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

BIPARTITNÍ GRAFY PRO ANALÝZU MIKROBIOMŮ
BIPARTITE GRAPHS FOR MICROBIOME ANALYSIS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR
Bc. Marcela Šafárová

VEDOUCÍ PRÁCE
SUPERVISOR
Mgr. Ing. Karel Sedlář

BRNO 2017

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. Marcela Šafárová ***ID:*** 147499
Ročník: 2 ***Akademický rok:*** 2016/17

NÁZEV TÉMATU:

Bipartitní grafy pro analýzu mikrobiomů

POKYNY PRO VYPRACOVÁNÍ:

1) Zpracujte literární rešerši o současných aplikacích metagenomických studií a možnostech vizualizace metagenomických dat. Zaměřte se na vizualizaci pomocí grafů/sítí. 2) Popište nejpoužívanější sekvenační technologie, metodu sekvenování amplikonů 16S rRNA a princip kvantitativního určení OTU ve vzorku. 3) Ve vhodně zvoleném jazyce vytvořte funkce pro konstrukci bipartitního grafu včetně předzpracování a čištění dat z kvantitativních metagenomických dat (tzv. OTU table). 4) Balíček rozšiřte o vhodné nástroje umožňující analýzu vytvořených grafů, např. detekci komunit, barvení vrcholů apod. 5) Nástroj doplňte o možnost exportu vytvořených grafů do vhodného formátu, který by byl dále použitelný se software na vizualizaci grafů (Gephi, Cytoscape). 6) Proveďte diskusi a zhodnocení výsledků.

DOPORUČENÁ LITERATURA:

[1] SEDLAR, Karel, Petra VIDENSKA, Helena SKUTKOVA, Ivan RYCHLIK a Ivo PROVAZNIK. Bipartite Graphs for Visualization Analysis of Microbiome Data. Evolutionary Bioinformatics. 12 (Suppl 1), s. 17-23.

[2] BITTNER, Lucie et al. Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. Biology Direct. 2010, 5 (1), s. 47.

Termín zadání: 6.2.2017 ***Termín odevzdání:*** 19.5.2017

Vedoucí práce: Mgr. Ing. Karel Sedlář
Konzultant:

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

UPOZORNĚNÍ:
Autor diplomové práce nesmí při vytváření diplomové práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

Abstrakt

Mikroorganismy se vyskytují ve velkém množství prakticky všude kolem nás. Některé přežívají dokonce i v našem těle a jsou nutné pro správné fungování organismu. Studium mikrobiálních společenstev na základě souboru jejich genetické informace se stalo velmi populární s rozvojem nových technologií umožňujících snadné čtení DNA či RNA. Klíčovou úlohou těchto studií je obvykle charakterizovat významné mikrobiální vzory prostředí. V současné době využívané vizualizační nástroje však mají pro takové analýzy mnoho nedostatků. Předmětem této práce je návrh R/Bioconductor balíčku pro tvorbu bipartitních grafů z mikrobiálních dat, které mají pro analýzu mikrobiomů mnoho výhod. Benefity této vizualizační metody jsou dále předvedeny na analýze hlavních parametrů ovlivňujících počítačové zpracování mikrobiálních dat.

Klíčová slova

studium mikrobiomů; vizualizace dat; bipartitní graf; detekce komunit; OTU picking; QIIME

Abstract

Microorganisms are all around us. Some of them even live in our body and are essential for our healthy being. Study of microbial communities based on their genetic content has become very popular with the development of new technologies, which enable easy reading of DNA or RNA. The key role of these studies is usually to characterize significant microbial patterns of an environment. However, currently used visualization tools have many drawbacks for such analyses. The subject of this thesis is to design a R/Bioconductor package for simple creation of bipartite graphs from microbial data. This type of visualization brings many advantages for microbiome analysis. Benefits of bipartite graphs are further demonstrated by analysis of main parameters affecting computer processing of microbial data.

Keywords

study of microbiome; data visualization; bipartite graphs; community detection; OTU picking; QIIME

ŠAFÁROVÁ, Marcela. *Bipartitní grafy pro analýzu mikrobiomů:* diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2016. 70 s. Vedoucí práce Mgr. Ing. Karel Sedlář.

Prohlášení

Prohlašuji, že svou diplomovou práci na téma Bipartitní grafy pro analýzu mikrobiomů jsem vypracovala samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením tohoto projektu jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 1. května 2017 podpis autora

Poděkování

Ráda bych poděkovala především vedoucímu mé diplomové práce Mgr. Ing. Karlu Sedláři za jeho velkou podporu, cenné rady a čas, který mi věnoval.

Děkuji také svým rodičům za vzdělání, které mi umožnili, a za slova povzbuzení během celého mého studia.

V Brně dne 1. května 2017 podpis autora

Obsah

Úvod 7
1 Metagenomika 8
2 Vývoj sekvenátorů 10
2.1 První generace sekvenování 10
2.2 Druhá generace sekvenování 12
2.3 Třetí generace sekvenování 15
3 Zpracování dat 17
3.1 Shotgun sekvenování 17
3.2 Cílené amplikonové sekvenování genů 18
3.3 Multiplexace dat 20
3.4 Operační taxonomická jednotka a shlukování 20
3.5 Referenční databáze 23
3.6 OTU tabulky 23
4 QIIME – Quantitative Insights Into Microbial Ecology 24
5 Vizualizace dat 26
5.1 Jednoduché bodové 2D grafy 26
5.2 Nástroje redukce dimenzí 28
5.3 Grafické znázornění rozložení celku 31
5.4 Stromové struktury 33
5.5 Sítě 34
6 Praktická realizace bipartitních grafů 37
6.1 Zpracování kvantitativních mikrobiálních dat 37
6.2 Algoritmy detekce komunit 37
6.3 Charakteristiky grafu 42
6.4 Doplňkové funkce pro práci s grafy 42
7 Analýza 43
7.1 Analýza funkcí balíčku bipartiteOTU 43
7.2 Analýza algoritmů detekce komunit 53
7.3 Analýza procesu OTU picking 56
Závěr 63
Literatura 64
Seznam zkratek 69
Seznam příloh 70

Seznam obrázků

Obr. 1: Nejpoužívanější metody studia genomu. [49] ... 8
Obr. 2: Sekvenování Sangerovou metodou ... 11
Obr. 3: Roche 454 Pyrosequencing – příprava vzorků před zahájením sekvenace. [30] ... 13
Obr. 4: Můstková PCR. [33] ... 14
Obr. 5: Schéma sekvenování ligací využívané sekvenátory SOLiD. [31] ... 15
Obr. 6: Replikace vláken DNA v jamkách zero-mode waveguides. [57] ... 16
Obr. 7: Oxford Nanopore – patrné jsou změny proudu při průchodu molekul. [32] ... 16
Obr. 8: Shotgun metoda sekvenování. ... 18
Obr. 9: Variabilní regiony 16S ribozomální RNA. [58] ... 19
Obr. 10: konzervativní (zelené) a variabilní (šedé) regiony 16S rRNA genu. [35] ... 20
Obr. 11: Přehled OTU picking metod. ... 21
Obr. 12: OTU tabulka ... 23
Obr. 13: Příklad formátování mapping file. ... 24
Obr. 14: Příklad formátování demultiplexovaného FASTA souboru. ... 25
Obr. 15: Teplotní mapa. [60] ... 26
Obr. 16: Rozmanitost lidského mikrobiomu. [38] ... 27
Obr. 17: Vývoj lidského mikrobiomu. [38] ... 28
Obr. 18: Principal coordinate analysis. [40] ... 29
Obr. 19: Vizualizace metagenomických dat pomocí t-SNE. [61] ... 30
Obr. 20: Porovnání Gaussova a Studentova rozdělení ... 31
Obr. 21: Sloupcový graf znázorňující zastoupení mikrobiálních kmenů ... 32
Obr. 22: Mikrobiální rody v různých částech lidského těla. [44] ... 32
Obr. 23: Fylogenetický strom s prstenci vytvořený pomocí nástroje GraPhlAn. [47] ... 33
Obr. 24: Bayesovská síť. [29]. ... 34
Obr. 25: Bipartitní graf – lidský mikrobiom. ... 35
Obr. 26: Princip detekce komunit pomocí metody propagace značky. [63] ... 39
Obr. 27: Princip detekce komunit pomocí algoritmu rychlého rozvíjení. [64] ... 40
Obr. 28: Detekované taxonomické říše ... 44
Obr. 29: Detekované taxonomické kmeny ... 45
Obr. 30: Detekované taxonomické druhy ... 46
Obr. 31: Detekované čeledi předzpracované porovnáním s prahovou hodnotou 100 ... 48
Obr. 32: Bipartitní graf porovnávající dodavatele. ... 50
Obr. 33: Detekované organismy spadající do třídy *Bacilli* ... 51
Obr. 34: Krabicový graf rozložení délky sekvencí ... 52
Obr. 35: Výsledek neváhované detekce komunit metodou optimálního rozložení ... 54
Obr. 36: Výsledek váhované detekce komunit optimálním rozložením. ... 55
Obr. 37: OTU picking: Nalezené taxonomické čeledi ... 57

Obr. 38: Práh podobnosti: Nalezené taxonomické čeledi 58
Obr. 39: Graf srovnávající abundanci OTU při různém nastavení prahové hodnoty..... 60
Obr. 40: Taxonomické čeledi nalezené použitím 3 odlišných databází. 61

Seznam tabulek

Tabulka 1: Transformace popisků 42
Tabulka 2: Výsledky redukce dat na základě taxonomického určení. 43
Tabulka 3: Výsledky redukce dat nastavením prahové hodnoty 47
Tabulka 4: Dostupná metadata 49
Tabulka 5: Parametry grafu získaného souborem tří redukčních funkcí 51
Tabulka 6: Parametry grafu zaměřeného na třídu Bacilli 52
Tabulka 7: Rozložení délky analyzovaných sekvencí 52
Tabulka 8: Výsledky detekce komunit mezi OTU 53
Tabulka 9: Výsledky detekce komunit mezi vzorky 55
Tabulka 10: Srovnání výsledků získaných odlišnými OTU picking metodami 56
Tabulka 11: Maximální počet odlišných bází pro přiřazení do shluku 57
Tabulka 12: Práh podobnosti: Rozdíly v zaznamenané četnosti jednotlivých organismů 59
Tabulka 13: Databáze: Rozdíly v zaznamenané četnosti jednotlivých organismů 62

Úvod

Mikroorganismy hrají nezastupitelnou roli v našich životech. Ačkoliv byly vždy součástí života lidí, až dnes začínáme odhalovat komplexitu jejich společenstev a způsob, jakým dokážou ovlivňovat náš svět. Ukazuje se, že i malé změny v diverzitě mikrobiálních společenstev mohou mít velký vliv na zdraví jedince či životní prostředí. Proto společnou snahou velké části současných studií je odhalovat vzájemné vazby mezi mikroorganismy uvnitř společenstev, a také vystopovat zajímavé asociace mezi konkrétními změnami v mikrobiálním společenství a jejich vlivem na prostředí.

Zajímavým trendem se stalo studium mikrobiálních společenstev přímým sekvenováním jejich genetické informace. Tento poměrně nový typ studia mikroorganismů se vyvinul díky značnému technologickému pokroku na poli sekvenátorů. Přestože studium mikrobiálních společenstev již dnes nenaráží na problém sběru dat, je analýza takového společenství náročná v důsledku nedostatku bioinformatických nástrojů pro jejich vizualizaci a analýzu.

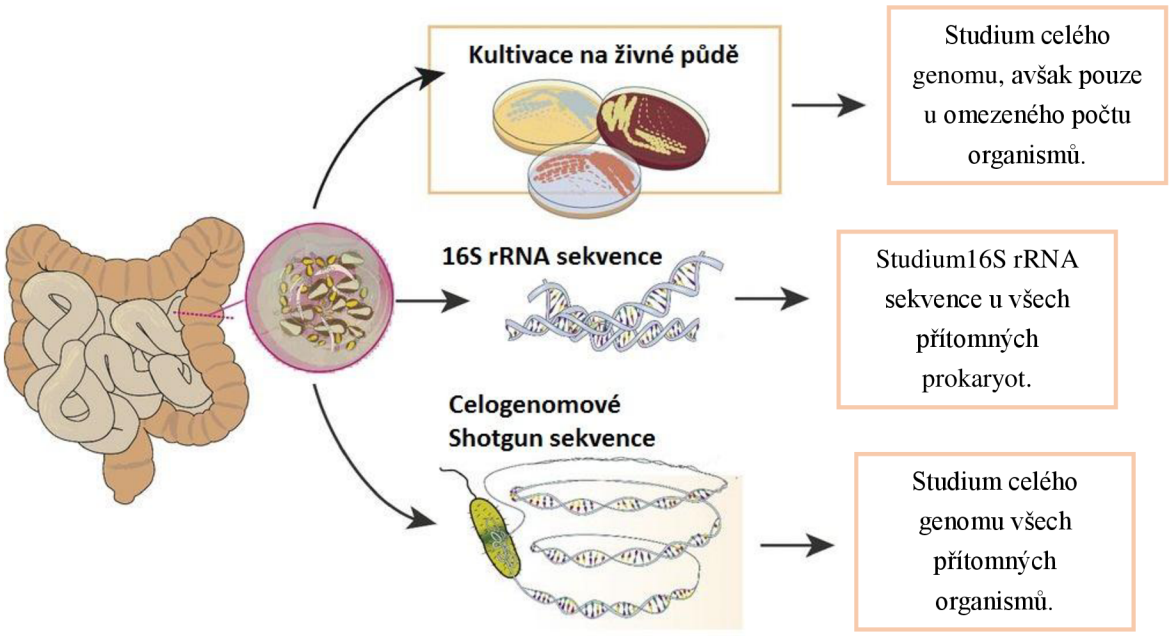
Předmětem této práce je návrh R/Bioconductor balíčku pro snadnou tvorbu bipartitních grafů z mikrobiálních dat. Právě tato vizualizační technika může mít pro analýzu společenstev mikroorganismů mnoho benefitů - možnost detekce důležitých komunit, přímé vykreslení taxonů, propojení vzorků z různých prostředí apod. Jednou z hlavních úloh při tvorbě bipartitního grafu z takto obsáhlého datasetu je pak vhodná volba redukčních technik, která ve výsledku umožní zvýšit výtěžnost analýzy.

Diplomová práce se dále zabývá analýzou jednoho z hlavních kroků zpracování genomických mikrobiálních dat, a to sice kroku zvaného OTU picking. Jsou porovnávány tři hlavní OTU picking přístupy – *de novo*, closed reference a open reference. Zároveň je také porovnáván vliv vstupních parametrů jako volba referenční databáze a práh podobnosti na dosažené výsledky. Pro tyto analýzy byl využit právě navržený R/Bioconductor balíček.

