobrazena přesnost klasifikace do taxonomických skupin pro vybrané metody výpočtu příznaků. Analýza byla provedena na datasetu SET01. Vzdálenostní metrikou použitou pro výpočet vzdálenosti mezi objekty je euklidovská a metodou shlukování Wardova metoda.



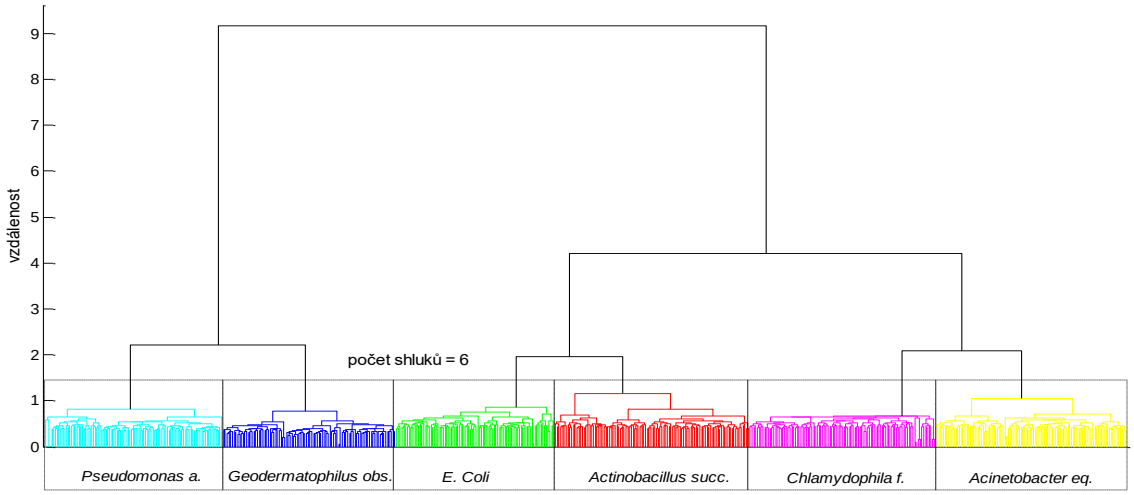
Obr. 5.5: Vliv délky *k*-meru na přesnost klasifikace (hierarchické shlukování)

Z grafu na Obr. 5.5 lze pozorovat, že úspěšnost algoritmu obvykle roste s délkou analyzovaných *k*-merů pro délku slova 2 až 5. Pro slova delší než 5 nebyl zaznamenán výrazný nárůst v úspěšnosti klasifikace. Lze tak potvrdit poznatky studie Higashi a kol. [65], že délka analyzovaných *k*-merů se jeví jako limitovaná délkou sekvence. V našem případě lze tedy usoudit, že pro sekvence o délce přibližně 1000 bp, není ve většině případů účinné analyzovat *k*-mery delší než 5. Výsledky korespondují také s poznatky Laczny a kol. [38], kdy uvádí jako nejvhodnější výpočet *k*-merů pro k=5, mimo jiné také s ohledem na výpočetní náročnost.

Studovány byly rovněž kombinace slov délky *k*=4, 5, 6 a *k*=5, 6. S využitím první z těchto kombinací (v grafu označená jako d4+d5+d6) lze vylepšit přesnost metod Ding a Tang. Kombinace d5+d6 dává lepší výsledek než samostatná analýza pentamerů pro metodu Ding. U dalších metod však nebylo zaznamenáno vylepšení oproti využití samostatné délky slova *k*=5.

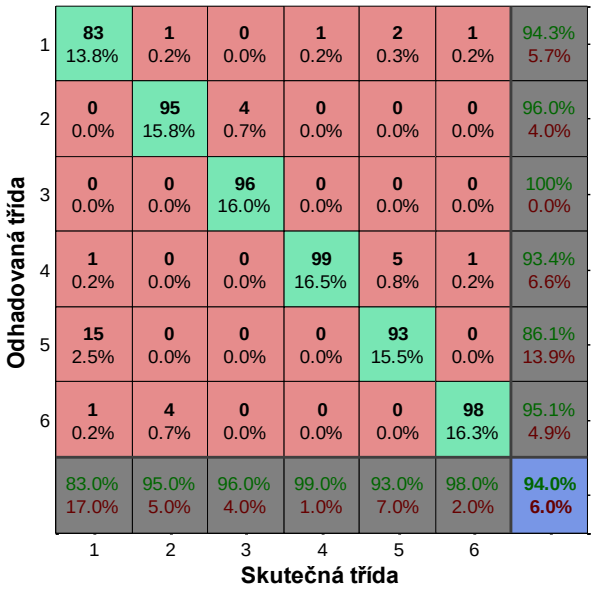
Celkově lze vyhodnotit, že nejlepší výsledky dávají algoritmy založené na výpočtu příznaků **frekvence *k*-merů**, **Yang** (pořadí seřazených četností *k*-merů) a **frekvence symetrizovaných *k*-merů**.

V následujícím grafu (Obr. 5.6) je uveden příklad analýzy 600 genomických fragmentů SET01 pro frekvence symetrizovaných 5-merů. Z dendrogramu je patrné, že dochází k tvorbě dobře definovatelných shluků. Při určení výsledného počtu shluků rovno 6 dojde ke klasifikaci genomických fragmentů do tříd odpovídajících jejich skutečné taxonomické příslušnosti pouze s minimem nesprávně klasifikovaných objektů (úspěšnost klasifikace 94 %). Přehled objektů zařazených do jednotlivých tříd spolu s jejich skutečnou příslušností k jednotlivým bakteriálním genomům jsou uvedeny s využitím konfuzní matice na Obr. 5.7.



Obr. 5.6: SET01 Analýza frekvencí symetrizovaných 5-merů

1	*Escherichia coli*
2	*Pseudomonas aeruginosa*
3	*Geodermatophilus obs.*
4	*Acinetobacter equi*
5	*Actinobacillus succ.*
6	*Chlamydophila felis*



Obr. 5.7: Označení tříd (vlevo) a konfuzní matice, klasifikace do 6 tříd (vpravo)

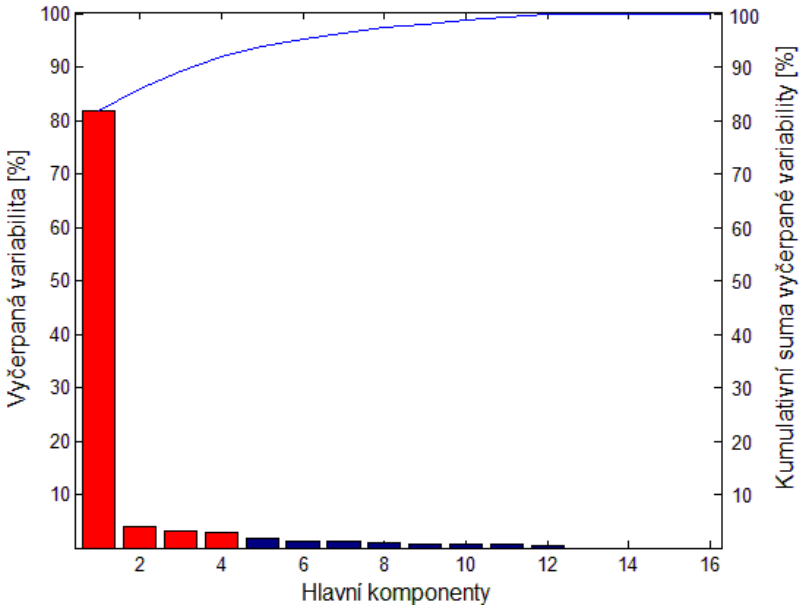
5.4 Analýza hlavních komponent a K-means

Vzhledem k analýze vysokodimenziálních vektorů příznaků reprezentujících genomické fragmenty je pro redukci dimenzionality využívána analýza hlavních komponent (PCA). Tato část se zabývá problematikou volby počtu hlavních komponent a jejich vlivem na procenta vyčerpané variability. Na zvoleném počtu hlavních komponent je následně provedeno shlukování algoritmem K-means.

5.4.1 Vliv počtu hlavních komponent na procenta vyčerpané variability

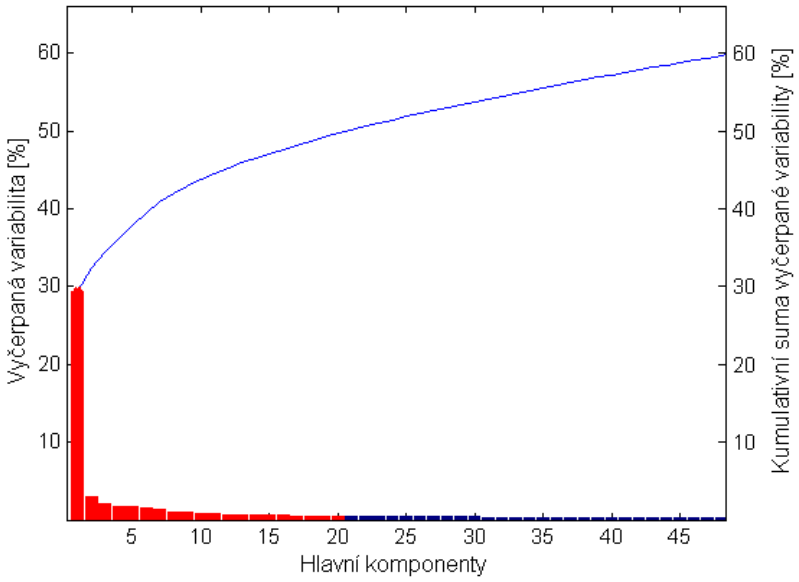
V analýze hlavních komponent je žádoucí redukovat data na co nejmenší počet hlavních komponent (dimenzí), které zároveň vyčerpávají co největší množství variability dat. Objektivním měřítkem takové volby je bod zlomu ve scree plotu. Vzhledem k tomu, že je tato analýza aplikována na různé sekvenční příznaky pro různou délku *k*-meru, je vhodné diskutovat optimální volbu hlavních komponent a vyčerpanou variabilitu, kterou tyto komponenty popisují.

V následujícím grafu (Obr. 5.8) je zobrazen scree plot pro příznaky frekvence dimerů (SET01). První 4 PC (vyznačené červeně) zde celkem vyčerpávají 91,95 % variability. Lze tedy usoudit, že originální data o 16 dimenzích lze dobře vysvětlit s využitím 4 hlavních komponent.