1 Metagenomika

Rozvoj sekvenátorů na přelomu tisíciletí způsobil rozmach nejen v sekvenování genomů, ale také v sekvenování celých metagenomů. Ukázalo se, že pouhým sekvenováním kultivovaných vzorků mikroorganismů dochází ke značnému podhodnocení mikrobiální diverzity, neboť mnoho mikroorganismů běžně se vyskytujících v přírodě nelze běžnými kultivačními metodami pěstovat [3]. Podstatou metagenomiky je studium veškerého genetického materiálu přítomného přímo ve vzorku prostředí bez nutnosti kultivace. Vzorkem prostředí může být nejen půda, voda, sliny, ale také například střeva, pokožka apod.

Obr. 1 ukazuje tři nejpoužívanější metody studia genomu. Kultivace na živné půdě sice umožňuje studium celého genomu, avšak pouze u omezeného množství organismů. Pro komplexní popis mikrobiálních společenstev se tak využívá metod, které nevyžadují předcházející kultivaci *in vitro* – amplikonového sekvenování genů (nejčastěji 16S rRNA genu) nebo shotgun sekvenování celého metagenomu. Oba přístupy jsou detailněji popsány v kapitolách 3.1 a 3.2.



Obr. 1: Nejpoužívanější metody studia genomu. [49]

Izolace genetického materiálu mikroorganismů přímo z jejich přirozeného prostředí vedla k objevu mnoha dosud neobjevených rodových linií. Metagenomika našla kromě mikrobiologie uplatnění rovněž v medicíně a veterinární medicíně, potravinářství, ekologii a biotechnologiích. [4] Mezi zajímavé projekty patří například výzkum střevní mikroflóry a její vliv na zdraví [5],[6],[7], výzkum mikrobiální diverzity v průduškách [8] nebo výzkum výskytu antibiotické rezistence napříč mikrobiomy [9]. V ekologii a biotechnologiích je pak studium mikrobiomu využíváno například pro analýzu biodegradačních schopností bakterií [10].

Rostoucí počet studií mikrobiomu dokazuje, že sekvenování genomů se stalo poměrně dostupným a výkonným. Avšak s nárůstem získávaných mikrobiálních dat roste také poptávka po nových nástrojích umožňujících jejich zpracování, analýzu a vizualizaci.

2 Vývoj sekvenátorů

Již od objevu struktury a funkce DNA se pozornost vědecké obce upírala na možnosti rozluštění pořadí nukleových bází v sekvencích DNA. Dlouhá léta zůstávalo sekvenování DNA velmi obtížné. Průlom nastal představením ve své době revoluční metody sekvenování pomocí dideoxynukleotidů. Metoda byla poprvé publikována F. Sangerem, S. Nicklenem a A. R. Coulsem v roce 1977, stala se velmi populární a umožnila jedno z prvních aplikovatelných čtení DNA. O této éře prvních sekvenátorů se dnes hovoří jako o první generaci sekvenování (viz kapitola 2.1).

Avšak problémem sekvenování stále zůstávala jeho časová a finanční náročnost. To se změnilo s nástupem druhé generace sekvenování, které je také nazývané sekvenování nové generace. Paralelní sekvenování vyvolalo v roce 2007 novou éru, ve které začaly ceny sekvenování prudce klesat [1]. Dnes je na trhu celá řada sekvenátorů lišících se používanou chemií, cenou a časem zpracování, výstupní kapacitou a dalšími parametry. Experimentuje se také s přímým sekvenováním jediné molekuly DNA bez nutnosti předešlé amplifikace – tzv. třetí generace sekvenování. [2]

2.1 První generace sekvenování

Mezi nejznámější sekvenační metody první generace patří Sangerova metoda a metoda Maxam-Gilbertova. Obě využívají k odečtení pořadí bází elektroforézu, liší se však v předešlém zpracování vzorku.

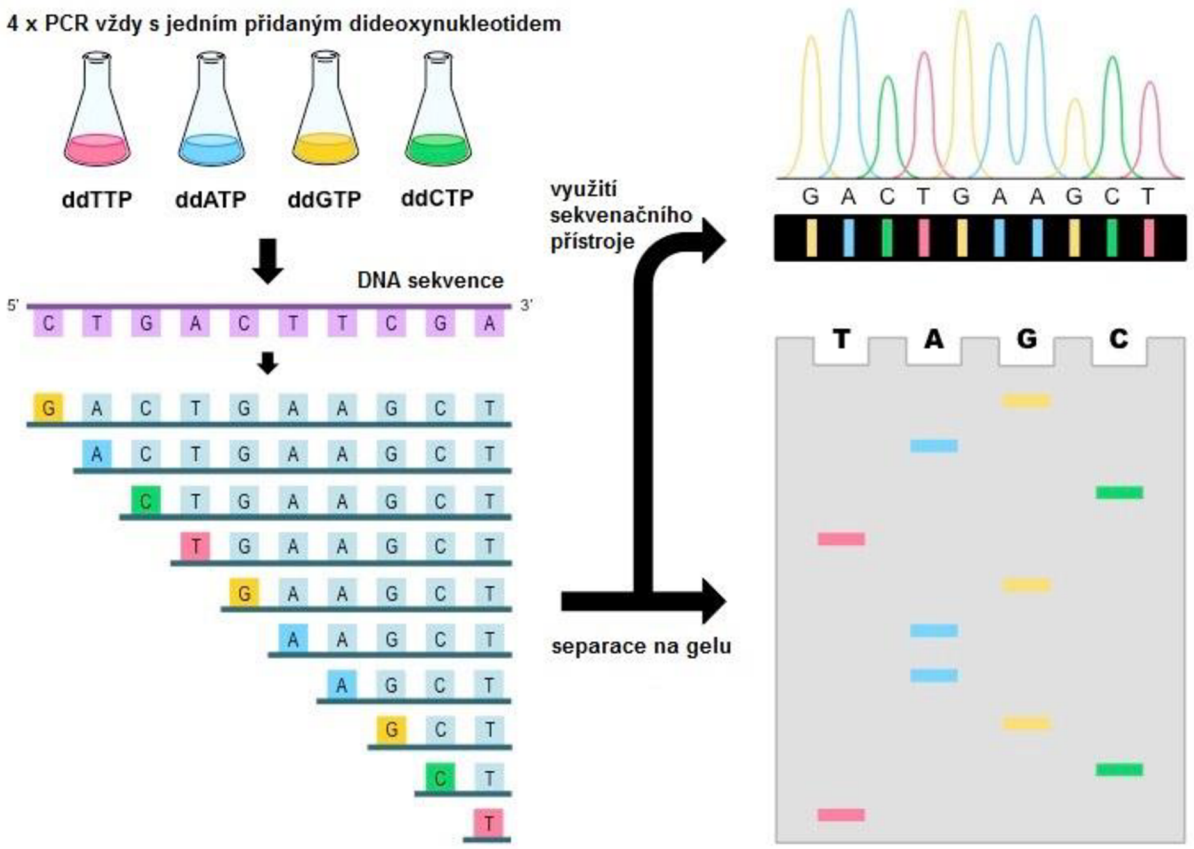
Maxam-Gilbertova metoda

Maxam-Gilbertova metoda pracuje s naštěpenými fragmenty DNA. Radioaktivně značený vzorek je napipetován do čtyř nádob se specifickými chemikáliemi, které štěpí sekvenci DNA v místě s určitou nukleovou bází. Konkrétně se pro rozkladné reakce využívá dimethylsuflátu, hydrazinu a piperidinu. Dimethylsulfát metyluje purinové báze, jež jsou pak nestabilní a snadno se rozpadají zahřátím při neutrálním pH. Pyrimidiny zase dobře reagují s hydrazinem, který štěpí glykosidickou vazbu mezi deoxyribózou a pyrimidinovou bází. Funkcí piperidinu je katalyzovat rozklad fosfodiesterových vazeb v místech, kde chybí purinová či pyrimidinová báze. Vzhledem k tomu, že metylovaný adenin je méně stabilní než metylovaný guanin, je možné pomocí dimethylsulfátu a piperidinu v kyselém prostředí rozložit pouze adeninové báze. Podobně pak hydrazin a piperidin v 1,5 M NaCl rozkládá pouze báze cytosinové. [17]

Těchto chemických reakcí je využito pro cílené štěpení DNA ve čtyřech různých nádobách. Vznikají různě dlouhé sekvence, které mohou být separovány pomocí gelové elektroforézy a využity pro odečtení pořadí nukleotidů.

Sangerova metoda

Sangerova metoda naproti tomu využívá čtyř speciálně připravených inhibitorů k terminaci elongace nově syntetizovaného řetězce DNA (viz obr. 2). Jedná se o dideoxyribonukleotidy (ddNTP) lišící se od příslušných deoxyribonukleotidů (dNTP) chybějící OH skupinou na uhlíku č. 3 cyklické cukerné složky. Absence této OH skupiny zabraňuje navázání dalších ddNTP či dNTP a další syntéza DNA tak již není možná. Opět vznikají různě dlouhé sekvence DNA, které mohou být obdobným způsobem jako u Maxam-Gilbertovy metody odečteny pomocí elektroforézy. [12]



Obr. 2: Sekvenování Sangerovou metodou

Ačkoliv má první generace sekvenátorů v historii genomiky své nezpochybnitelné místo, pro potřeby studia mikrobiomu je rozhodně lepší sáhnout po sekvenačních metodách druhé nebo dokonce třetí generace. Hlavním nedostatkem sekvenátorů první generace je nedostatečná výstupní kapacita. Sekvenovat celé genomy pomocí těchto sekvenátorů je časově i finančně příliš náročné. Dnes se zautomatizovaná verze Sangerovy metody využívající fluorescenčních barviv a kapilární elektroforézy využívá pro svou nízkou chybovost především k charakterizaci sporných oblastí ve výstupech novějších metod.

2.2 Druhá generace sekvenování

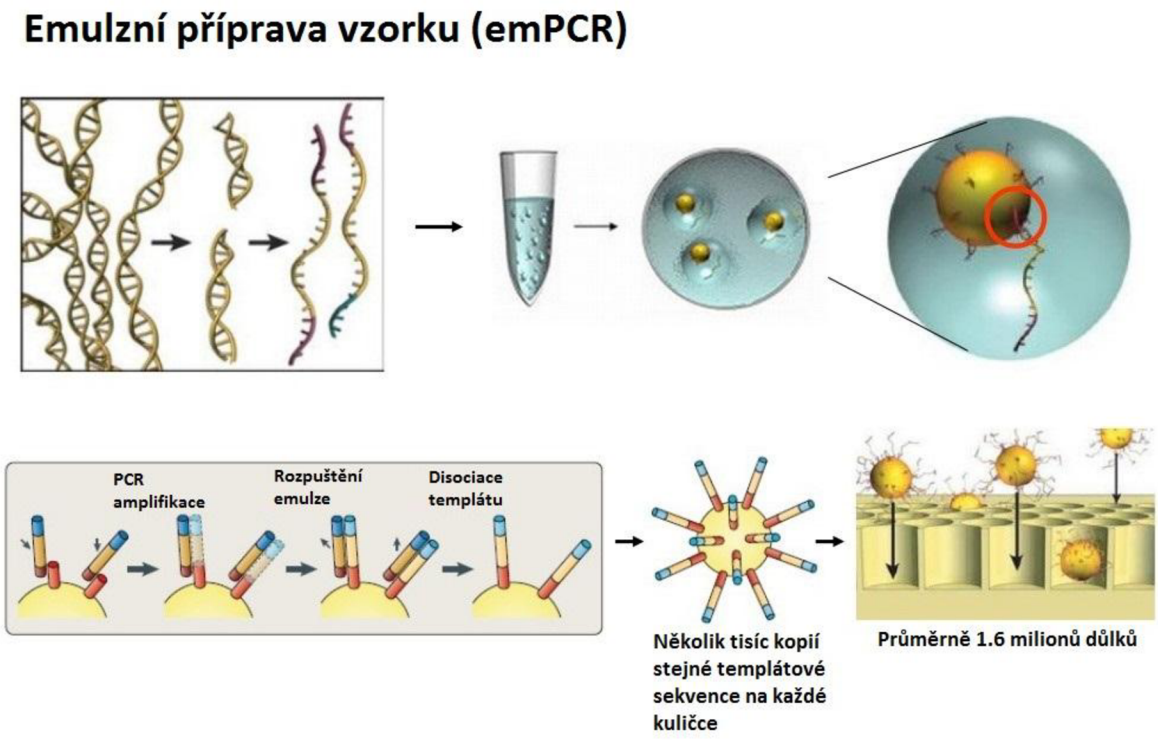
Na začátku 21. století se výroby sekvenátorů chopily soukromé firmy jako například Roche, Illumina nebo Life Technologies. Komerční firmy přišly s rozmanitými přístupy k sekvenování, avšak je možné vysledovat hlavní trend druhé generace, a to sice zavedení souběžného sekvenování více molekul DNA naráz. Tento faktor se významně podílel na snížení ceny a času potřebného ke zpracování vzorků. Druhé generaci sekvenování se také říká sekvenování nové generace (NGS z anglického „Next Generation Sequencing“).

454 pyrosekvenování

Mezi dříve hojně používané sekvenátory patří například 454 pyrosekvenátory dodávané firmou Roche. Přestože se firma rozhodla k roku 2016 ukončit podporu všech platforem těchto sekvenátorů, považuji za důležité 454 pyrosekvenace zmínit, neboť data sesbíraná na těchto přístrojích jsou stále hojně využívána.

454 pyrosekvenace pomocí systému FLX^{TM} může být rozdělena do tří kroků:

1) Příprava knihovny DNA fragmentů – Dvoušroubovice DNA je rozfragmentována na menší fragmenty délky cca 400 – 600bp. V další fázi jsou ligací připojeny adaptory na oba konce DNA fragmentů a dvoušroubovice DNA je rozdělena na jednotlivá vlákna.
2) Navázání DNA vzorku na nosič – Nosičem jsou speciální kuličky, na které je pomocí adaptéru imobilizován vždy jeden fragment DNA. Po přidání PCR reagencií a emulzního oleje ke kuličkám dochází k tzv. emulzní PCR, jejímž výsledkem je přibližně 10 milionů identických kopií DNA navázaných na každé z kuliček. Tato amplifikace umožní významné zesílení signálu získaného ve třetím kroku (viz obr. 3).
3) Sekvenování – Kuličky s navázanou a amplifikovanou DNA jsou umístěny do PicoTiter^{TM} destiček, které jsou pokryty důlky o velikosti v řádu mikrometrů. Každý z důlku je tak vyplněn maximálně jednou kuličkou. Ke každé kuličce jsou také přidány enzymy podílející se na následné syntéze komplementárního řetězce k navázané jednovláknové DNA. Celá destička je poté postupně omývána reagenciemi a nukleotidy A, C, G a T. Při začlenění některého z nukleotidů je spuštěn sled chemických reakcí vedoucích k uvolnění světelné energie. Světelné záblesky jsou detekovány CCD kamerou z každého důlku zvlášť, a může tak být sekvenováno několik vláken DNA naráz. [18]

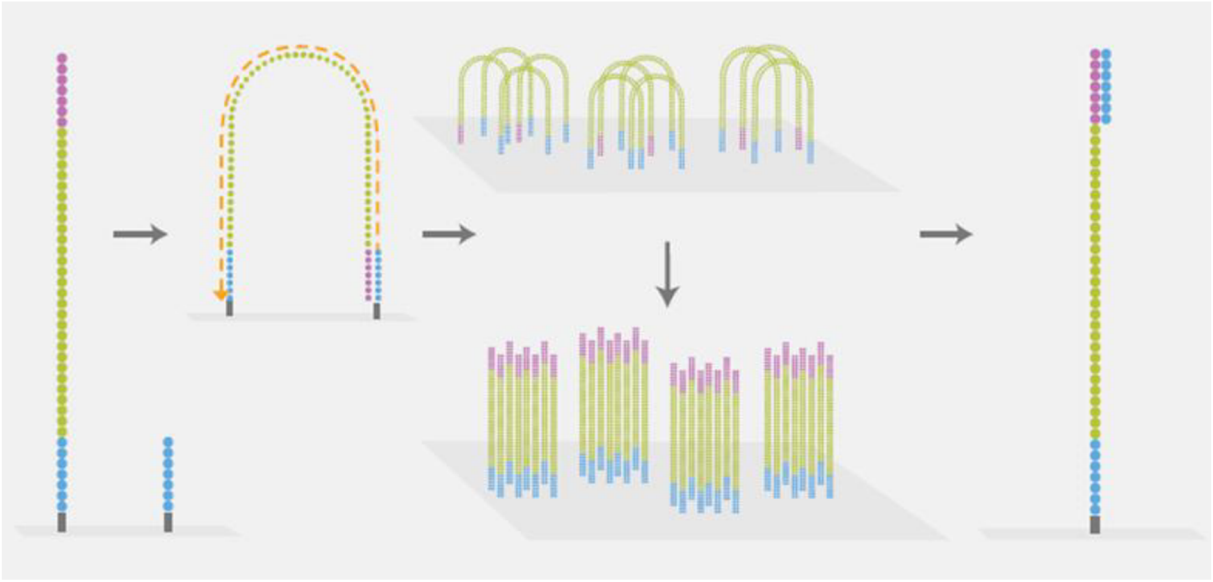


Obr. 3: Roche 454 Pyrosequencing – příprava vzorků před zahájením sekvenace. [30]

Hlavní výhodou 454 sekvenátoru oproti jiným NGS přístrojům je jeho schopnost nabídnout sekvenování o délce čtení až 1000bp, tedy srovnatelnou s délkou čtení Sangerova sekvenování [19]. Délka čtení je ovšem vykoupena vyšší chybovostí při čtení homopolymerů (oblasti složené z jediného, mnohokrát se opakujícího nukleotidu), a také vyšší cenou chemikálií oproti ostatním NGS přístrojům.

Sekvenátory Illumina

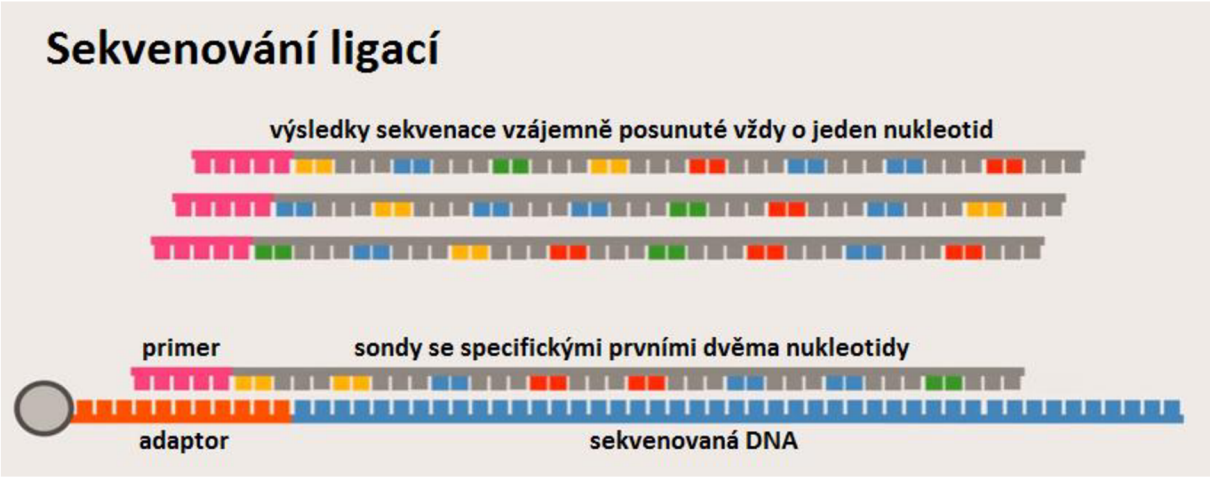
Sekvenování syntézou využívají také sekvenátory firmy Illumina. Příprava knihovny DNA je podobná jako u 454 pyrosekvenátoru – molekula DNA je rozfragmentována a ligací jsou navázány adaptory nutné pro PCR amplifikaci a sekvenování. Amplifikace probíhá na destičce s vazebnými místy pro adaptory. Molekuly DNA jsou zde amplifikovány metodou můstkové PCR a vznikají shluky kopií původních fragmentů (viz obr. 4). Následnou replikací DNA a postupným začleňováním fluorescenčně značených komplementárních nukleotidů do nově syntetizovaného vlákna DNA jsou fragmenty sekvenovány. Používané nukleotidy mají inaktivovanou 3‘-OH skupinu a blokují včleňování dalších nukleotidů, takže syntéza vlákna DNA může být regulována v závislosti na rychlosti čtení detektoru. Tato řízená replikace DNA umožňuje snížení chybovosti při sekvenování homopolymerů. Nevýhodou sekvenátorů Illumina je jejich poměrně nízká délka čtení v důsledku zvyšujícího se šumu s délkou sekvenovaného fragmentu. [34]



Obr. 4: Můstková PCR – DNA fragmenty jsou navázány adaptory k Flow Cell, kde probíhá amplifikace. Pro každý DNA fragment vzniká shluk přibližně jednoho miliónu identických kopií DNA. [33]

Sekvenátory SOLiD

Sekvenátory SOLiD (z anglického „Sequencing by Oligo Ligation and Detection“) firmy Thermo Fisher Scientific (dříve Life Technologies) přistupují k procesu čtení DNA jinak. Příprava DNA knihovny a amplifikace je sice podobná přípravě knihovny u 454 pyrosekvenování, avšak proces sekvenování není závislý na DNA polymeráze a začleňování nukleotidů. Místo toho se používají krátké fluorescenčně značené oligonukleotidové sondy, které hybridizují k DNA. Sondy jsou osm bází dlouhé, definovány prvními dvěma nukleotidy. Zbylých šest bází je degenerovaných, tzn. mají schopnost párovat se s jakýmikoliv nukleotidy templátové sekvence. Pokud se shodují první dva nukleotidy sondy s nukleotidy templátu, může proběhnout hybridizace a následná ligace. Nenavázané sondy jsou vymyty a je detekováno fluorescenční záření specifické pro jednotlivé typy sond. Poslední tři báze sondy včetně fluorescenčního značení jsou odstraněny a hybridizace s ligací pokračují podél vlákna DNA dál. Celkem existuje 16 možných kombinací dinukleotidů, ale používá se jen čtyř barev fluorescenčního záření. To znamená, že pro celkové osekvenování DNA musí celý proces proběhnout alespoň 5krát s primerem navázaným na pozici o jedna menší než předešlý primer (viz obr. 5). Tato metoda bohužel umožňuje pouze sekvenování krátkých čtení s maximální délkou cca 120 bází. Co ovšem dělá sekvenátory SOLiD výjimečné je jejich vysoká přesnost, díky které jsou vhodnou volbou pro detekci jednonukleotidových polymorfismů a inzercí či delecí. [20]

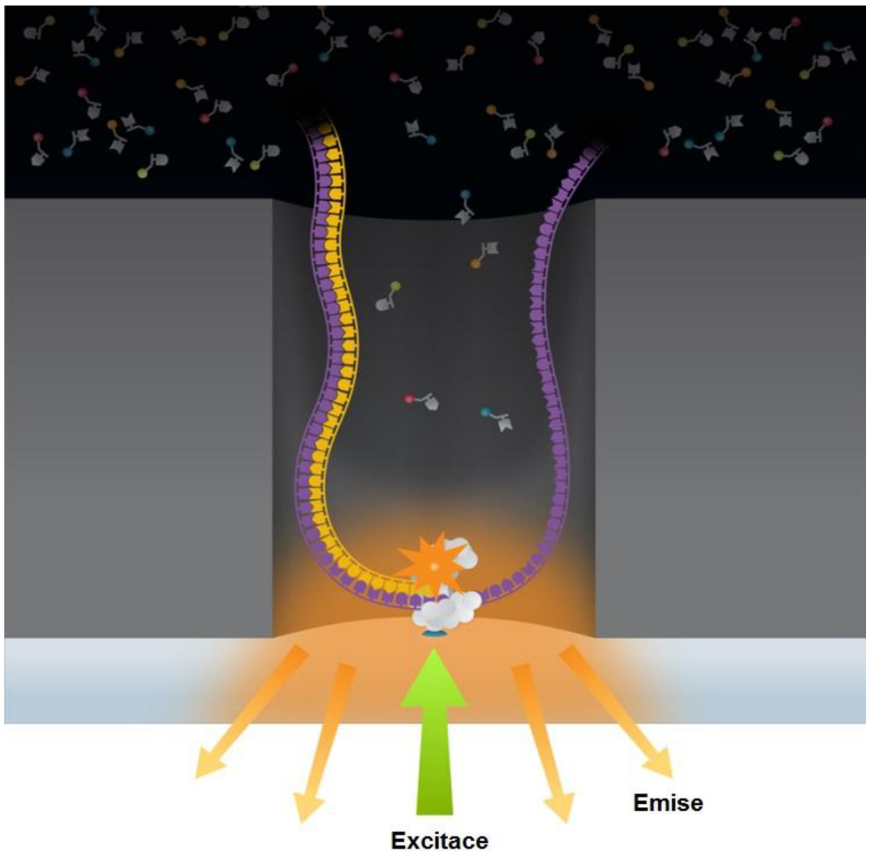


Obr. 5: Schéma sekvenování ligací využívané sekvenátory SOLiD. [31]

2.3 Třetí generace sekvenování

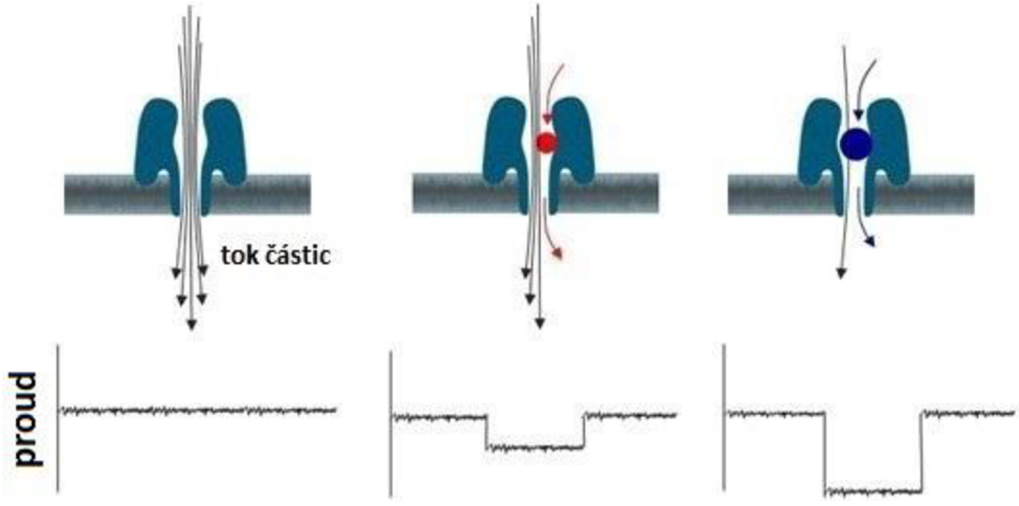
Po druhé generaci, zaměřující se na paralelizaci dat, přichází třetí generace snažící se uplatnit přístup sekvenování pouze jediné molekuly DNA (SMS z anglického „Single molecule sequencing“). Společným znakem je vynechání PCR a amplifikace. Do popředí se také dostává myšlenka real-time sekvenování a je znatelný jistý odklon od postupů zavedených první a druhou generací sekvenování.

V současné době je nejrozšířenější technologií třetí generace metoda single molecule real time (SMRT) sekvenování. Podobně jako přístroje Illumina využívá fluorescenčně značených nukleotidů a detekce emitovaného záření. Avšak chceme-li se obejít bez předešlé amplifikace DNA, je potřeba značně snížit signál šumu. SMRT technologie nabízejí řešení, a to pomocí dvou hlavních inovací. První z nich jsou tzv. zero-mode waveguides (ZMWs). Jedná se o jamky o rozměrech v řádu nanometrů pohlcující světlo excitačního paprsku. Excitační paprsek je tak schopen pronikat maximálně do spodních 30 nm každého ZMWs, kde také probíhá replikace jednotlivých vláken DNA (viz obr. 6) [21]. Druhou inovací je odštěpení fluoroforu od nukleotidu při začleňování nukleotidu do nově vznikajícího řetězce DNA. Fluorofor tak může difundovat pryč z detekčního objemu a později se nepodílí na vzniku šumu. Tyto nové techniky umožnily značné snížení šumu a detekci emitovaného záření i při velmi malých koncentracích [22].



Obr. 6: Replikace vláken DNA v jamkách zero-mode waveguides. [57]

Slibný pokrok v DNA sekvenování se očekává i od technologie nanopore. Princip metody vychází z faktu, že jednovláknová DNA procházející proteinovým nanopórem v membráně vyvolává změny proudu toku částic (viz obr. 7). Na membránu je přivedeno napětí, které způsobuje pohyb částic skrz nanopór. Při průchodu jednotlivých bází jsou vyvolávány charakteristické změny proudu, jež jsou zaznamenávány, a následnou analýzou takto získaného signálu je rekonstruována primární struktura DNA. Vývojem nanopore sekvenátorů se zabývá například firma Oxford Nanopore Technologies.