Obr. 5.8: Scree plot, analýza frekvencí 2-merů

Další ukázkou na Obr. 5.9 je scree plot pro příznaky frekvence pentamerů. Zde je na první pohled možno pozorovat, že první hlavní komponenta vyčerpává znatelně menší procento variability (29 %) než je tomu v předchozím případu. Z kumulativní sumy variability vyčerpané hlavními komponentami (modrá křivka) je patrné, že ani volba 50 PC nevyčerpá více než 60 % variability dat. Vzhledem k charakteru analyzovaných dat (1024 dimenzí) byl v tomto případě zvolen počet hlavních komponent roven 20.



Obr. 5.9: Scree plot, analýza frekvencí 5-merů

Během analýzy je tedy nutno čelit problému volby optimálního počtu hlavních komponent. V případě, že je užito příliš málo hlavních komponent, nedostatečné vysvětlení dat vede ke ztrátě informace. Pokud je do modelu PCA zahrnuto příliš mnoho PC, může být zahrnut také šum, což má za následek větší chybovost v úspěšnosti klasifikace. Tento problém byl řešen opětovným rozborem scree plotu a rovněž testováním vlivu zahrnutých komponent na následnou úspěšnost klasifikace. Z tohoto testování lze uvést následující poznatky. Doporučená kumulativní suma variability, která by měla být vyčerpána hlavními komponentami spolu s doporučeným počtem hlavních komponent pro jednotlivé nastavení jsou uvedeny v tabulce (Tab. 5.7). Počet PC uvedený pro frekvence *k*-merů je platný též pro metody Yang, Ding a Tang vzhledem ke stejné délce vektorů příznaků charakterizujících genomické fragmenty. Pro délky *k*-merů rovno 6 a 7 nebyla stanovena doporučení týkající se procent vyčerpané variability, vzhledem k velmi malému podílu vyčerpané variability jednotlivými hlavními komponentami. Analýzou scree plotu však lze doporučit využití 20 PC a pro příznaky frekvence symetrizovaných *k*-merů 10 PC.

Tab. 5.7: Doporučení pro volbu počtu hlavních komponent

	***k*=2**	***k*=3**	***k*=4**	***k*=5**	***k*=6**	***k*=7**
Doporučená vyčerpaná variabilita popsaná PC	80 %	70 %	60 %	50 %	-	-
Počet PC (frekvence *k*-merů)	4	6	10	20	20	20
Počet PC (frekvence symetr. *k*-merů	2	3	5	10	10	10

5.4.2 Nastavení K-means shlukování

Algoritmus K-means je dále aplikován na zvolený počet hlavních komponent dle výše uvedených pravidel.

Vzhledem k tomu, že metrikou využívanou pro výpočet asociační matice pro hierarchické shlukování je euklidovská, je tato metrika využívána i v algoritmu K-means.

Podstatným vstupním parametrem tohoto algoritmu je volba počtu shluků (tříd), do kterých má algoritmus rozdělit data. Při analýze simulovaného metagenomického dataset SET01 je žádoucí rozdělení dat do 6 shluků dle počtu taxonů, které se v něm vyskytují.

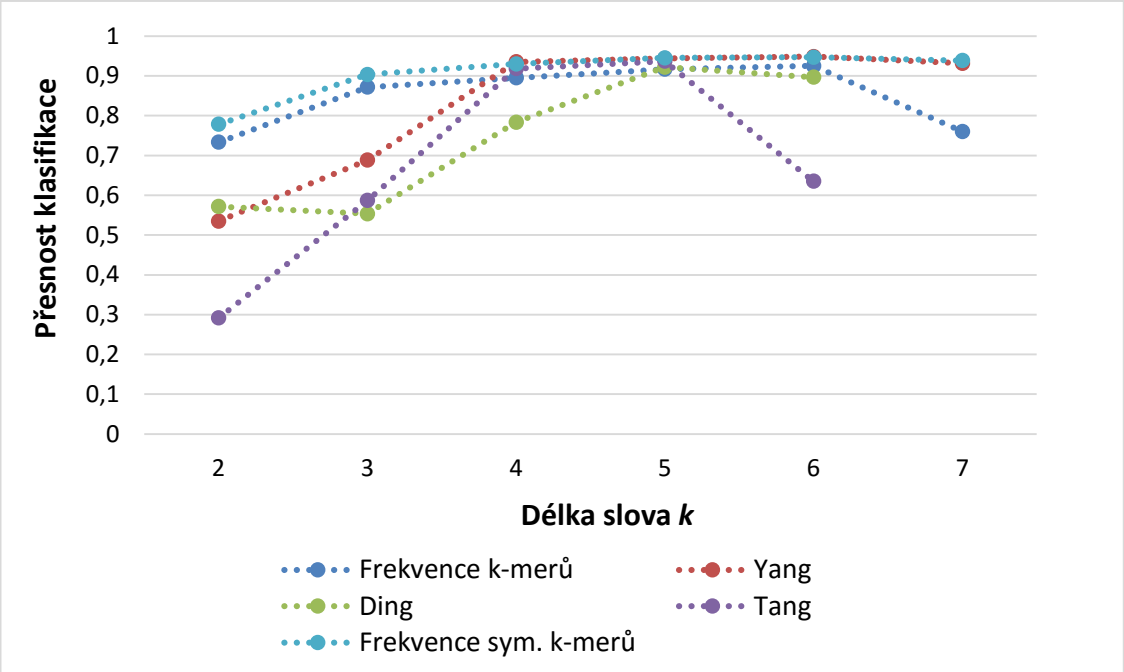
Následně jsou iterativně hledány centroidy shluků. Počáteční pozice centroidů jsou zvoleny náhodně. Pro zamezení uvíznutí centroidů v lokálních minimech je vhodné nastavit větší počet replikací. Algoritmus pak vybere nejlepší řešení z pozic centroidů na základě minimální vnitroshlukové sumy vzdáleností objektů k centroidu. V analyzovaných případech bylo zvoleno s ohledem na výpočetní náročnost 5 replikací.

5.4.3 Analýza přesnosti klasifikace pro různou délku *k*-meru

Na Obr. 5.10 je uveden graf znázorňující vliv použité délky *k*-meru na přesnost klasifikace vybraných metod. S výjimkou metody Ding lze pozorovat nárůst v přesnosti klasifikace pro délku slova 2 až 5. Metody Yang a frekvence symetrických *k*-merů pak dávají nejlepší výsledky pro délku slova *k*=6. Tyto výsledky jsou konzistentní s poznatky získanými analýzou hierarchickým shlukováním.

Metody Ding a Tang obvykle nevyhovují doporučením ohledně procent vyčerpané variability popsané zvoleným počtem hlavních komponent, viz Tab. 5.8. Tyto metody také netvoří dobře identifikovatelné shluky. Při využití délky slova *k*=7 pak dochází ke shlukování dat do jednoho výrazného shluku.

S ohledem na nejmenší variabilitu v úspěšnosti klasifikace lze usoudit, že optimální délka *k*-meru pro aplikaci metody PCA a následovného K-means shlukování je *k*=5.

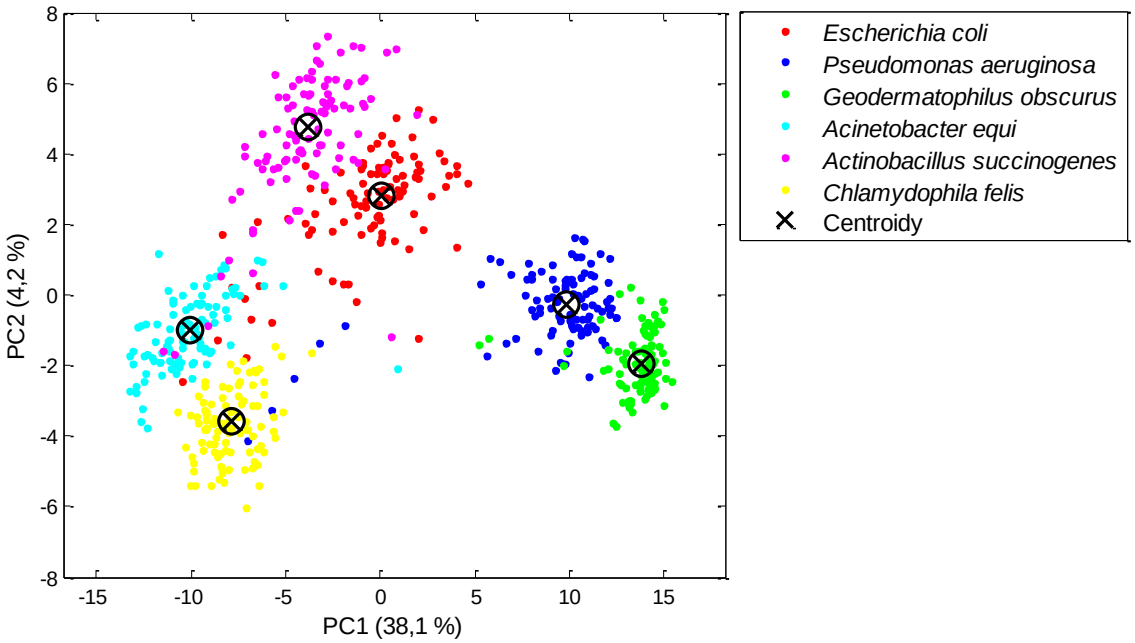


Obr. 5.10: Vliv délky *k*-meru na přesnost klasifikace (PCA a K-means)

Tab. 5.8: Kumulativní suma vyčerpané variability [%]

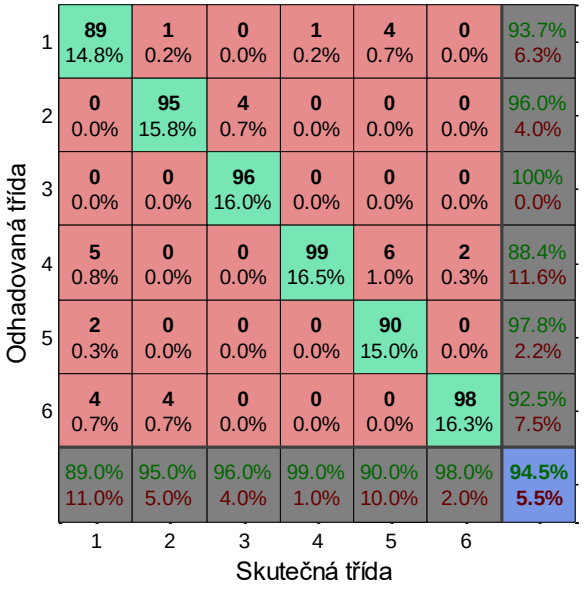
Délka slova *k*	**2**	**3**	**4**	**5**	**6**	**7**
Frevence *k*-merů	91,95	82,05	68,37	50,16	26,61	13,632
Yang	82,77	76,63	61,64	52,36	20,96	11,71
Ding	92,2752	45,49	21,09	22,7	14,56	12,32
Tang	60,2538	55,12	43,62	31,59	18,74	15,19
Frekvence symetrizovaných *k*-merů	94,59	88,6887	72,46	51,26	28,40	14,05

Na následující ukázce (Obr. 5.11) jsou vizualizována data SET01 pro sekvenční příznaky frekvence symetrizovaných 5-merů s využitím prvních dvou hlavních komponent. Jak lze pozorovat, data tvoří relativně dobře rozlišitelné shluky. V grafu jsou rovněž vyznačeny centroidy shluků. Konfuzní matice na Obr. 5.12 (vpravo) znázorňuje klasifikaci objektů do tříd a jejich skutečnou příslušnost do tříd. Aplikací PCA a následném K-means shlukování na 20 PC lze tato data klasifikovat do 6 tříd dle příslušnosti k jednotlivým taxonům s 94,5% úspěšností klasifikace.