Obr. 7: Oxford Nanopore – patrné jsou změny proudu při průchodu molekul. [32]

3 Zpracování dat

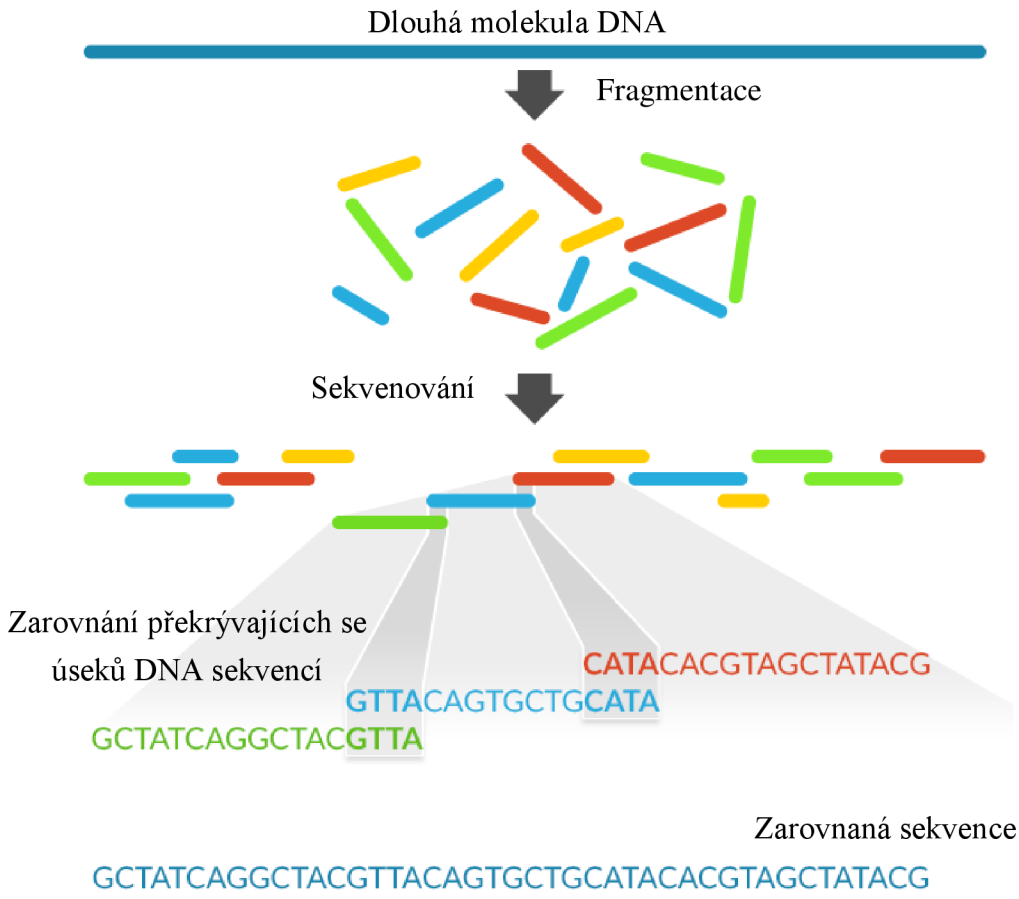
Pojem metagenomika se zdá jasný, jak se ale liší zpracování metagenomických dat od studia jediného organismu? Odlišnosti najdeme již při přípravě vzorků před sekvenováním. Zatímco při studiu jediného organismu je obvykle vzorek setřen tampónem, rozočkován na Petriho miskách a jednotlivé kolonie jsou poté manuálně vybrány k osekvenování, při zpracování metagenomických dat je po odběru vzorek ihned zpracován extrakcí DNA a amplifikací. Je tedy vynechán krok rozočkování, který se podílí na značné redukci počtu mikroorganismů ve vzorku. [3]

V následujících kapitolách si popíšeme jednotlivé kroky zpracování mikrobiálních dat. Jsou popsány dva hlavní přístupy k analýze mikrobiomu, a to sice shotgun a cílené amplikonové sekvenování. Dále jsou vysvětleny pojmy multiplexace, barcoding a trimování dat, popsány nejčastěji používané shlukovací algoritmy a referenční databáze, definován pojem OTU a vysvětlen proces tvorby OTU tabulky. Všechny tyto body jsou důležitými prvky zpracování mikrobiálních dat a odlišné přístupy v některém z dílčích kroků mohou vést k odhalení odlišných vazeb mezi jednotlivými mikrobiomy.

3.1 Shotgun sekvenování

Shotgun sekvenování je metoda určená pro sekvenování velmi dlouhých vláken DNA nebo celých genomů. Molekuly DNA jsou náhodně rozštěpeny do menších fragmentů, které jsou dále sekvenovány některým ze sekvenačních přístrojů popsaných v 2. kapitole. Osekvenované fragmenty jsou pomocí počítačového softwaru zarovnány na základě svých překryvů a slouží k sestavení celého původního vlákna.

Metod rozštěpení DNA na menší fragmenty je celá řada. Za všechny jmenujme např.: nebulizaci (rozštěpení pomocí stlačeného vzduchu), enzymatické štěpení nebo štěpení pomocí ultrazvuku, jehož vibrace mohou pomocí kavitačních jevů způsobit fragmentaci molekuly DNA [11]. Delší fragmenty vedou k přesnějšímu zarovnání díky delším překryvům, avšak s délkou čtení může narůstat chybovost.



Obr. 8: Shotgun metoda sekvenování.

Samotné zarovnávání takto získaných a upravených sekvencí je výpočetně poměrně náročný proces zvyšující se s počtem fragmentů. Nicméně vyšší počet čtení se podílí na korekci chybovosti, která mohla při sekvenování vzniknout. Mezi problematické oblasti zarovnání patří především repetitivní úseky DNA. Tento problém může být značně redukován využitím referenčního genomu.

3.2 Cílené amplikonové sekvenování genů

Populárnějším přístupem k získávání mikrobiálních dat je metoda tzv. amplikonového sekvenování pouze jediného genu, nikoliv celého genomu organismu.

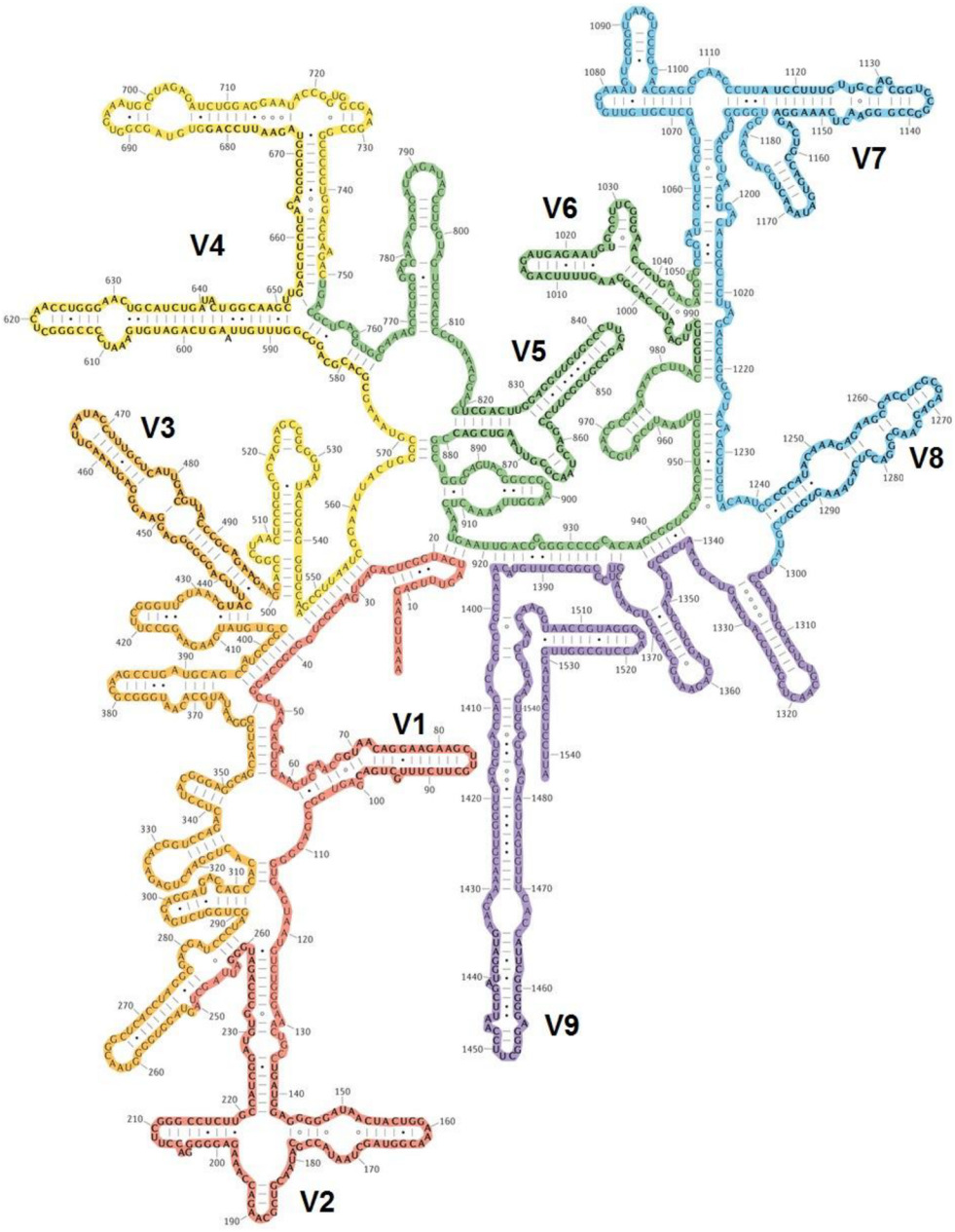
Gen používaný pro amplikonové sekvenování by měl splňovat tři základní požadavky:

- měl by být přítomen ve všech zkoumaných organismech,
- měl by obsahovat úseky, které jsou napříč organismy totožné,
- a dále takové úseky, které jsou napříč organismy velmi rozdílné a umožňují identifikaci organismu.

Ukázalo se, že všechny výše popsané požadavky dobře splňuje například gen kódující 16S rRNA zobrazený na obrázku 9. Průkopníky ve využívání 16S rRNA genu v oblasti

fylogenetiky jsou Carl Woese a George Fox, kteří v roce 1977 sekvenováním právě tohoto genu dokázali vymezit tři hlavní domény života, a to sice archea, bakterie a eukaryota. [13]

16S rRNA tvoří složku menší 30S podjednotky ribozomu prokaryot (tedy archeí a bakterií). Vzhledem k funkci ribozomu můžeme předpokládat, že se různé formy tohoto genu budou vyskytovat ve všech žijících mikroorganismech (jejím protějškem u eukaryot je 18S rRNA). Zároveň nesprávná funkce ribozomu vede ke smrti organismu, tudíž v kódujícím genu najdeme sekvence, které nejsou napříč organismy nijak pozměněny mutacemi, protože mutace v těchto klíčových regionech vede k syntéze nefunkčního ribozomu a zániku jedince. Ale také naopak, v genu se rovněž vyskytují regiony, kde byla variabilita možná bez fatálních následků.



Obr. 9: Variabilní regiony 16S ribozomální RNA. [58]

Pro potřeby studia mikrobiologie není dokonce nutné sekvenování celého genu 16S rRNA. Často se pracuje se čtvrtým variabilním regionem tohoto genu a konzervativními regiony, které jej vymezují (viz obr. 10). Avšak volba regionu by měla sledovat konkrétní cíle studie. Znalost konzervativních regionů je nutná pro návrh metody PCR k amplifikaci variabilních úseků.

Obr. 10: konzervativní (zelené) a variabilní (šedé) regiony 16S rRNA genu. [35]

3.3 Multiplexace dat

Za účelem finanční i časové úspornosti se pro soubory vzorků používá tzv. multiplexace. Jedná se v podstatě o smíchání většího množství vzorků do jedné sekvenační směsi a následné souběžné sekvenování všech vzorků najednou.

Pro potřeby multiplexace je nejprve nutné označení vzorků pomocí barcodů, což jsou speciální markery vážící se na začátek DNA sekvencí a sloužící pro následnou identifikaci a přiřazení sekvence ke konkrétnímu vzorku. Teprve poté je možné smíchání vzorků a zahájení souběžného sekvenování.

Výstupy ze sekvenátoru jsou před dalším zpracováním roztříděny ke konkrétním vzorkům na základě svých barcodů a zároveň jsou tyto barcody (a také primerové sekvence) odstraněny pomocí počítačových metod procesem trimování. Celý proces demultiplexace a trimování může být proveden například pomocí softwaru QIIME (více o software QIIME v kapitole 4). [14]

3.4 Operační taxonomická jednotka a shlukování

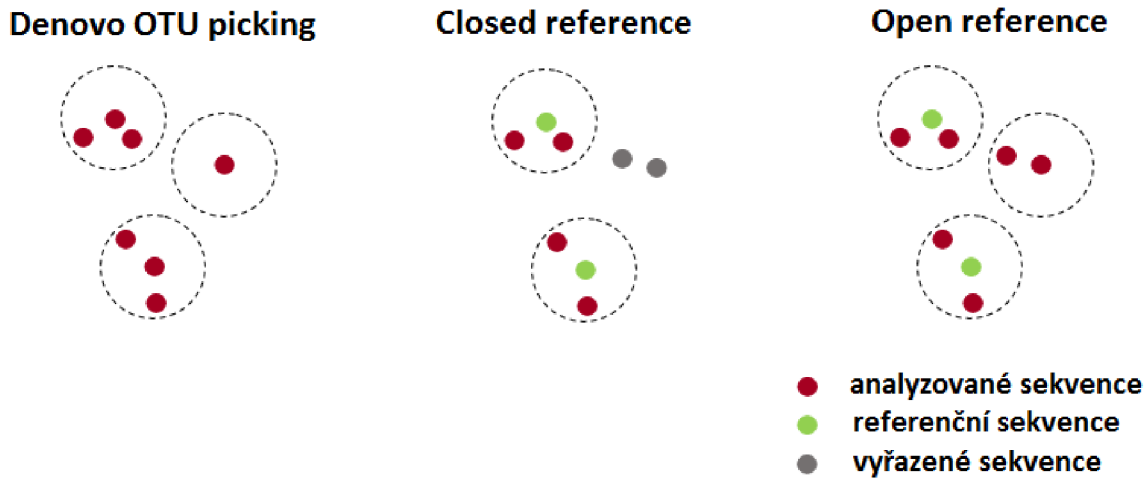
Demultiplexované a trimované sekvence jsou shlukovány některým z clusterovacích algoritmů a vznikají tzv. operační taxonomické jednotky. Pojmem operační taxonomická jednotka neboli OTU (z anglického „Operational Taxonomic Unit“) je tedy obvykle rozuměn shluk organismů seskupených na základě podobnosti DNA sekvencí. Proces tvorby shluků je nazýván OTU picking a je typický pro amplikonové sekvenování.

Clustering umožňuje nejen kompresi dat z několika milionů sekvencí do cca tisícovky shluků, čímž značně usnadňuje další zpracování, ale také se podílí na odstraňování chyb, které mohly vzniknout sekvenováním nebo samotnou amplifikací.

Ideálem při vytváření OTU je identifikace všech známých i nových druhů nacházejících se ve vzorku. Ovšem v praxi může být splnění tohoto cíle značně komplikované. Existují tři

základní přístupy k OTU shlukování. Jejich volba obvykle sleduje cíle analýzy, charakter vzorků a výpočetní náročnost:

- *De novo* clustering – Shlukování čtení na základě podobností mezi sebou.
- Closed reference clustering – Porovnávání čtení s referenční databází a na základě podobností třídění do OTU. Čtení, která nejsou přiřazena žádné sekvenci v databázi, jsou z dalšího zpracování vyřazena.
- Open reference clustering – Čtení jsou porovnávána s referenční databází podobně jako v closed reference clustering, avšak nezařazená čtení nejsou vyřazena, ale shlukována mezi sebou *de novo*.



Obr. 11: Přehled OTU picking metod.

Využití *de novo* shlukování je nutné v případě, kdy neexistuje žádná referenční databáze. Jeho hlavní nevýhodou je velká výpočetní náročnost, avšak na rozdíl od closed reference shlukování umožňuje nalezení nových druhů. Naopak využití closed reference clustering může být výhodné, pokud je k dispozici komplexní a kvalitní referenční databáze. Obecně nejdoporučovanějším přístupem je open reference clustering, který spojuje výhody obou předcházejících metod. [16]

Nezávisle na zvoleném přístupu shlukování je možné modifikovat práh podobnosti. Obvykle se používá práh minimálně 97% podobnosti, avšak je možné klasifikovat data také na 99% nebo jiném prahu podobnosti.

Výsledky shlukování může také ovlivnit zvolená shlukovací metoda. Těch existuje celá řada, jen namátkou můžeme zmínit shlukování k centroidům, metodu nejbližšího souseda či metodu nejvzdálenějšího souseda. Obecně lze říci, že vhodná shlukovací metoda by měla splňovat požadavek velké podobnosti uvnitř vytvořených shluků a zároveň co nejmenší podobnost mezi shluky.

Vybrané shlukovací algoritmy

Jedním z rozšířených shlukovacích algoritmů je UCLUST. Algoritmus využívá shlukování k centroidům, kdy je nejdříve vybrána náhodná sekvence a stává se centroidem. Poté jsou postupně vybírány další sekvence z datasetu a porovnávány s tímto centroidem. Je-li splněna podmínka minimální podobnosti, je sekvence přiřazena k centroidu do shluku, není-li podmínka splněna, stává se nová sekvence dalším centroidem, se kterým budou další sekvence porovnávány. UCLUST je takzvaný greedy nebo také hladový algoritmus a jeho výsledek je závislý na pořadí vstupních sekvencí. Drží se předpokladu, že první náhodně vybraná sekvence je centroidem, avšak již tento předpoklad může být značně zavádějící.

Zarovnávání každé sekvence s centroidy je výpočetně náročný proces, proto UCLUST využívá pro porovnání podobností USEARCH. USEARCH pracuje s k-mery a porovnává kolik k-merů mají sekvence s centroidem společné. K samotnému zarovnání a porovnání podobnosti je pak vybráno pouze několik sekvencí s nejvyšším počtem shodných k-merů.

Hlavní výhoda UCLUST algoritmu v porovnání s ostatními shlukovacími algoritmy je jeho rychlost a nízká výpočetní náročnost. Avšak patrné jsou rovněž nevýhody algoritmu. Patří sem nejen již zmíněná závislost na pořadí vstupních sekvencí, ale také fakt, že ne všechny sekvence jsou porovnávány se všemi centroidy. Tato skutečnost může vést k nezařazení sekvence do správného shluku, ačkoliv si je s daným centroidem velmi podobná. [23]

Trváme-li na porovnání všech sekvencí se všemi danými centroidy, nabízí se možnost využití shlukovacího a porovnávacího algoritmu CD-HIT. Tento algoritmus pracuje podobně jako UCLUST, avšak nevyužívá odhadu podobnosti sekvencí pomocí k-merů. Metoda byla původně navržena pro shlukování proteinů a později rozšířena také na nukleotidové sekvence. Od UCLUSTu se liší počátečním seřazením sekvencí v sestupném pořadí. Nejvyšší důležitost je tak přikládána nejdelší z analyzovaných sekvencí. Pro snížení výpočetní náročnosti je využito filtru krátkých slov a tabulky indexů. Algoritmus vychází z úvahy, že pro danou podobnost musí mít sekvence alespoň jedno slovo délky x. Například pro podobnost dvou sekvencí nad 90 % musí mít tyto sekvence identické alespoň jedno slovo délky 10 bází. Podobně je tomu i u dalších hodnot prahu podobnosti. Dvojice sekvencí nesplňující tuto podmínku nejsou zarovnávány. [24]

Dalším algoritmem využitelným pro *de novo* shlukování je mothur. Pracuje s maticí distancí, pomocí které poté shlukuje sekvence metodou nejbližšího, průměrného nebo nejvzdálenějšího souseda. Avšak již samotné získání matice distancí pro všechny sekvence je pro větší objemy dat problematické. [25]

Posledním algoritmem, který si zmíníme, je swarm. Algoritmus v prvním kroku vybere náhodnou sekvenci označovanou jako semínko. Následně hledá sekvence, které se od semínka liší o zadanou vzdálenost d a přidá je do shluku. Verze swarm 2.0 používá defaultně

nastaveného $d = 1$. V dalším kroku hledá sekvence, které se liší od předešlých sekvencí o vzdálenost d. Takovým způsobem shluk roste, dokud jsou k dispozici sekvence splňující zmíněnou podmínku. Nevýhodou této metody je její předpoklad, že jednotlivé shluky mají mezi sebou velké vzdálenosti. [59]

3.5 Referenční databáze

Využívání referenčních databází se stalo důležitým prvkem analýzy dat. Uplatňují se nejen při shlukování sekvencí, ale slouží také k taxonomické klasifikaci. Velké veřejné databáze jako GenBank se velmi rychle rozrůstají, avšak mohou obsahovat nevalidovaná data nízké kvality. Pro analýzu 16S ribozomální RNA se tak více uplatňují specializované databáze jako Greengenes, SILVA nebo Ribosomal Database Project.

Výběr konkrétní databáze by měl zohledňovat jejich aktuálnost a specializaci. Zároveň, je-li cílem studie porovnání výsledků s jinou studií, je nutné zvážit, zda použití odlišné databáze nepovede ke zkreslené interpretaci výsledků.

3.6 OTU tabulky

Je zvykem výsledky sekvenování a výše popsaného zpracování dat převádět do podoby tzv. OTU tabulek. Jedná se o tabulky poskytující informaci o celkovém počtu nalezených sekvencí v každé skupině OTU u každého vzorku. Příklad takovéto OTU tabulky je na obr. 12. Tento zápis je výhodný pro prostý rozbor dat v Excelu, vizualizaci i složitější analýzy.

	116	117	118	119	120	121	122	123	124
s__cinerea	25	18	11	0	4	2	0	11	1
s__depolymerans	0	1	0	3	0	0	0	0	0
s__formigenes	1	0	0	6	29	318	0	69	148
s__subflava	9143	7648	1005	23	362	2319	44	1393	36

Obr. 12: OTU tabulka – v prvním řádku jsou uvedeny čísla vzorků a v prvním sloupci identifikátory OTU

4 QIIME – Quantitative Insights Into Microbial Ecology

QIIME (z anglického „Quantitative Insights Into Microbial Ecology“) je open source bioinformatický software pro analýzu mikrobiomu. Uživatelům umožňuje analýzu surových dat a jejich grafické či statistické vyjádření. Mezi funkce QIIME softwaru patří demultiplexace dat, OTU picking, taxonomická klasifikace, fylogenetické rekonstrukce, analýzy diverzity a vizualizace. V této práci je využito softwaru QIIME k procesu OTU picking a tvorbě OTU tabulek. [15]

Vstupem softwaru jsou obvykle sekvenovaná data a mapping file obsahující dodatečné informace o vzorcích, tzv. metadata (např.: z jakého mikrobiomu vzorky pocházejí). Často je potřeba také vyplnit parametry sekvenování či doplnit analýzy skórem kvality. Oba soubory (sekvenovaná data i mapping file) musí dodržovat jistý formát.

Mapping file

Mapping file je textový soubor s příponou txt, jehož formát je patrný z obr. 13. Uvedeme si tři klíčová pravidla pro tvorbu těchto souborů, podrobnější informace mohou být získány v [36]:

1) Hlavička souboru začíná znakem křížku (*#*) a polem „SampleID“. Další požadovaná pole jsou „BarcodeSequence“, „LinkerPrimerSequence“ a „Description“. Všechna pole jsou oddělena tabulátorem.
2) Případná volitelná pole (na obr. 13 pole „Treatment“ a „DOB“) musí následovat za polem „LinkerPrimerSequence“. Pole „Description“ je pak posledním.
3) Datová pole nezačínají znakem křížku, jsou oddělena tabulátory a mohou obsahovat pouze alfanumerické znaky a některé speciální symboly (viz [36]). Veškeré řádky následující hlavičku a začínající symbolem křížku jsou považovány za poznámky a QIIME je ignoruje.

Tvůrci platformy QIIME doporučují pro testování kompatibility mapping file využít předpřipraveného QIIME skriptu *validate_mapping_file.py*.

```
#SampleID       BarcodeSequence LinkerPrimerSequence    ReversePrimer   Dodavatel   ZemePuvodu    Description
#Priklad mapping file pro praci v QIIME
21      ACGAGTGCGT      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     B    Vietnam Paracheirodon_axelrodi
22      ACGCTCGACA      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     B    Vietnam Poecilia_sphenops
23      AGACGCACTC      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     B    Vietnam Hypostomus_plecostomus
29      AGCACTGTAG      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     D    Peru    Otocinclus_affinis
32      ATCAGACACG      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     D    Peru    Otocinclus_affinis
44      CGTGTCTCTA      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     E    Vietnam Plecostomus_Gold
46      CTCGCGTGTC      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     E    Vietnam Gold_Black_Molla
55      TAGTATCAGC      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     F    Vietnam V4
28      ATACGACGTA      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     C    Peru    Panaqolus_changae
43      TCACGTACTA      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     E    Thaisko Poecilia_sphenops
30      TCTACGTAGC      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     D    Peru    Surubim_Lima
40      ACGACTACAG      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     E    Vietnam Sewelia
61      TACGAGTATG      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     G    Vietnam Xiphophorus_maculatus
39      ACTACTATGT      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     E    Cina    Xenopus_laevis_Albin
57      AGACTATACT      GGAGGCAGCAGTRRGGAAT     CTACCRGGGTATCTAATCC     G    Vietnam Gyrinocheilus
```

Obr. 13: Příklad formátování mapping file.

FASTA soubor

Jak již bylo zmíněno v kapitole 3.3, multiplexovaná data musí pro další analýzu projít demultiplexací a trimováním. Ať už se rozhodneme tyto kroky provést pomocí QIIME nebo jiného programu, výsledný FASTA formát by měl odpovídat následujícím požadavkům:

1. Každá sekvence má svou hlavičku začínající identifikátorem vzorku (shodný s identifikátorem z mapping file). Poté následuje podtržítko a celé číslo identifikující sekvenci v daném vzorku (viz obr. 14).
2. Žádná z hlaviček sekvencí se nesmí vyskytovat duplikovaně.
3. Sekvence jsou vypsány vždy do jediného řádku pod hlavičkou. (Pozn.: V některých textových editorech se mohou sekvence jevit jako rozepsané do více řádků, jedná se však pouze o vnitřní předpis zkracování dlouhých slov textového editoru.)

```
>21_1 H885UFZ01BPA1G orig_bc=ACGAGTGCGT new_bc=ACGAGTGCGT bc_diffs=0
ATTGGACAATGGGCGCAAGCCTGATCCAGCAATGCCGCGTGAGTGATGAAGGCCTTCGGGTCGTAAAGCTCTTTTACC
>29_2 H885UFZ01AWLQE orig_bc=AGCACTGTAG new_bc=AGCACTGTAG bc_diffs=0
ATTGGGCAATGGATGAAAGTCTGACCCAGCCATGCCGCGTGCCGGATGAAGGCGCTCTGCGTTGTAAACGGCTTTTAT
>55_3 H885UFZ01ALBSV orig_bc=TAGTATCAGC new_bc=TAGTATCAGC bc_diffs=0
CTTGCGCAATGGGGGCAACCCTGACGCAGCGACGCCGCGTGAGTGACGAAGGCCTTCGGGTTGTAAAGCTCTGTGGAG
>28_4 H885UFZ01BFJHR orig_bc=ATACGACGTA new_bc=ATACGACGTA bc_diffs=0
ATTGGACAATGGGCGGAAGCCTGATCCAGCCATGCCGCGTGAAGGAATACGGTCCTATGGATTTTAAACTTCTTTTGT
>57_5 H885UFZ01A1PG7 orig_bc=AGACTATACT new_bc=AGACTATACT bc_diffs=0
TTTGGACAATGGACGCAAGTCTGATCCAGCCATGCCGCGTGCGGGAAGAAGGCCTTCGGTTGTAACCGCTTTTGTCAG
>32_6 H885UFZ01BMOC0 orig_bc=ATCAGACACG new_bc=ATCAGACACG bc_diffs=0
ATTGGTCAATGGAGGCAACTCTGAACCAGCCATGCCGCGTGCAGGAAGACAGCCCTCTGGGTCGTAAACTGCTTTTAT
>40_7 H885UFZ01ADV3V orig_bc=ACGACTACAG new_bc=ACGACTACAG bc_diffs=0
ATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGAGTGATGAAGGCCTTAGGGTTGTAAAGCTCTTTCACC
```

Obr. 14: Příklad formátování demultiplexovaného FASTA souboru.

V případě nejistot, zda FASTA soubor splňuje požadavky QIIME na demultiplexovaný FASTA soubor, je možné využít předpřipraveného QIIME skriptu *validate_demultiplexed_fasta.py*, který validitu souboru ověří.