Obr. 5.11: SET01, frekvence symetrizovaných 5-merů, PCA

1	*Escherichia Coli*
2	*Pseudomonas aeruginosa*
3	*Geodermatophilus obs.*
4	*Acinetobacter equi*
5	*Actinobacillus succ.*
6	*Chlamydophila felis*

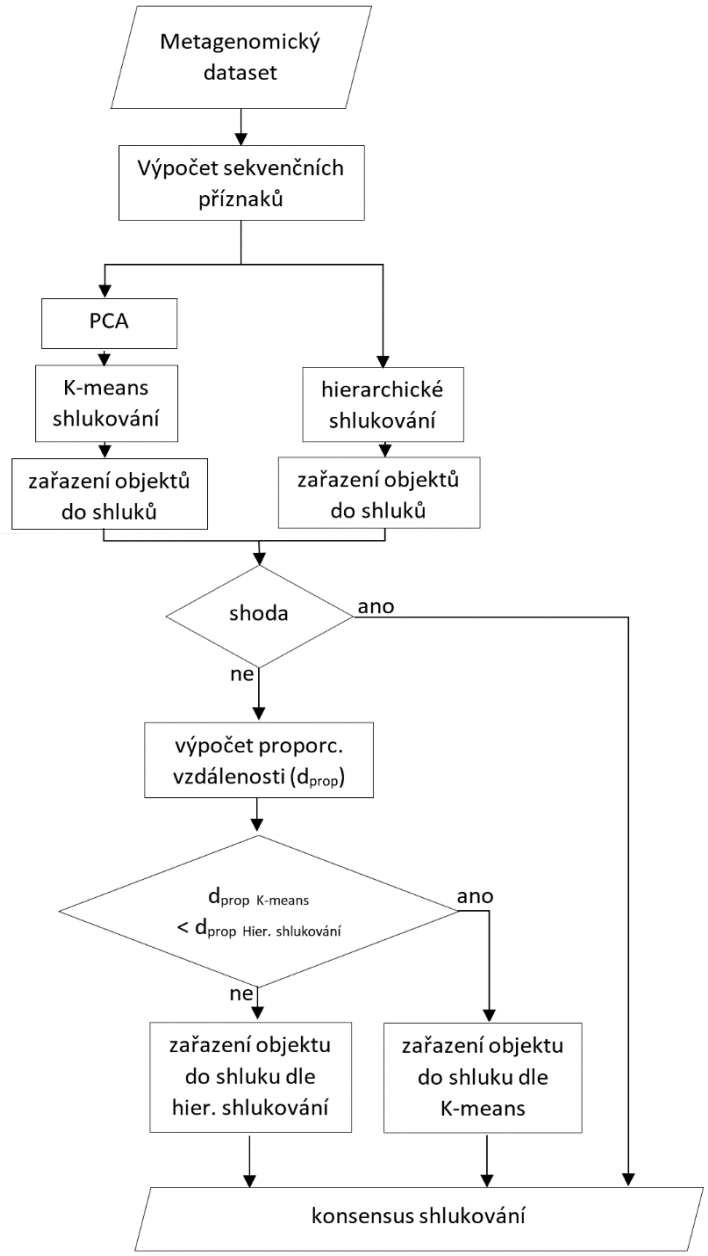


Obr. 5.12: Označení tříd (vlevo) a konfuzní matice, klasifikace do 6 tříd (vpravo), PCA + K-means shlukování

5.5 Konsensus shlukování

Dalším přístupem klasifikace je vytvoření konsensu výsledků hierarchického shlukování (na datech o původním počtu dimenzí) a K-means shlukování (na redukovaném počtu dimenzí).

Logika metody spočívá v tom, že kombinuje poznatky získané hierarchickým a K-means shlukováním pro rozhodnutí o zařazení objektů do shluků. Schematický postup tohoto algoritmu je znázorněn na Obr. 5.13.

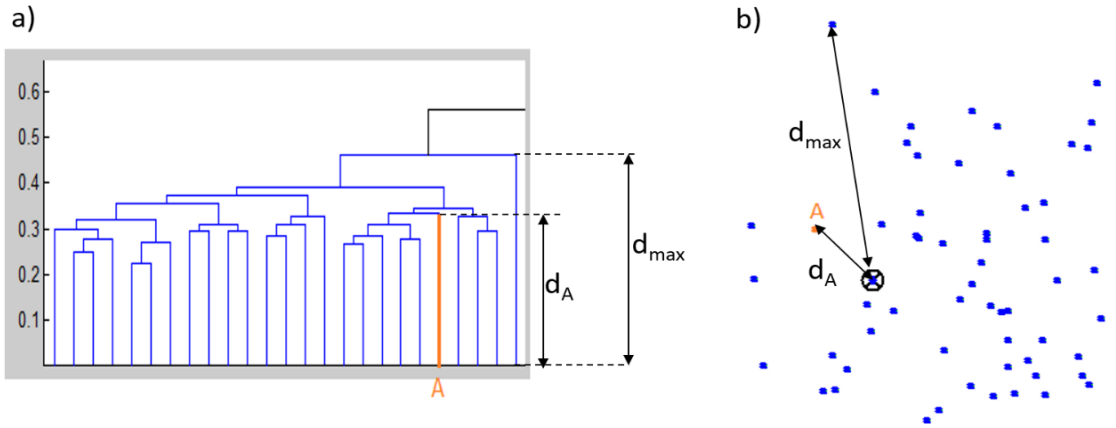


Obr. 5.13: Blokové schéma přístupu konsensus shlukování

Podstatné kroky této metody jsou následující. Nejprve je porovnáváno zařazení objektů do tříd (shluků). V případě, že nejsou u obou algoritmů zařazeny do stejného shluku, je počítána proporcionální vzdálenost objektů. Proporcionální vzdálenost pro objekt A se pak vypočítá dle vzorce (5.1):

$$d_{prop} = \frac{d_A}{d_{max}}. \tag{5.1}$$

Jako proporcionální vzdálenost rozumíme délku větve dendrogramu vedoucí k objektu A ku maximální délce ve shluku, viz Obr. 5.14 vlevo. V případě K-means shlukování je to vzdálenost objektu A k centroidu shluku (d_A) děleno maximální vzdáleností k centroidu v daném shluku (d_{max}), konkrétně vyznačeno na Obr. 5.14 vpravo. Metoda, která dává menší proporcionální vzdálenost je pak využita pro přiřazení objektu do shluku.



Obr. 5.14: Znázornění vzdálenosti objektu A: a) hierarchické shlukování, b) K-means shlukování

Metoda obvykle rozhoduje o zařazení objektů, které se nachází na hranici dvou blízkých, nepříliš dobře odlišitelných shluků. Srovnání tohoto přístupu s klasifikací algoritmem K-means a hierarchickým shlukováním bude uvedeno v závěrečné části této kapitoly.

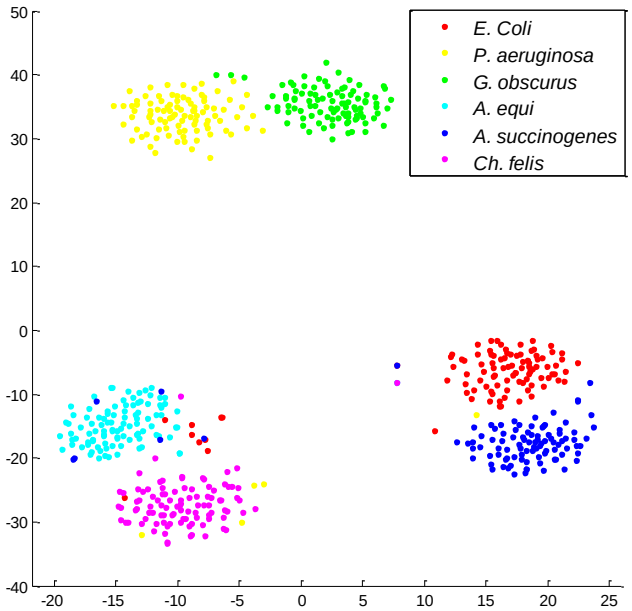
5.6 Vizualizace dat

Výsledek hierarchického shlukování je možné vizualizovat prostřednictvím dendrogramu. Topologie dendrogramu obvykle umožňuje rozlišit jednotlivé shluky a následně rozhodnout o počtu požadovaných shluků, a to buď intuitivně s ohledem na předchozí znalost o analyzovaném datasetu nebo s využitím matematických metod, např. metoda siluety.

Vizualizace dat je jednou z limitací týkající se aplikace algoritmu K-means. Analýza shlukováním je ve všech případech provedena na větším počtu hlavních komponent (4 a více) než je možné vizualizovat. Dostupným řešením je využít první dvě hlavní komponenty, jako bylo ukázáno na Obr. 5.11, případně 3 PC pro 3D vizualizaci. Avšak zanedbání dalších komponent může zavádět zkreslení a nemusí na první pohled vypovídat o skutečném počtu shluků, který data tvoří.

Pro srovnání a doplnění analýzy je v této souvislosti prezentován jeden z aktuálních přístupů vizualizace dat, algoritmus t-SNE. V následující analýze bylo použito defaultní nastavení tohoto algoritmu s počtem vstupních dimenzí redukovaných PCA na 50, jako je doporučeno v [62].