Proces OTU picking

Software QIIME nabízí všechny tři základní přístupy k OTU picking – *de novo* OTU picking, closed i open reference OTU picking. Slouží k tomu QIIME skripty *pick_de_novo_otus.py*, *pick_closed_reference_otus.py* a *pick_open_reference_otus.py*.

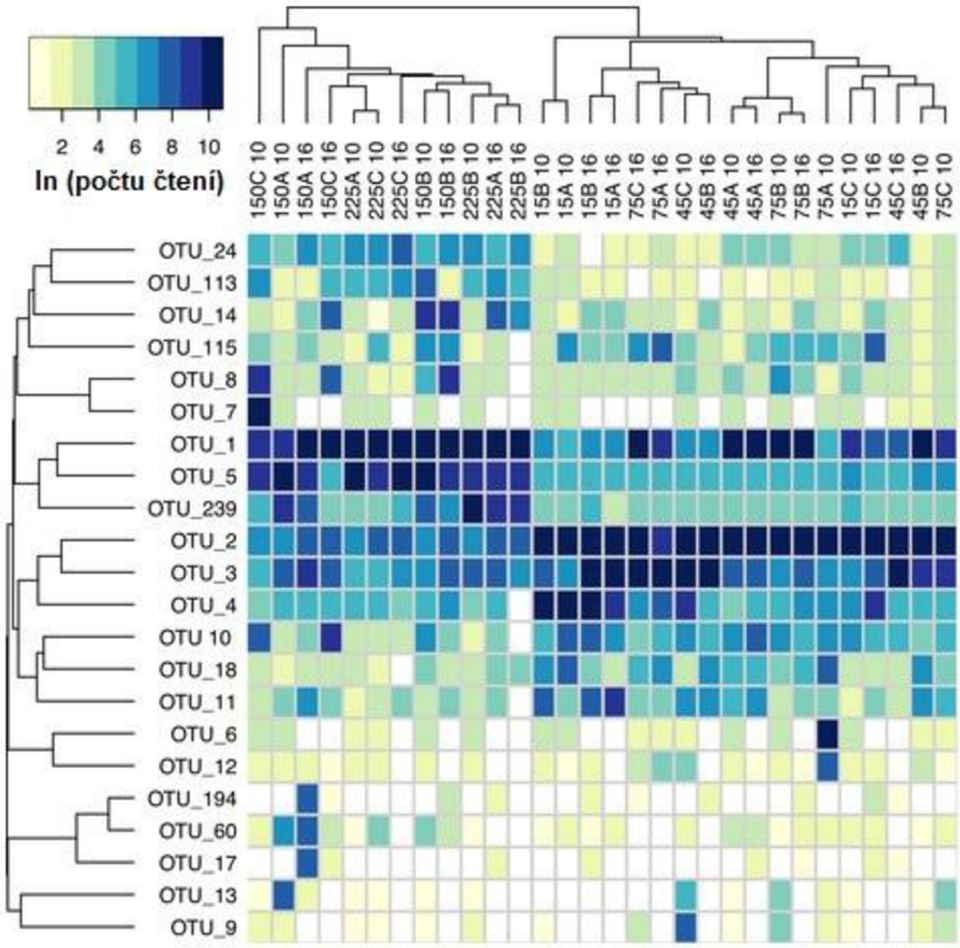
Skripty nabízejí celou řadu nastavitelných parametrů. Jedním z nich je například volba clusterovací metody. V současné době jsou v QIIME implementovány clusterovací algoritmy cd-hit, Mothur, blast, uclust, swarm a další. Dále QIIME nabízí také možnost volby referenční databáze. Pro 16S rRNA geny se nabízí volba Greengenes nebo SILVA.

Výstupem je pak nejen textový soubor popisující vytvořené shluky, ale také FASTA soubor obsahující jednu reprezentativní sekvenci vybranou pro každou OTU, soubor zarovnaných sekvencí, taxonomického zařazení jednotlivých OTU, fylogenetický strom a OTU tabulku v biom formátu.

5 Vizualizace dat

Jak již vyplynulo z předešlých kapitol, mikrobiální data jsou velmi komplexní, multidimenzionální, a analýza prostým okem je takřka nemožná. Výběr efektivní vizualizační metody může odhalit skryté vazby mezi mikrobiomy a je důležitý pro efektivní zkoumání, interpretaci a zpracování tak bohatého datasetu. Nejjednodušší vizualizační metodou je například prosté převedení OTU tabulky na teplotní mapu (heatmapu). Ukázka takové teplotní mapy je na obrázku 15. Vzorky i OTU jsou podrobeny shlukové analýze a seřazeny tak, aby se podobné vzorky a OTU nacházely v OTU tabulce vedle sebe. Odstíny v tabulce jsou dány přirozeným logaritmem počtu nalezených čtení.

Ačkoliv tato metoda nabízí jednodušší orientaci v datech než prostá numerická reprezentace, její aplikace na mnohorozměrné prostory je obtížná.

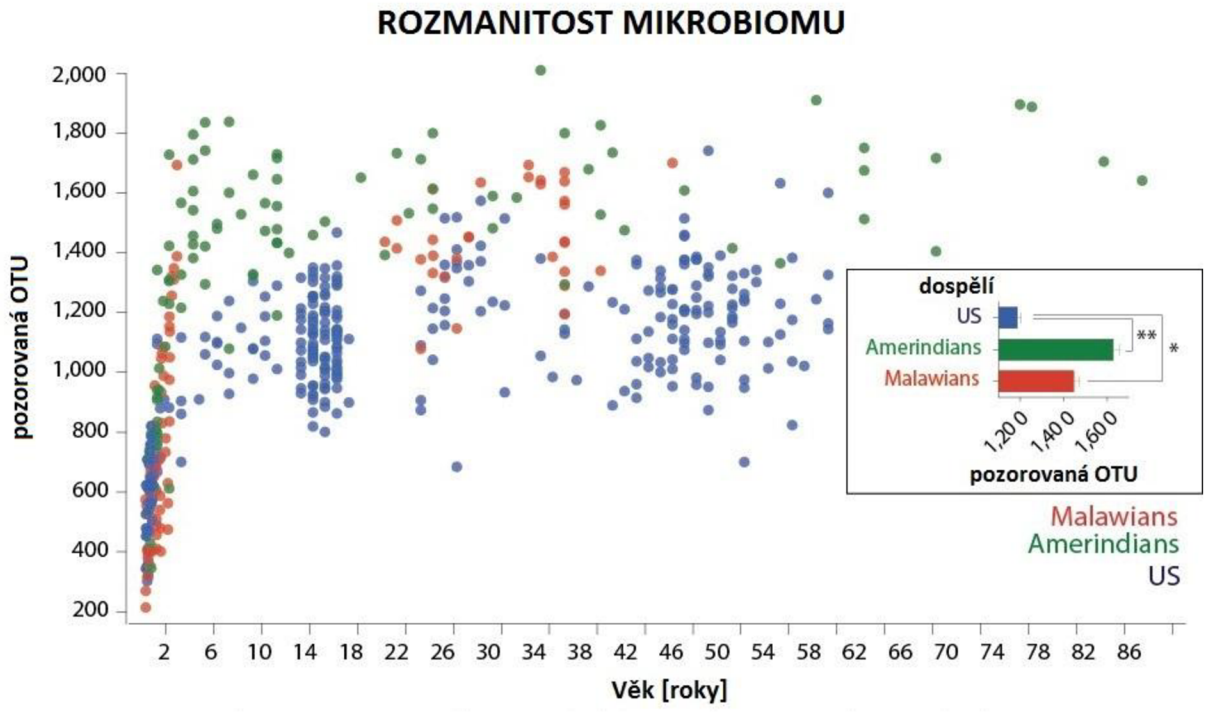


Obr. 15: Teplotní mapa. [60]

5.1 Jednoduché bodové 2D grafy

Jednoduché bodové 2D grafy a správná volba jejich os jsou důkazem, že vizuální analýza mikrobiálních dat nemusí být vždy složitá, aby odhalila zajímavé trendy v datech. Jako příklad zde uvádím dva bodové typy grafů použité v [38]. K analýze střevní mikroflóry zde bylo využito jednoduchého grafu, kde osa y odpovídala počtu pozorovaných OTU ve vzorku

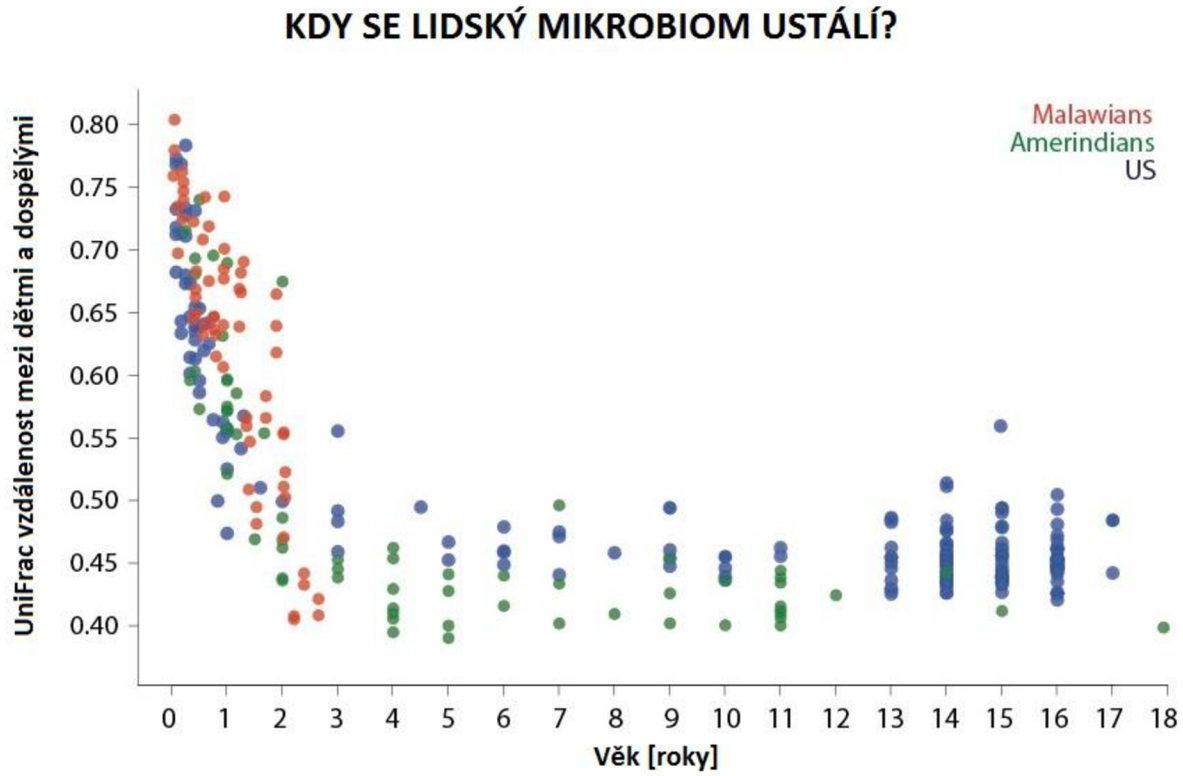
a osa x věku probandů (viz obr. 16). I tento velmi jednoduchý přístup dokázal odhalit vzrůstající počet OTU v organismu člověka v prvních dvou letech života. Díky barevnému značení je také možné porovnávat alfa diverzitu (diverzitu jednoho společenství) mezi jednotlivými národnostmi.



Obr. 16: Rozmanitost lidského mikrobiomu. [38]

Druhým hodnotným grafem použitým ve studii Tanyi Yatsunenko a kolektivu je obr. 17. Tento graf je skvělým příkladem využití referenční databáze k analýze vzorků. V tomto případě sloužila jako referenční databáze skupina vzorků dospělých participantů, avšak volba reference se mění s cílem projektu. Jednotlivé vzorky z datasetu jsou porovnávány se všemi referenčními vzorky, je vypočtena průměrná vzdálenost a ta je zanesena do grafu. Z obr. 17 je možno vyčíst beta diverzitu (diverzitu mezi společenstvy) jednotlivých vzorků vypočtenou pomocí UniFrac (Unique Fraction) metriky [26]. Tato metrika umožňuje měření vzdálenosti dvou mikrobiomů na základě jejich fylogenetické informace. Nejdříve je ze sekvencí genetické informace sestaven fylogenetický strom a poté podstromy ze všech dvojic zkoumaných mikrobiomů. Vzdálenost mikrobiomů je vypočtena jako frakce délky větví náležící jednomu nebo druhému mikrobiomu, ale ne oběma zaráz.

Tento způsob vizualizace dat je velmi přehledný a účinně redukuje multidimenzionalitu dat. Důležitým faktorem ovlivňujícím analýzu je zde volba referenční databáze. Nevhodně zvolená referenční skupina může vést k opomenutí důležitých informací a snížení celkové výtěžnosti analýzy.

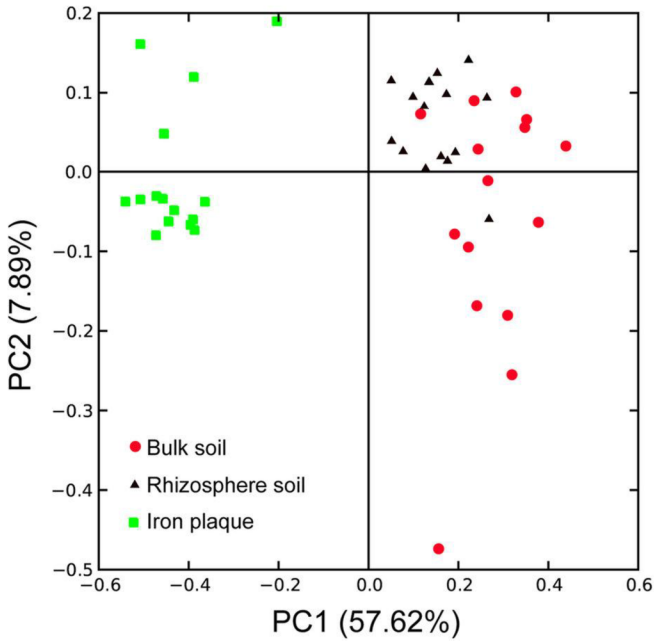


Obr. 17: Vývoj lidského mikrobiomu. [38]

5.2 Nástroje redukce dimenzí

Principal coordinate analysis

Jedním z typických nástrojů mikrobiálních studií je vizualizace pomocí analýzy hlavních koordinát (PCoA z anglického „principal coordinate analysis“) s UniFrac metrikou. PCoA je podobná známější analýze hlavních komponent (PCA z anglického „principal components analysis“) avšak není vázána Euklidovskou metrikou, ale umožňuje využití i jiných metrik. PCoA tedy nabízí možnost pracovat s metrikami zohledňujícími fylogenetickou informaci obsaženou v mikrobiálních datech, např.: se zmíněnou UniFrac metrikou (viz obr. 18).