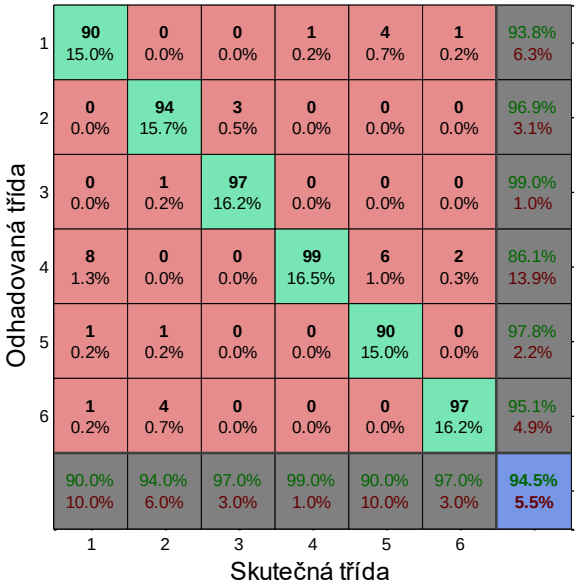
Na Obr. 5.15 je vizualizován dataset SET01 s vyžitím sekvenčních příznaků frekvencí symetrizovaných 5-merů. Lze pozorovat 6 vzájemně se nepřekrývajících disjunktních shluků, které odpovídají taxonomickému složení tohoto simulovaného metagenomického datasetu.



Obr. 5.15: t-SNE vizualizace SET01, použité příznaky frekvence symetrizovaných 5-merů

Automatická klasifikace do předem předem zvoleného počtu shluků je následně aplikována s využitím algoritmu K-means. Pro výše uvedený případ lze tak dosáhnout přesnosti klasifikace 94,5 % (viz konfuzní matice na Obr. 5.16), tedy stejné jako byla dosažena s využitím aplikace PCA a K-means algoritmu.

1	*Escherichia Coli*
2	*Pseudomonas aeruginosa*
3	*Geodermatophilus obs.*
4	*Acinetobacter equi*
5	*Actinobacillus succ.*
6	*Chlamydophila felis*



Obr. 5.16: Označení tříd (vlevo) a konfuzní matice, klasifikace do 6 tříd (vpravo), t-SNE + K-means shlukování, příznaky frekvence symetrizovaných 5-merů

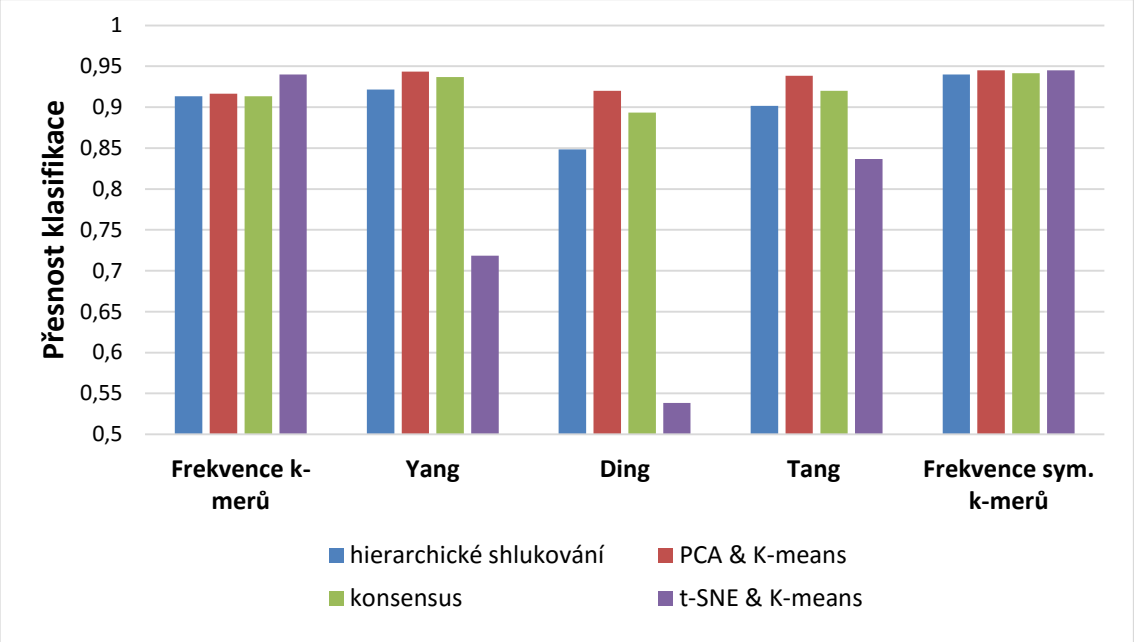
5.7 Srovnání metod klasifikace

V této části bude uveden přehled úspěšnosti klasifikace pro aplikované metody. Do tohoto srovnání byly zahrnuty metody extrakce příznaků genomických fragmentů založené na studiu pentamerů (k=5). Z analýzy uvedené výše vyplývá, že se jedná o vhodnou délku k-merů s ohledem na konzistentnost přesnosti klasifikace pro studované metody.

Z grafu na Obr. 5.17 lze vyhodnotit, že hierarchické shlukování dává nejlepší výsledky pro metody frekvence k-merů, Yang (pořadí seřazených četností k-merů) a frekvence symetrizovaných k-merů. Nejlepších výsledků pro PCA redukci dimenzí a následné K-means shlukování je dosaženo metodami výpočtu příznaků Yang (20 PC) a frekvence symetrizovaných k-merů (10 PC), a to až 94% úspěšnosti klasifikace. Konsensus shlukování pak slouží k validaci zařazení objektů s využitím obou metod. U žádné z metod výpočtu příznaků však tímto způsobem nelze zlepšit přesnost klasifikace objektů do tříd.

Realizované metody klasifikace jsou srovnávány s algoritmem pro vizualizaci t-SNE, na jehož výstup je aplikováno K-means shlukování. Z přehledu úspěšnosti klasifikace analyzovaných příznaků (Obr. 5.17) lze říci, že tento algoritmus je nevhodný pro metodu Ding. Rovněž pro metody Yang a Tang dosahuje kombinace t-SNE a K-means shlukování menší přesnosti klasifikace než srovnávané metody. Zlepšení bylo naopak zaznamenáno pro příznaky frekvence 5-merů. V případě analýzy frekvencí symetrizovaných 5-merů dosahuje algoritmus srovnatelné úspěšnosti jako kombinace PCA analýzy a K-means shlukování

Celkově lze posoudit, že frekvence symetrizovaných 5-merů jsou příznaky, se kterými je možné dosáhnout největší úspěšnosti v klasifikaci, tedy více než 94 % pro všechny z hodnocených metod klasifikace.



Obr. 5.17: Srovnání metod klasifikace

5.8 Limitace četnostních metod

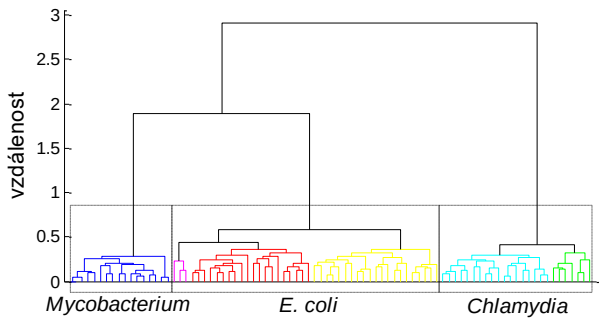
V této části bude analýzou SET02 představena limitace algoritmů založených na studiu četností *k*-merů v souvislosti s rozlišením shluků u taxonomicky příbuzných organismů. SET02 obsahuje simulovaná metagenomická čtení o délce 1000 bp. Taxonomické složení a číselné označení těchto genomických fragmentů je představeno v Tab. 5.9

Tab. 5.9: Označení genomických fragmentů SET02

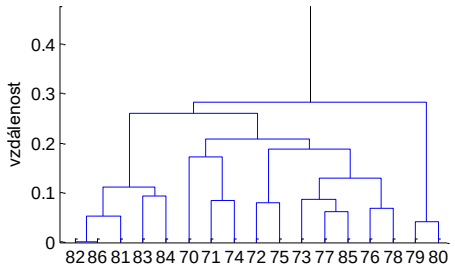
Taxon	**Číselné označení fragmentů**
Escherichia coli str. K-12	1-26
Escherichia coli IAI39	27-43
Chlamydia psittaci 01DC12	44-59
Chlamydia avium 10DC88	60-69
Mycobacterium tuberculosis H37Rv	70-79
Mycobacterium avium subsp. paratuberculosis str. k10	80-87

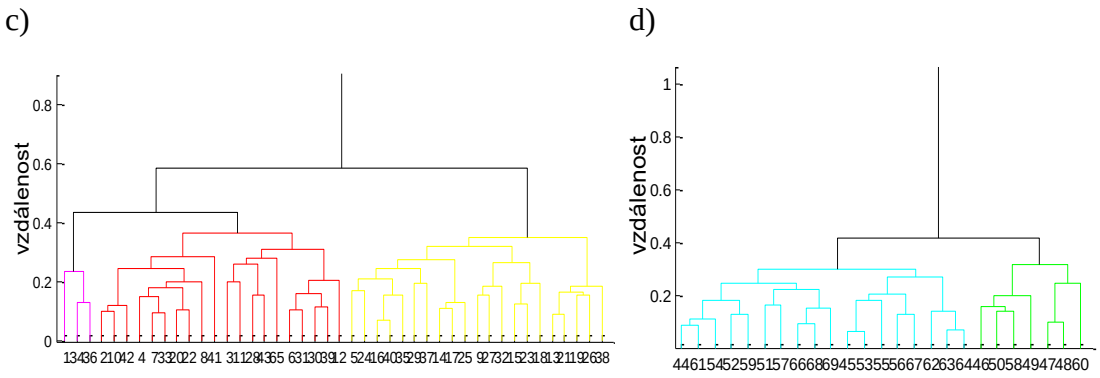
Na Obr. 5.18a je prezentován dendrogram pro sekvenční příznaky frekvence symetrizovaných 5-merů. Počet genomů zařazených v tomto datasetu je 6, přičemž zahrnuje tři dvojice vzájemně příbuzných organismů. Při nastavení počtu shluků rovno 3 dojde k vytvoření shluků, které odpovídají rozlišení na úrovni bakteriálních rodů. Při nastavení požadovaného počtu shluků rovno 6 (barevné kódování v dendrogramu) však dojde k rozdělení objektů do skupin, které neodpovídají skutečnému taxonomickému původu analyzovaných sekvencí. Na Obr. 5.18b je ukázka výřezu shluku odpovídajícího fragmentům rodu *Mycobacterium*. Všechny fragmenty byly zařazeny správně do shluku odpovídajícího rodu, avšak z hlediska druhu je není možné rozlišit. Na Obr. 5.18c byly genomické fragmenty naopak zařazeny do 3 shluků, což ale neodpovídá jejich skutečné příslušnosti k různým bakteriálním kmenům *Escherichia coli*. U genomických fragmentů dvou zástupců rodu *Chlamydia* došlo správně k vytvoření 2 shluků, avšak ty neodpovídají jejich referenčnímu taxonomickému zařazení.

a)



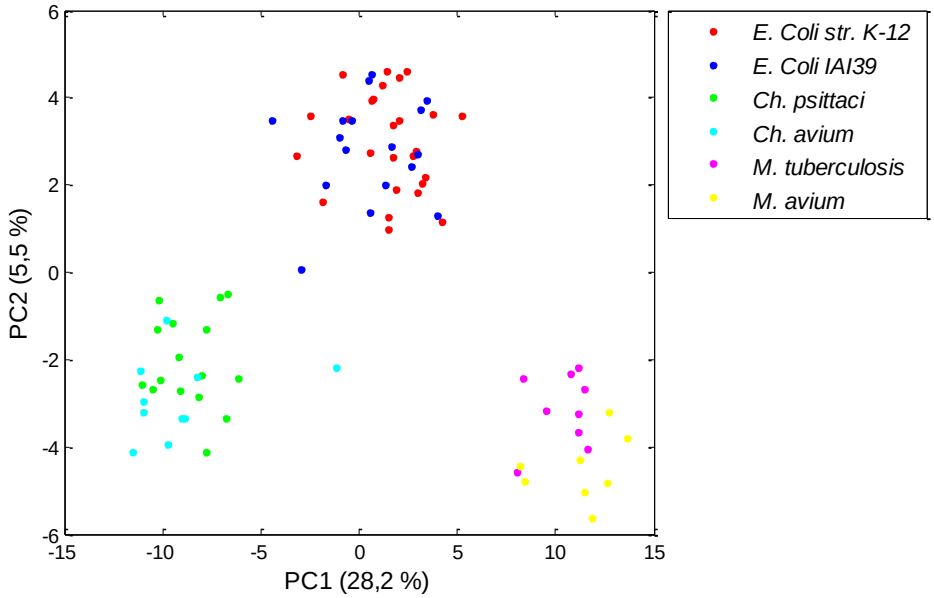
b)





Obr. 5.18: SET02, frekvence symetrizovaných 5-merů, a) celkový vzhled dendrogramu, b) výřez shluku genomických fragmentů rodu *Mycobacterium*, c) výřez shluku genomických fragmentů druhu *Escherichia coli*, d) výřez shluku genomických fragmentů rodu *Chlamydia*

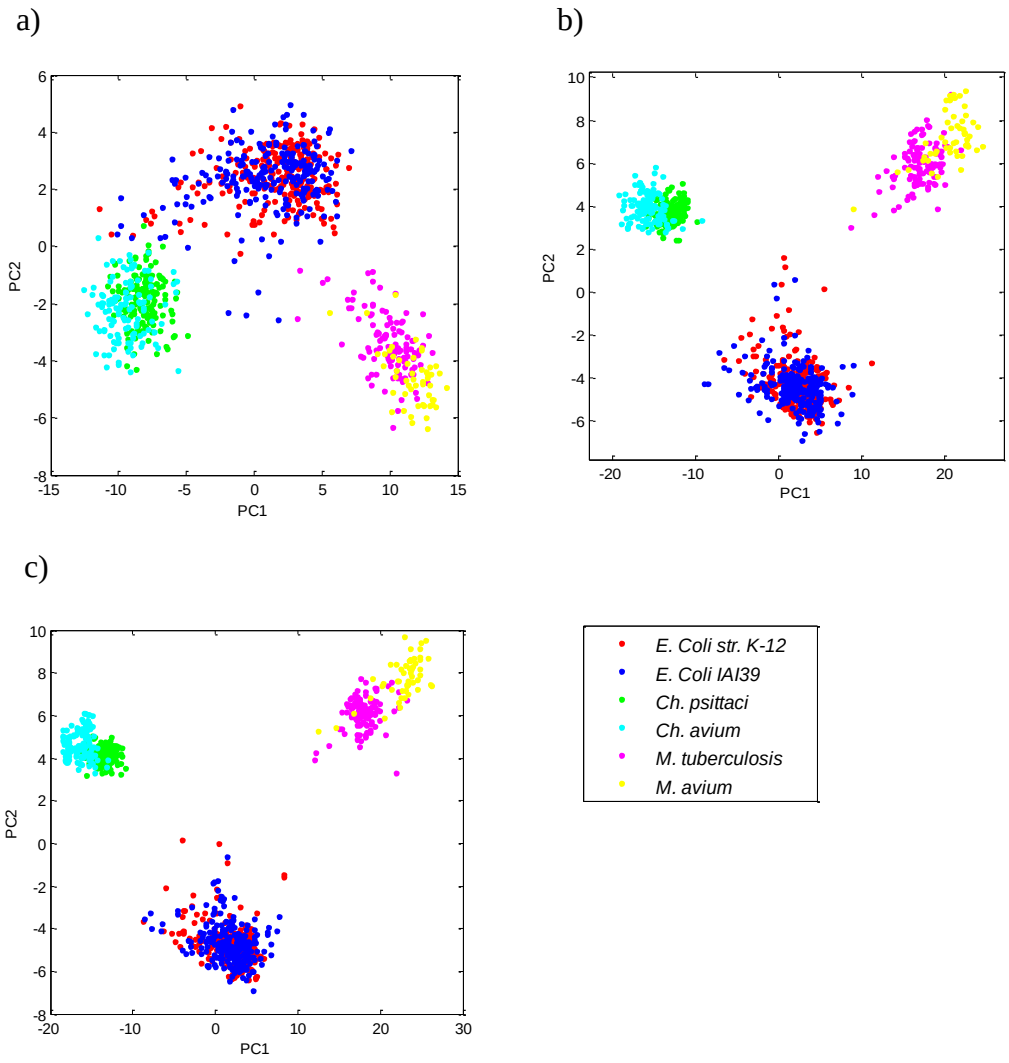
Dále byla provedena PCA analýza. Z vizualizace prvních dvou hlavních komponent lze na první pohled usoudit, že data tvoří tři zřetelné shluky. Dle teoretických předpokladů tvoří jeden ze shluků genomické fragmenty kmenů bakterie *Escherichia coli*. Dále je možno pozorovat, že genomické fragmenty organismů *Chlamydia psittaci* a *Chlamydia avium*, netvoří vzájemně disjunktní shluky, nýbrž se překrývají a tvoří jeden shluk, který však odpovídá rodové příslušnosti. Stejně tak je tomu i pro genomické fragmenty organismů *Mycobacterium tuberculosis* a *Mycobacterium avium.*



Obr. 5.19: Vizualizace SET02, PCA, frekvence symetrizovaných *k*-merů

Obr. 5.20a pak zobrazuje PCA analýzu datasetu obdobného taxonomického zastoupení jako SET02, je však tvořen 915 genomickými fragmenty. Výše uvedené poznatky je možné potvrdit. Lze tedy usuzovat, že pro genomické fragmenty o délce 1000 bp nejsou aplikované příznaky založené na výpočtu četností *k*-merů druhově specifické.

Testování PCA analýzy na genomických fragmentech o délce 8000 bp a 15 000 bp (pozn. délky čtení reálné pro sekvenační technologie třetí generace) ukazuje, že shluky přestavující genomické fragmenty jednotlivých bakteriálních druhů rodu *Mycobacterium* a *Chlamydia* jsou již lépe rozlišeny, viz Obr. 5.20b,c. Analýza je doplněna přehledem vnitroshlukové sumy vzdáleností objektů k centroidu shluku v Tab. 5.10. Lze pozorovat, že se vzrůstající délkou analyzovaných segmentů klesá vnitroshluková suma.



Obr. 5.20: Vizualizace genomických fragmentů o délce a) 1000 bp, b) 8000 bp, c) 15 000 bp. Taxonomická příslušnost organismů je pro všechny případy zakódována barevně dle legendy.

Tab. 5.10: Přehled vnitroshlukové sumy

	1 000 bp	**8 000 bp**	**15 000 bp**
Ch. psittaci	3544,94	892,78	696,06
Ch. avium		936,96	661,96
M. tuberculosis	2203,02	1491,32	853,31
M. avium		468,30	449,51

Lze tedy předpokládat, že délka genomického fragmentu determinuje taxonomické rozlišení jednotlivých shluků. Z tohoto hlediska je patrné předpokládané využití studia obsahu *k*-merů v metagenomických datech získaných ze sekvenátorů 3. generace, které produkují delší čtení.

6 ANALÝZA REÁLNÝCH METAGENOMICKÝCH DAT

Vedle aplikace vybraných metod na simulované metagenomické datasety byla analýza testována také na reálných metagenomických datech gastrointestinálního traktu člověka. Zvolené metody byly aplikovány na tento metagenomický dataset za účelem vizualizace dat a diskuse možného taxonomického složení. Přesnost klasifikace byla ověřena mapováním zvolených kontigů s využitím nástroje BLAST oproti referenční databázi.