Obr. 18: Principal coordinate analysis odvozená pomocí nevážené UniFrac metriky mezi 16S rRNA geny mikrobiálních komunit různých druhů půd. [40]

Cílem této analýzy je redukce n-dimenzionálního prostoru do nižšího počtu dimenzí (tzv. hlavních koordinát), které poskytují maximum informací o analyzovaných datech. Zpravidla se jedná o dimenze vykazující nejvyšší variabilitu dat. [28] Variabilita dat v jednotlivých dimenzích může být analyzována pomocí tzv. scree plotů znázorňujících procenta variability vyčerpané jednotlivými koordinátami. Mohou mít mnoho podob. Mezi nejjednodušší patří vyjádření pomocí sloupcových diagramů. Scree ploty mohou usnadnit proces výběru optimálního počtu dimenzí, avšak pro možnosti vizualizace se obvykle používá 2D nebo 3D prostor, neboť vykreslení většího počtu dimenzí je pro člověka těžko uchopitelné. Ideálním výstupem PCA a PCoA je co nejnižší počet výstupních dimenzí za současně co nejvyšší vyčerpané variability dat. [39]

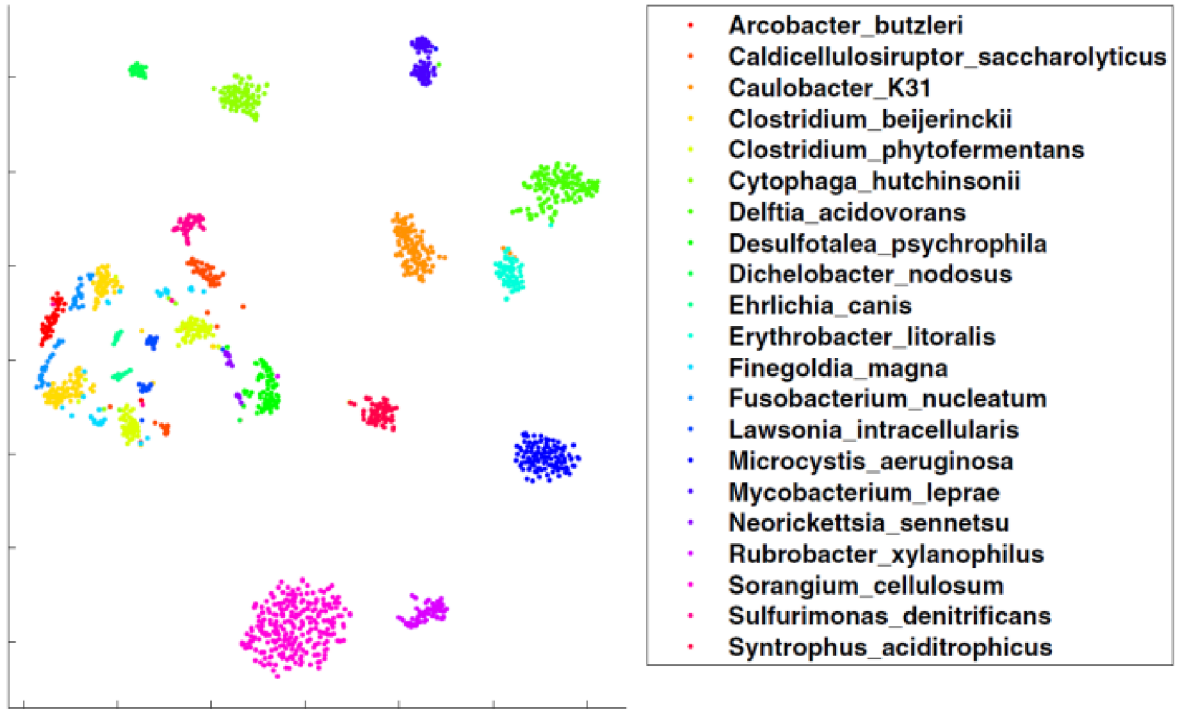
Tento přístup umožňuje vizualizovat vztahy mezi jednotlivými mikrobiomy a odhalit mikrobiální diverzitu, nicméně redukce dat ve smyslu výběru pouze dvou popř. tří hlavních koordinát a zahození zbylých může vést k nedostatečné charakterizaci dat. Poměrně nový přístup t-SNE z roku 2008 se snaží o úplnější popis dat.

T-distributed stochastic neighbor embedding

Metoda zvaná t-distributed stochastic neighbor embedding nebo také zkráceně t-SNE byla vyvinuta autory Laureansem van der Maatenem a Goeffrey Hintonem. Jedná se o variaci redukční techniky Stochastic Neighbor Embedding z roku 2002, avšak nabízí snazší čitelnost vizualizovaných dat, neboť redukuje trend shlukování bodů do středu grafu. [42]

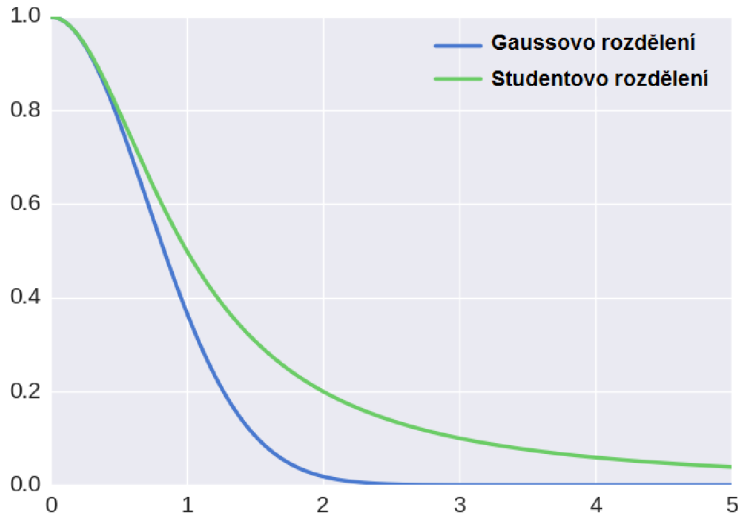
Jak již název napovídá, využívá t-SNE Studentova rozdělení. Uvažujme, že každý bod prostoru může být střední hodnotou Gaussova rozdělení s rozptylem σ. Výpočtem vzdálenosti

bodů od střední hodnoty Gaussova rozdělení tak může být vypočtena pravděpodobnost, která v podstatě popisuje podobnost dvou bodů v n-dimenzionálním prostoru. Takto je vytvořena matice podobností *p*. Poté, co jsou body umístěny do 2D či 3D prostoru, je obdobným způsobem vypočítána podobnost bodů v tomto nízkodimenzionálním systému, avšak místo Gaussova rozdělení je využito Studentova rozdělení s jedním stupněm volnosti. Získáváme tak matici podobností *q*. Výsledný graf je získán minimalizací Kullback-Leiblerovy divergence mezi vzdálenostmi *p* a *q* mezi jednotlivými body. Ukázka t-SNE je na obr. 19. [41]



Obr. 19: Vizualizace metagenomických dat pomocí t-SNE. [61]

Srovnáme-li distribuční funkci Gaussova rozdělení a Studentova rozdělení s jedním stupněm volnosti (viz obr. 20), je jasné, že použití Studentova rozdělení pro nízkodimenzionální prostory umožňuje vykreslení méně podobných bodů do větších vzdáleností a výsledné grafy se tak stávají mnohem přehlednějšími, než tomu bylo u použití původního SNE algoritmu. [42]



Obr. 20: Porovnání Gaussova rozdělení a Studentova rozdělení s jedním stupněm volnosti. [62]

Redukce dimenzí pomocí t-SNE i PCoA vede ke ztrátě informace o jednotlivých OTU a vykreslené body reprezentují soubor více OTU. Analýza zastoupení jednotlivých taxonů tak není pomocí těchto nástrojů možná.

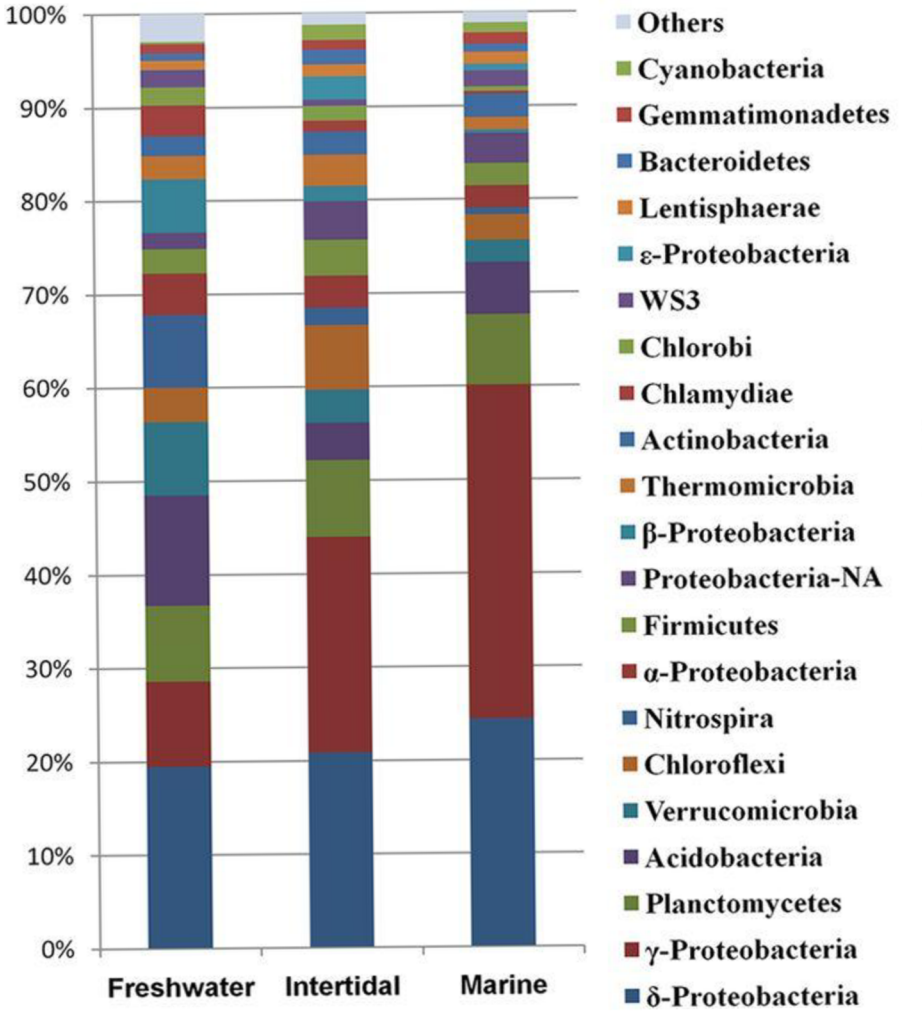
5.3 Grafické znázornění rozložení celku

Vhodnou volbou pro analýzu zastoupení jednotlivých taxonů může být grafické znázornění rozložení celku. Obvykle se pro tento účel využívá prstencový či koláčový graf. Obr. 21 ukazuje i variantu se sloupcovým grafem. Ať už je zvolen jakýkoliv útvar, vždy platí, že jeho celková plocha znázorňuje celý soubor (např. celý jeden mikrobiom) a jeho jednotlivé části představují procentuální podíl složek celku (např. procentuální zastoupení daných taxonomických řádů v mikrobiomu).

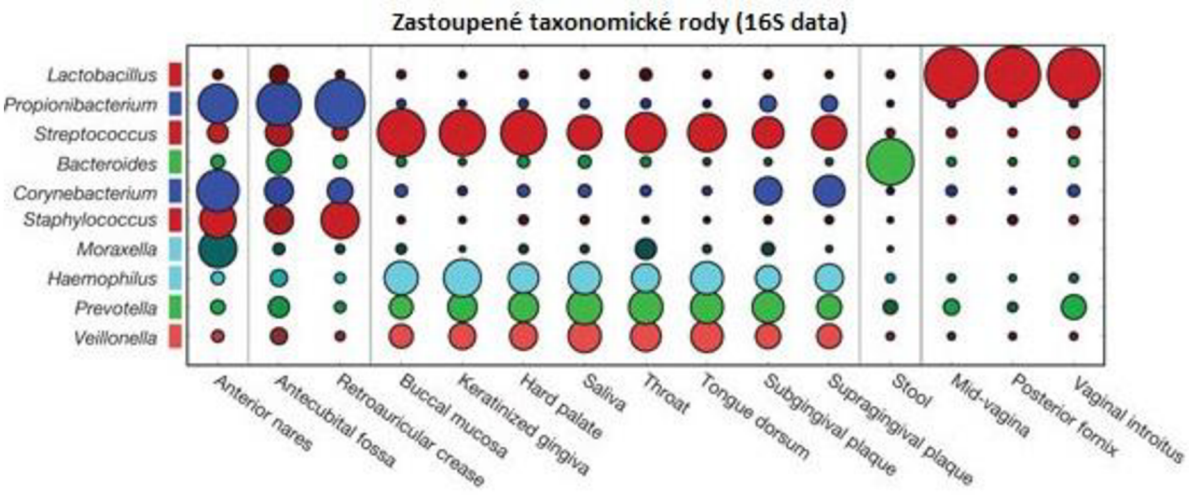
Srovnáním více takovýchto grafů je možné získat dobré povědomí o shodnosti nebo naopak rozdílnosti zastoupení organismů napříč mikrobiomy. Zároveň se nabízí možnost výběru taxonomické úrovně, pro kterou je graf rozdělen do složek, anebo vykreslení pouze OTU spadajících pod určitou taxonomickou úroveň (např. vykreslení všech zastoupených taxonomických rodů patřících do řádu *Lactobacillales*).

Jako hlavní nevýhodu tohoto typu grafu vnímám chybějící propojenost grafů mezi sebou. Propojení grafů by mohlo významně zjednodušit vzájemné srovnávání mikrobiomů. Rovněž porovnávání zastoupení jednotlivých složek může být bez doplňující číselné informace obtížné. Všimněme si v obr. 21 zastoupení kmenu *Chlorobi* ve vzorcích sladké vody, intertidalu a vodě mořské. Vzhledem k tomu, že se jednotlivé složky nacházejí v rozdílné výšce, je jejich vzájemné porovnání problematické. Zajímavé řešení bylo použito v Human Microbiom Project [44]. Jedná se v podstatě také o graf znázorňující rozložení celku, avšak

jednotlivé složky jsou zobrazeny pod sebou a jejich vzájemné porovnání je tím usnadněno (viz obr. 22).