6.1 Popis dat

Metagenomická data střevního mikrobiomu člověka byla získána z webové stránky projektu zaměřeného na studium enterotypů střevního mikrobiomu a jejich variací u osob různých národností. [66]

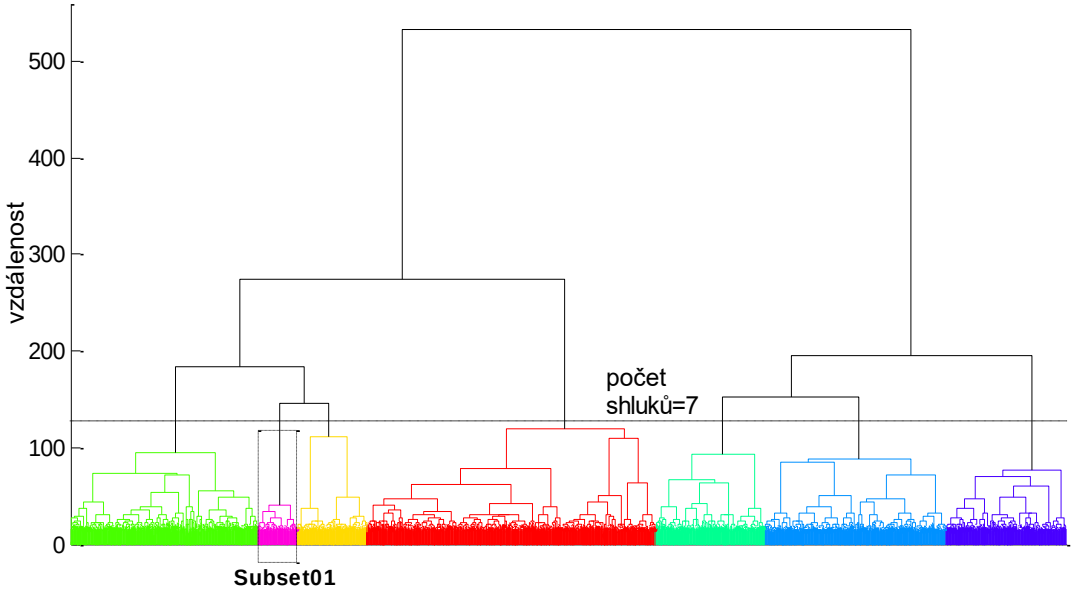
Ze sekce *Assembled contigs* byly zvoleny kontigy vzorku A (A.contigs.fa.gz). Tato metagenomická data byla předzpracována a filtrována za účelem odstranění sekvencí o nízké kvalitě a kontaminací a následně assemblována s využitím nástroje Arachne. Vybrány byly kontigy o délce 1000 až 1500 včetně. Celkem tedy bylo analyzováno 6061 sekvencí s celkovou délkou 7 352 589 bp.

6.2 Vizualizace dat

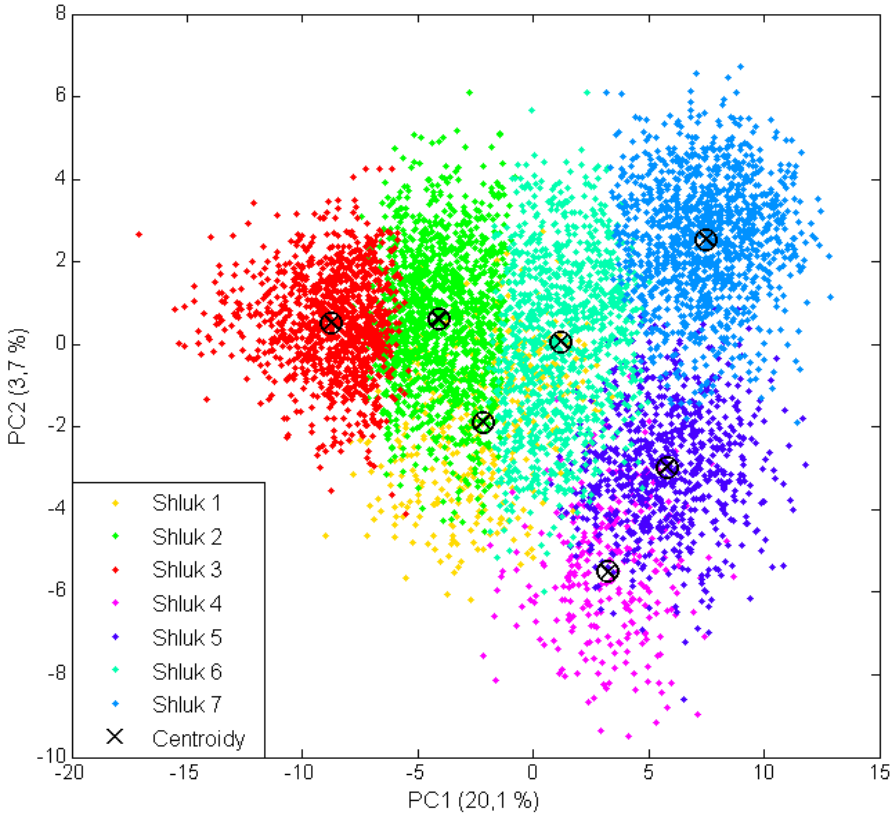
Pro každou ze sekvencí byly vypočítány vektory frekvenčních příznaků. Frekvence symetrizovaných 5-merů byly zvoleny jako příznaky charakterizující sekvence vzhledem k předchozím výsledkům analýzy na simulovaných datech.

Na Obr. 6.1 je vyobrazen dendrogram charakterizující tento dataset. Z hlediska vizualizace lze dobře identifikovat 7 shluků, které byly získány po nastavení prahové vzdálenosti rovno 120.

Stejný počet požadovaných shluků byl nastaven i pro K-means analýzu na příznacích PCA redukovaných sekvenčních příznacích. Na Obr. 6.2 jsou zobrazeny první dvě hlavní komponenty tohoto datasetu a barevně vyznačeny shluky spolu s centroidy.



Obr. 6.1: Dendrogram, frekvence symetrizovaných 5-merů



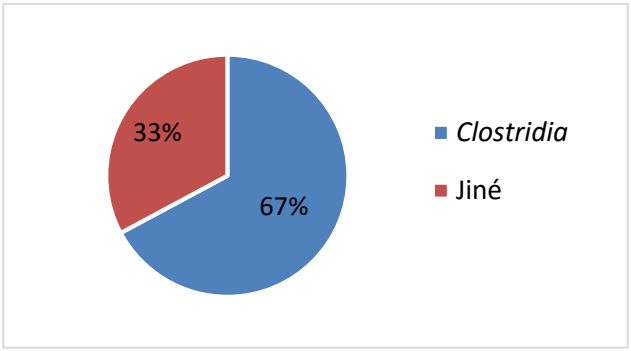
Obr. 6.2: PCA vizualizace dvou hlavních komponent a centroidů vytvořených shluků, frekvence symetrizovaných 5-merů

6.3 Klasifikace

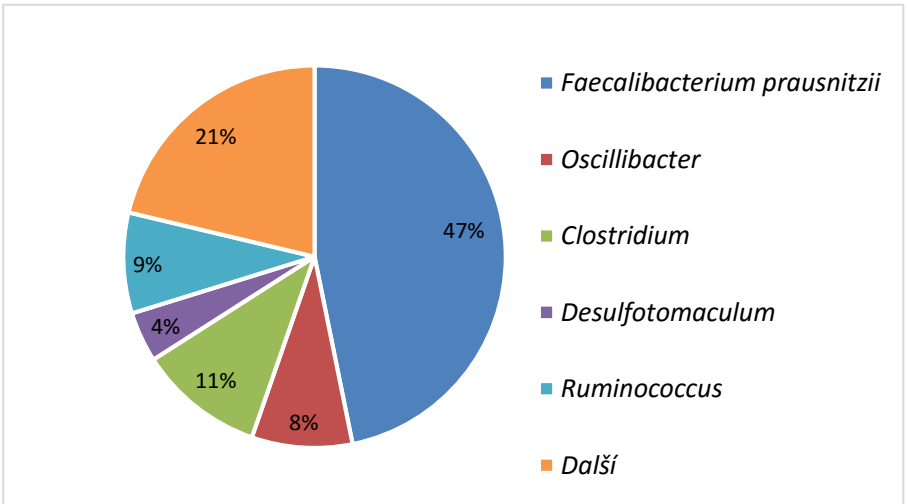
Jako reprezentativní ukázka byl zvolen Subset01, viz Obr. 6.1, který se skládá z 239 kontigů. Velmi podobný soubor lze získat také výběrem shluku 4 podle PCA vizualizace (Obr. 6.2, purpurová). Tyto sekvence byly analyzovány oproti referenční databázi *Nucleotide collection* s využitím nástroje Standard Nucleotide BLAST a algoritmu zarovnání Mega BLAST. [67] Pro analýzu byly ponechány defaultní parametry s výjimkou nastavení maximálního počtu cílových sekvencí rovno 10. Nalezené výsledky byly seřazeny podle maximálního skóre. Jako referenční genom byl zvolen výsledek s maximálním skóre.

Z celkového počtu 239 kontigů bylo možné nalézt 67 významných referencí. Z tohoto počtu se téměř v 70 % jedná o organismy třídy *Clostridia*, jak je možné pozorovat v grafu Obr. 6.3a. Na Obr. 6.3b je pak uvedeno procentuální zastoupení jednotlivých genomů třídy *Clostridia* s výrazným zastoupením taxonu *Faecalibacterium prausnitzii.*

a)



b)



Obr. 6.3: Procentuální zastoupení nalezených referencí, a) celkově, b) třída *Clostridia*

Je patrné, že pro rozsáhlou analýzu a ověření přesnosti klasifikace pro jednotlivé vytvořené shluky chybí dostatek anotovaných záznamů v databázích.

Lze však předpokládat, že vytvořené shluky reprezentují sekvence patřící k taxonomicky příbuzným organismům. Vzhledem k počtu vytvořených shluků pravděpodobně klastrují takovou komunitu na úrovni bakteriální tříd, případně kmenů.

7 ZÁVĚR

Tato práce s názvem *Numerické metody pro klasifikaci metagenomických dat* se zabývá metagenomikou a výpočetními metodami využívanými pro zpracování metagenomu.

Pro pochopení problematiky bylo nezbytné nastudovat charakteristiky metagenomických dat, přístupů sekvenování a využívaných sekvenačních technologií, od kterých se odvíjí i volba analytických nástrojů.

Byla vypracována literární rešerše metod pro klasifikaci organismů na základě taxonomicky specifických četností nukleotidových slov (*k*-merů) v DNA sekvenci. Zahrnuty jsou jednak metody, které již byly publikovány v souvislosti se zpracováním metagenomických dat, jako frekvence *k*-merů a frekvence symetrizovaných *k*-merů, ale také metody používané ve srovnávací genomice.

V rámci praktické části práce byly aplikovány vybrané metody extrakce sekvenčních příznaků založené na četnostech nukleotidových slov. Podrobná analýza byla provedena na sadách simulovaných metagenomických čtení o délce 1000 bp. Data byla klasifikována s využitím hierarchického shlukování v originálním datovém prostoru. Testována byla rovněž redukce dimenzionality dat analýzou hlavních komponent a následné K-means shlukování. Třetím přístupem klasifikace dat je konsensus výsledků hierarchického a K-means shlukování.

Úspěšnost klasifikace byla hodnocena jako tzv. přesnost, tedy počet správně zařazených objektů do odpovídajících tříd ku počtu všech objektů. Ve vybraných případech bylo zařazení objektů do tříd vyhodnoceno s využitím konfuzní matice.

Bylo analyzováno dílčí nastavení hierarchického shlukování. Z testování plyne, že volba vzdálenostní metriky nemá signifikantní vliv na úspěšnost klasifikace. Doporučenou metrikou výpočtu vzdáleností mezi vektory sekvenčních příznaků je standardní euklidovská. Stěžejní je však volba shlukovacího algoritmu, který ovlivňuje vzhled výsledného dendrogramu i rozdělení objektů do shluků a následnou úspěšnost klasifikace. Nejvhodnější metodou pro tuto aplikaci byla vyhodnocena Wardova metoda, která tvoří dobře rozlišitelné shluky. Dále byl zhodnocen vliv délky analyzovaných *k*-merů na přesnost klasifikace pro k=2,…,7. Z tohoto hlediska lze obvykle pozorovat nárůst úspěšnosti pro délku *k*-meru 2 až 5. Pro delší slova nebylo zaznamenáno významné zlepšení v přesnosti klasifikace.

Analýza hlavních komponent byla aplikována pro redukci dimenzionality vektorů sekvenčních příznaků. Vzhledem k proměnlivosti procent variability vyčerpané hlavními komponentami pro různou délku analyzovaných *k*-merů byl na základě analýzy scree

plotu a testování úspěšnosti vyhodnocen doporučený počet PC. Pro K-means shlukování na PCA redukovaných datech byla zvolena euklidovská vzdálenost a 5 replikací náhodného generování poloh centroidů. Z hodnocení úspěšnosti analýzy sekvenčních příznaků pro různou délku *k*-meru byl u tohoto přístupu obvykle pozorován nárůst pro délku slova 2 až 5, případně 6. Z hlediska vizualizace má tento přístup limitaci, jelikož pro žádnou délku analýzy *k*-merů není možné data popsat dvěmi, případně třemi PC.

Vyhodnocení a srovnání aplikovaných metod bylo provedeno pro délku slova *k*=5, jelikož pro obě studované metody shlukování obvykle tato délka vykazuje nejlepší výsledky klasifikace a nejmenší variabilitu mezi jednotlivými metodami. Hierarchické shlukování dosahuje největší přesnosti klasifikace pro sekvenční příznaky frekvence 5-merů, Yang (pořadí seřazených četností 5-merů) a frekvence symetrizovaných 5-merů. Analýza hlavních komponent a K-means shlukování dává nejlepší výsledky pro metody Yang (20 PC) a frekvence symetrizovaných 5-merů (10 PC), a to až 94% úspěšnost klasifikace. Přístup výpočtu konsensu shlukování umožňuje porovnat a ověřit zařazení objektů do tříd s využitím obou metod klasifikace. U žádné z metod výpočtu příznaků však tímto způsobem nebylo dosaženo nejlepší přesnosti klasifikace. Realizované přístupy byly také srovnávány s jedním ze současných algoritmů vizualizace dat t-SNE, na jehož výstup bylo aplikováno K-means shlukování. Tento přístup analýzy dat překonává srovnávané metody v úspěšnosti klasifikace sekvenčních příznaků frekvence 5-merů. Dosahuje srovnatelné úspěšnosti klasifikace jako kombinace PCA analýzy a K-means shlukování pro příznaky frekvence symetrizovaných 5-merů.

Celkově je možné vyhodnotit, že frekvence symetrizovaných 5-merů jsou příznaky, které umožňují dosáhnout největší úspěšnosti klasifikace, tedy více než 94 % pro všechny z hodnocených metod klasifikace. Z hlediska operační náročnosti jsou výhodné také frekvence 5-merů, ze kterých ostatní analyzované metody příznaků vychází.

Pro analýzu reálných metagenomických dat byla zvolena data střevního mikrobiomu člověka. Pro assemblované kontigy o délce 1000-1500 bp byly vypočítány frekvence symetrizovaných 5-merů a následně analyzovány hierarchickým shlukováním a K-means shlukováním. Vybraný shluk byl srovnán s referenční databází pomocí webového nástroje BLAST.

Výše uvedenou analýzou byl ověřen potenciál metod založených na analýze četností nukleotidových slov pro studium metagenomických dat. Výhodou těchto přístupů je, že nevyžadují zarovnání a znalost referenční databáze, které jsou často nekompletní. Rovněž je lze řadit mezi nesupervizované metody nevyžadující téměř žádné předchozí znalosti o daném datasetu. Limitací metod prezentovaných v této práci může být horší rozlišení při hledání shluků v datech obsahujících příbuzné sekvence např. na úrovni bakteriálních

rodů a druhů. Je však patrné, že tyto metody odráží taxonomickou strukturu dat, a proto se jeví jako vhodný nástroj pro předzpracování objemných metagenomických datasetů. Námětem pro další studium v této oblasti je vývoj metod založených na studiu *k*-merů a jejich kombinací, které by umožňovaly charakterizovat metagenomická data i z hlediska nižších taxonomických úrovní.

LITERATURA

[1] HODKINSON, B. a E. GRICE. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in Wound Care*. 2015, **4**(1), s. 50-58 ISSN 2162-1918.

[2] NELSON, K. a B. WHITE. *Metagenomics: Theory, Methods and Applications*. Metagenomics and its applications to the study of the human microbiome. Norfolk: Caister Academic Press, 2010, s. 171-182. ISBN 978-1-904455-54-7.

[3] SCALES, B. a G. HUFFNAGLE. The microbiome in wound repair and tissue fibrosis. *Journal of Pathology* [online]. Chichester, UK: John Wiley, 2013, **229**(2), s. 323-331 [cit. 2016-01-02]. DOI: 10.1002/path.4118. ISSN 00223417.

[4] BASHIR, Y., S. PRADEEP SINGH a B. KUMAR KONWAR. Metagenomics: An Application Based Perspective. *Chinese Journal of Biology* [online]. Hindawi Publishing Corporation, 2014 [cit. 2016-01-02]. DOI: 10.1155/2014/146030.

[5] SeqOmics Biotechnology Ltd. In: *Metagenomics* [online]. 2013 [cit. 2015-11-05]. Dostupné z: <http://www.seqomics.hu/en/contents/metagenomics>.

[6] SUENAGA, H. Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environmental Microbiology*. Oxford, UK: Blackwell Publishing Ltd, 2012, **14**(1), s. 13-22 ISSN 14622912.

[7] METZKER, M. Sequencing technologies — the next generation. *Nature Reviews Genetics*. 2010, **11**(1), s. 31-46 [cit. 2015-12-05]. ISSN 14710056.

[8] KANEHISA, M., S. GOTO, M. FURUMICHI, M. TANABE a M. HIRAKAWA. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research* [online]. 2010, (38) [cit. 2015-12-30]. DOI: 10.1093/nar/gkp896.

[9] TATUSOV, R., N. FEDOROVA, J. JACKSON, a kol. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* [online]. BioMed Central, 2003, **4**, 41-41 [cit. 2015-12-30]. DOI: 10.1186/1471-2105-4-41.

[10] VAN DIJK, E. L., H. AUGER, Y. JASZCZYSZYN a C. THERMES Ten years of next-generation sequencing technology. *Trends in Genetics*. Elsevier Ltd, 2014, **30**(9), s. 418-426. ISSN 01689525.

[11] KOUBKOVÁ, L., B. VOJTĚŠEK a R. VYZULA Sekvenování nové generace a možnosti jeho využití v onkologické praxi. *Klinická onkologie* [online]. Brno, 2014, (27), 61-68 [cit. 2015-12-05]. DOI: 10.14735/amko20141S6.

[12] ANSORGE, W. Next-generation DNA sequencing techniques. *New Biotechnology*. 2009, **25**(4), 195-203. ISSN 18716784

[13] Virtual Genetics Education Centre: Diagrams. In: *University of Leicester* [online]. Leicester, 2012 [cit. 2015-12-05]. Dostupné z: <http://www2.le.ac.uk/departments/genetics/vgec/diagrams/>.

[14] LIU, L., Y. LI, S. LI, a kol. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* [online]. New York: Hindawi Publishing Corporation, 2012 [cit. 2016-05-18]. ISSN 11107243.

[15] DAWN, F., G. GARRITY, T. GRAY et al.. The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*. Nature Publishing Group, 2008, **26**(5), 541 [cit. 2015-12-05]. ISSN 10870156.

[16] KUNIN, V., A. COPELAND, A. LAPIDUS, K. MAVROMATIS a P. HUGENHOLTZ. A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews: MMBR* [online]. 2008, **72**(4), 557 [cit. 2015-12-05]. DOI: 10.1128/MMBR.00009-08.

[17] REDDY, T., A. THOMAS, D. STAMATIS a kol. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic acids research* [online]. [cit. 2015-12-05]. DOI: 10.1093/nar/gku950. Dostupné z: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384021/>.

[18] LEINONEN, R., H. SUGAWARA a M. SHUMWAY The Sequence Read Archive. *Nucleic Acids Research*. 2010, **39**, s. 19-21 [cit. 2015-12-30]. ISSN 0305-1048.

[19] HUSON, D., A. AUCH, J. QI a S. SCHUSTER. MEGAN analysis of metagenomic data. *Genome research*. 2007, **17**(3), s. 377. ISSN 10889051.

[20] MITCHELL, A., F. BUCCHINI, G. COCHRANE a kol. EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research* [online]. 2015 [cit. 2015-12-05]. DOI: 10.1093/nar/gkv1195. Dostupné z: <http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1195>.

[21] KUCZYNSKI, J., Ch. L. LAUBER, W. A. WALTERS, L. W. PARFREY, J. C. CLEMENTE, D. GEVERS a R. KNIGHT. Experimental and analytical tools for studying the human microbiome. *Nature Reviews Genetics*. Nature Publishing Group, 2011, **13**(1), s. 47. ISSN 14710056.

[22] KIM, M., K.-H. LEE, S.-W. YOON, B.-S. KIM, J. CHUN a H. YI. Analytical tools and databases for metagenomics in the next-generation sequencing era. *Genomics & informatics*. 2013, **11**(3). ISSN 1598866X.

[23] SHARPTON, Thomas. An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science* [online]. 2014, (5), 209 [cit. 2015-12-05]. DOI: 10.3389/fpls.2014.00209.

[24] CAPORASO, J. G., J. KUCZYNSKI, J. STOMBAUGH a kol. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. Nature Publishing Group, 2010, **7**(5), s. 335 [cit. 2015-12-30]. ISSN 15487091.

[25] SCHLOSS, P., S. WESTCOTT, T. RYABIN a kol. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and environmental microbiology AEM*. American Society for Microbiology, 2009, **75**(23), s. 7537-7541 ISSN 00992240.

[26] MEYER, F., D. PAARMANN, M. D'SOUZA a kol. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* [online]. BioMed Central, 2008, (9), s. 386-386 [cit. 2015-12-05]. DOI: 10.1186/1471-2105-9-386.

[27] ALTSCHUL, S., W. GISH, W. MILLER, E. MYERS a D. LIPMAN. Basic local alignment search tool. *Journal of Molecular Biology*. 1990, **215**(3), s. 403-410 ISSN 00222836.

[28] COLE, J.R., Q. WANG, E. CARDENAS a kol.. Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research* [online]. 2009, **37**(1), 141-145 [cit. 2015-12-30]. ISSN 03051048.

[29] LOZUPONE, C., M. LLADSER, D. KNIGHTS, J. STOMBAUGH a R. KNIGHT. UniFrac: an effective distance metric for microbial community comparison. *The ISME Journal*. 2010, **5**(2), s.169-172. ISSN 1751-7362.

[30] NIMUSIIMA, J., M. KÖBERL, J. TUMUHAIRWE, J. KUBIRIBA, Ch. STAVER a G. BERG. Transgenic banana plants expressing Xanthomonas wilt resistance genes revealed a stable non-target bacterial colonization structure. *Scientific Reports*. 2015, **5**. DOI: 10.1038/srep18078.

[31] FRANZÉN, O., J. HU, X. BAO, S. ITZKOWITZ, I. PETER a A. BASHIR. Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. *Microbiome*. 2015, **3**(1). ISSN 2049-2618.

[32] GERLACH, W., S. JÜNEMANN, F. TILLE, A. GOESMANN a J. STOYE. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* [online]. BioMed Central, 2009, (10), s. 430-430 [cit. 2015-12-05]. DOI: 10.1186/1471-2105-10-430.

[33] MCHARDY, A., H. G. MARTIN, A. TSIRIGOS, P. HUGENHOLTZ a I. RIGOUTSOS. Accurate phylogenetic classification of DNA fragments based on sequence composition. *Nature Methods* [online]. Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US), 2006, **4**, s. 63-72. - Report Number: LBNL-60414.

[34] SANDBERG, R., G. WINBERG, C. BRÄNDEN, A. KASKE, I. ERNBERG a J. CÖSTER. Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome research* [online]. 2001, **11**(8), s.1404. ISSN 10889051.

[35] TEELING, H., J. WALDMANN, T. LOMBARDOT, M. BAUER a F. GLÖCKNER. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* [online]. BioMed Central, 2004, **5**, 163-163 [cit. 2015-12-30]. DOI: 10.1186/1471-2105-5-163. Dostupné z: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC529438/>.

[36] ABE, T., H. SUGAWARA, M. KINOUCHI, S. KANAYA a T. IKEMURA. Novel Phylogenetic Studies of Genomic Sequence Fragments Derived from Uncultured Microbe Mixtures in Environmental and Clinical Samples. *DNA Research*. Oxford University Press, 2005, **12**(5), 281-290 ISSN 13402838.

[37] LACZNY, C., T. STERNAL, V. PLUGARU a kol. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* [online]. BioMed Central, 2015, **3**(1) [cit. 2015-11-30]. DOI: 10.1186/s40168-014-0066-1. Dostupné z: <http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-014-0066-1>.

[38] LACZNY, C. C., N. PINEL, N. VLASSIS a P. WILMES. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Scientific Reports* [online]. Nature Publishing Group, 2014, (4) [cit. 2015-12-05]. DOI: 10.1038/srep04516.

[39] SUN, S., J. CHEN, W. LI a kol. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic acids research* [online]. 2011, **39**() [cit. 2015-12-30]. DOI: 10.1093/nar/gkq1102.

[40] MARKOWITZ, V., N. IVANOVA, E. SZETO a kol. IMG/M: A data management and analysis system for metagenomes. *Nucleic Acids Research*. Ernest Orlando Lawrence Berkeley NationalLaboratory, Berkeley, CA (US), 2007, **36**. ISSN 03051048.

[41] HUSON, D., S. MITRA, H.-J. RUSCHEWEYH, N. WEBER a S. SCHUSTER Integrative analysis of environmental sequences using MEGAN4. *Genome Research*. 2011, **21**(9), s.1552-1560. DOI: 10.1101/gr.120618.111. ISSN 10889051.

[42] MULLER, J., D. SZKLARCZYK, P. JULIEN a kol. EggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research* [online]. Oxford University Press, Oxford Journals, 2010, **38**(1) [cit. 2015-12-30]. DOI: 10.1605/01.301-0010071006.2010.

[43] Gene Ontology Consortium: going forward. *Nucleic acids research* [online]. 2015, **43** [cit. 2015-12-30]. DOI: 10.1093/nar/gku1179.

[44] TEELING, H., A. MEYERDIERKS, Margarete BAUER, Rudolf AMANN a Frank GLÖCKNER. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*. Oxford, UK: Blackwell Science Ltd, 2004, **6**(9), s. 938-947 [cit. 2015-12-05]. ISSN 14622912.

[45] DICK, G., A. ANDERSSON, B. BAKER, S. SIMMONS, B. THOMAS, A. YELTON a J. BANFIELD. Community-wide analysis of microbial genome sequence signatures. *Genome Biology*. BioMed Central, 2009, **10**(8). ISSN 14656906.

[46] BLAISDELL, B. A Measure of the Similarity of Sets of Sequences not Requiring Sequence Alignment. *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences of the United States of America, 1986, **83**(14), s. 5155-5159. ISSN 00278424.

[47] KARLIN, S., J. MRAZEK a A. CAMPBELL. Compositional biases of bacterial genomes and evolutionary implications. *Journal of Bacteriology*. Washington: American Society for Microbiology, 1997, **179**(12), s. 3899-3913. ISSN 00219193.

[48] 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.

[49] VINGA, S. aj. ALMEIDA. Alignment-free sequence comparison - a review. *Bioinformatics (Oxford, England)*. 2003, **19**(4). ISSN 13674803.

[50] YANG, X. a T. WANG. A novel statistical measure for sequence comparison on the basis of k-word counts. *Journal of Theoretical Biology*. 2013, (318), s. 91-100 ISSN 00225193.

[51] DING, S., Y. LI, X. YANG a T. WANG. A simple k-word interval method for phylogenetic analysis of DNA sequences. *Journal of Theoretical Biology*. 2013, **317**, s. 192-199. ISSN 00225193.

[52] TANG, J., K. HUA, M. CHEN, R. ZHANG a X. XIE. A novel k-word relative measure for sequence comparison. *Computational Biology and Chemistry*. Elsevier Ltd, 2014, **53**, 331-338. ISSN 14769271.

[53] 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.

[54] GISBRECHT, A., B. HAMMER, B. MOKBEL a A. SCZYRBA. Nonlinear Dimensionality Reduction for Cluster Identification in Metagenomic Samples. In: *Information Visualisation (IV), 2013 17th International Conference* [online]. IEEE, 1307, s. 174-179. ISSN 15506037.

[55] QI, X., E. FULLER, Q. WU a C.-Q. ZHANG Numerical Characterization of DNA Sequence Based on Dinucleotides. *The Scientific World Journal* [online]. 2012 [cit. 2016-01-01]. DOI: 10.1100/2012/104269. Dostupné z: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3349307/>.

[56] 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>.

[57] 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>.

[58] 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>.

[59] JARKOVSKÝ, J. a S. LITTNEROVÁ. Vícerozměrné statistické metody: Shluková analýza [přednáška předmětu FSTA Pokročilé statistické metody]. In: *IBA MU* [online]. Brno, 2015 [cit. 2016-05-12]. Dostupné z: <http://www.iba.muni.cz/esf/res/file/bimat-prednasky/vicerozmerne-statisticke-metody/VSM-05.pdf>.

[60] JARKOVSKÝ, J. a S. LITTNEROVÁ. Vícerozměrné statistické metody: Ordinační analýzy – principy redukce dimenzionality [přednáška předmětu FSTA Pokročilé statistické metody]. In: *IBA MU* [online]. Brno, 2015 [cit. 2016-05-12]. Dostupné z: <http://www.iba.muni.cz/esf/res/file/bimat-prednasky/vicerozmerne-statisticke-metody/VSM-06.pdf>.

[61] MELOUN, M. Určení struktury a vazeb v proměnných a objektech. In: *Univerzita Pardubice* [online]. Pardubice, 2016 [cit. 2016-05-12]. Dostupné z: <http://meloun.upce.cz/docs/research/chemometrics/methodology/4cmetody.pdf>.

[62] VAN DER MAATEN, L. a G. HINTON Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*. 2008, **9**, s. 2579-2605. ISSN 15324435.

[63] VAN DER MAATEN, L. *T-SNE* [online]. [cit. 2016-05-12]. Dostupné z: <https://lvdmaaten.github.io/tsne/>.

[64] HAMILTON, H. *Confusion matrix* [online]. In: . Regina: University of Regina, 2012 [cit. 2016-05-12]. Dostupné z: <http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html>.

[65] HIGASHI, S., A. BARRETO, M. CANTãO a A. DE VASCONCELOS Analysis of composition-based metagenomic classification. *BMC Genomics*. [online]. 2012, **13**(5). DOI: 10.1186/1471-2164-13-S5-S1.

[66] ARUMUGAM, M., RAES, J. a kol.: Enterotypes of the human gut microbiome. *EMBL* [online]. Heidlberg, 2011 [cit. 2016-05-13]. Dostupné z: <http://www.bork.embl.de/Docu/Arumugam_et_al_2011/downloads.html>.

[67] ZHANG, Z., S. SCHWARTZ, L. WAGNER a W. MILLER A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology*. 2000, **1-2**(7), 203-214. ISSN 10665277.

SEZNAM SYMBOLŮ A ZKRATEK

NGS	*Next-Generation Sequencing,* sekvenovací technologie nové generace
16S rRNA	gen kódující RNA malé podjednotky
bp	z angl. *base pair,* pár bazí
OTU	z angl. *operational taxonomic unit,* definice skupiny organismů
BLAST	*Basic Local Alignment Search Tool*, algoritmus pro lokální zarovnání
k-mer	nukleotidové slovo
K-means	metoda k-průměrů
PCA	z angl. *Principal Component Analysis,* analýza hlavních komponent
PC	hlavní komponenta
t-SNE	*t-distributed Stochastic Neighbor Embedding*, metoda redukce dimenzí
scree plot	indexový graf úpatí
kontig	řetězec po sobě jdoucích fragmentů DNA vytvářejících spojitou sekvenci

SEZNAM PŘÍLOH

A CD

Obsah CD

- Elektronická verze práce Tereza_Vaneckova_DP.pdf
- Skripty a funkce pro analýzu dat
- Data