Obr. 21: Sloupcový graf znázorňující zastoupení mikrobiálních kmenů v různých typech vod. [43]



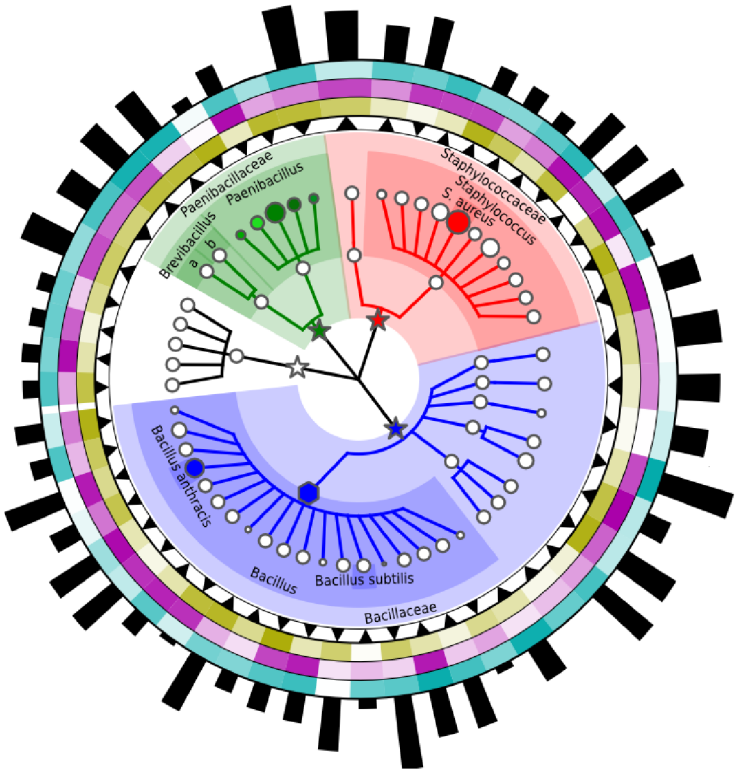
Obr. 22: Graf znázorňující zastoupení některých mikrobiálních rodů v různých částech lidského těla. [44]

5.4 Stromové struktury

Se studiem vztahů mezi organismy jde ruku v ruce tvorba stromů, a to se samozřejmě promítlo i do vizualizačních metod mikrobiálních dat. Dnes existuje celá řada přístupů k tvorbě stromů. Biologové rozlišují mnoho kladogramů a fylogenetických stromů, které se liší v detailech používaných metrik, délkách větví či topologii. Pro účely této diplomové práce stačí vnímat stromy jako prostředek pro vyjádření jisté příbuznosti mezi organismy.

Vizualizační platformy jako iTOL [45] nebo ETE toolkit [46] nabízejí celou řadu modifikací fylogenetických stromů, jejich barvení a zobrazení aditivních informací například pomocí tzv. prstenců. Pro studium mikrobiomů je zajímavé využití prstenců takovým způsobem, kde jeden prstenec odpovídá jedinému mikrobiomu. Intenzita zbarvení prstence pak odpovídá prevalenci daného organismu. Celková hojnost zastoupení daného organismu může být vykreslena v dalším prstenci zvlášť pomocí barplotu.

Obr. 23 je ukázkou výstupu výpočetního nástroje GraPhlAn [47]. Je zde patrné barvení uzlů fylogenetického stromu i jejich pozadí, rozličné velikosti jednotlivých uzlů i přídatné prstence. Program nabízí mnoho metod modifikace výsledného obrazu: změny velikosti stromu, barvení a přizpůsobování velikosti uzlů a větví, barvení pozadí jednotlivých částí stromu, popisky, přidání prstenců a mnoho dalšího.

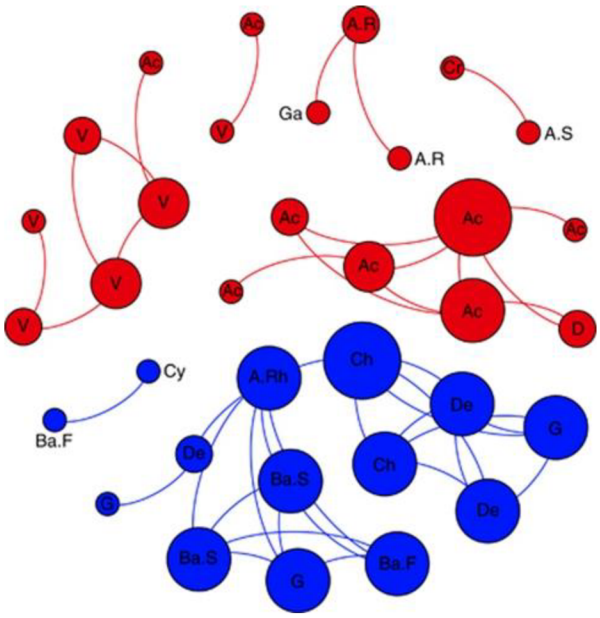


Obr. 23: Fylogenetický strom s prstenci vytvořený pomocí nástroje GraPhlAn. [47]

Zmíněné nástroje nabízejí vysokou personalizaci a vykreslení velkého množství dat do jediného grafu. Nicméně vzhledem k nejednotným pravidlům tvorby takových grafů mohou být na první pohled těžce srozumitelné, a proto je vhodné doplnit je slovním popisem.

5.5 Sítě

OTU tabulky mohou být rovněž vizualizovány pomocí sítí. Barberán a kolektiv [29] se zaměřují více na prozkoumání vzájemných asociací výskytu organismů v prostředí než na alfa či beta diverzitu mikrobiomu. Pro tento záměr si vybrali vizualizaci pomocí Bayesovské sítě, kde hrany sítí odpovídají silným a zároveň statisticky významným korelacím mezi uzly (tedy mezi jednotlivými OTU). Jak uvádějí ve svém článku, jako silná korelace je uvažován výsledek Spearmanova korelačního testu vyšší než 0.6. Jako hladina významnosti bylo zvoleno $\alpha = 0.01$. Pro popis topologie sítě byly kalkulovány také dodatečné výpočty, jako například proporcionální velikost uzlů k počtu hran s uzlem spojených (viz obr. 24).



Obr. 24: Bayesovská síť – vazby mezi OTU vycházející z korelační analýzy. Hrany značí silnou a statisticky významnou korelaci. [29]

Takto ohodnotit je možno nejen Bayesovské sítě, ale grafy obecně. Vážení uzlů může být spojeno nejen s výše zmíněným počtem vstupujících hran do uzlu, ale také například s kvantitativním zastoupením daného OTU. Váhování hran se obvykle projeví rozdílnou šířkou hran. Dalším prvkem, který se do značné míry podílí na čitelnosti grafu, je filtrace dat. Zvláště u rozsáhlých mikrobiálních dat je filtrace velmi žádoucí, neboť může značně zjednodušit proces interpretace odstraněním méně důležitých vazeb.

Bipartitní grafy

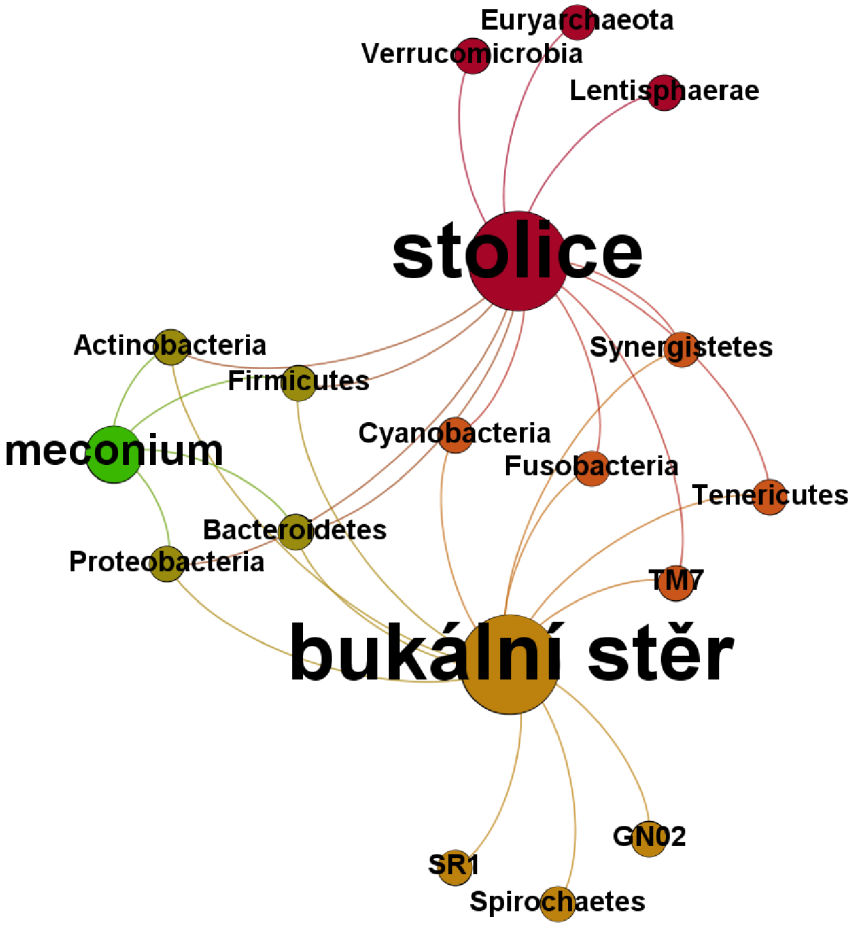
Speciálním typem grafu je bipartitní graf. Je charakterizován množinou vrcholů V, která může být rozdělena na dvě disjunktní množiny V_1 a V_2 takovým způsobem, že žádné dva vrcholy ze stejné množiny nejsou spojeny hranou e náležící do množiny hran E bipartitního grafu:

$$V = V_1 \cup V_2, V_1 \cap V_2 = \emptyset \quad (1)$$

$$\forall e = \{u, v\}, e \in E: u \in V_1 \wedge v \in V_2 \quad (2)$$

Disjunktní množiny V_1 a V_2 se také nazývají partity. OTU tabulky jsou vždy tvořeny dvěma partitami – první je reprezentována OTU a druhá jednotlivými vzorky. Přičemž každý vzorek může být propojen hranou pouze s OTU a naopak. [27]

Příklad takového grafu je na obr. 25. Byl sestaven na základě 16S rRNA dat získaných z lidského mikrobiomu. Je dobře patrné, které mikrobiální kmeny byly přítomny ve všech třech prostředích – například kmeny *Firmicutes* nebo *Actinobacteria*. A naopak, které kmeny byly charakteristické jen pro určité prostředí – například kmeny *Verrucomicrobia* a *Euryarchaeota* charakteristické pro vzorky odebrané ze stolice.



Obr. 25: Bipartitní graf – lidský mikrobiom.

Umístění jednotlivých uzlů podléhá alfa a beta diverzitě. Podíváme-li se na obr. 25 podrobněji, všimneme si, že kmeny jako *Lentisphaerae* nebo *Verrucomicrobia*, které nijak nepřispívají k alfa diverzitě mikrobiálního společenství slin (uzel bukální stěr), jsou tímto společenstvím jakoby odpuzovány. Naopak kmeny *Firmicutes* a *Bacteroidetes,* přítomné ve stolici, meconiu i slinách, zvyšují alfa diverzitu všech tří mikrobiomů, a jsou tak přitahovány do jejich pomyslného těžiště.

Podobné zákonitosti platí také pro beta diverzitu a umístění uzlů v grafu. Porovnáme-li beta diverzitu mikrobiomů meconium-stolice a meconium-bukální stěr, dá se předpokládat, že dostaneme dvě podobná čísla. Stejně tak vzdálenost mezi uzly meconium a stolice bude podobná vzdálenosti meconium-bukální stěr.

6 Praktická realizace bipartitních grafů

V programovacím prostředí R byl vytvořen balíček bipartiteOTU umožňující zpracování a čištění kvantitativních mikrobiálních dat a následnou analýzu pomocí bipartitních grafů. Balíček nabízí také možnost modifikace grafů pomocí váhování hran a uzlů, detekce komunit a barvení vrcholů a hran. Bipartitní grafy mohou být vykresleny přímo v prostředí R nebo uloženy v Graph Modeling Language formátu (GML formátu) a importovány do celé řady vizualizačních programů. R balíček je dostupný na www.github.com/safma/bipartiteOTU a jednoduše instalovatelný, včetně zabudované nápovědy, v prostředí R pomocí devtools příkazu *install_github("safma/bipartiteOTU")*.

6.1 Zpracování kvantitativních mikrobiálních dat

Přínosy redukce a zpracování kvantitativních mikrobiálních dat již byly popsány v kapitol 5.5. Navržený balíček bipartiteOTU nabízí čtyři možné přístupy ke zpracování OTU tabulek, jež je možné mezi sebou libovolně kombinovat.

Funkce *taxonomy_focus* umožňuje zaměřit se pouze na určitou taxonomickou skupinu a odstraňuje z tabulky OTU, které do této skupiny nespadají. Tento postup může být výhodný nezajímají-li nás mikrobiomy jako celek, ale chceme-li se více zaměřit na specifické organismy.

Další možností zpracování dat na základě taxonomie je shlukování OTU shodujících se v požadovaném taxonomickém určení. V balíčku je tato metoda umožněna funkcí *merge_taxonomy*. V dané taxonomické kategorii jsou jednotlivé OTU porovnávány mezi sebou a OTU se stejným zařazením sečteny a přiřazeny novému OTU nazvanému podle taxonomického určení, které je spojuje. To znamená, že při požadované taxonomické úrovni kmene jsou spojeny všechny OTU se stejným taxonomickým kmenem do jednoho. Tento přístup může významně redukovat OTU tabulky i na úrovni druhu, neboť i jeden druh přísluší v referenční databázi často mnohačetným OTU.

Podobným způsobem je možno data také třídit na základě metadat, tedy dodatečných informací o vzorcích. V balíčku k tomu slouží funkce *merge_metadata.* K vypsání nabízených metadat, jsou-li nějaká k dispozici, je pak možno využít funkce *get_metadata.*

Poslední metodou redukce OTU tabulek je prosté porovnávání četnosti pozorování (počtu čtení) jednotlivých OTU se zadaným prahem. Hodnoty nižší než práh jsou funkcí *threshold_drop* nastaveny na 0. Vyskytují-li se v OTU tabulce po tomto kroku nějaká OTU, jejichž četnosti pozorování jsou pro všechny vzorky nulové, jsou z tabulky automaticky odebrány.

6.2 Algoritmy detekce komunit

Účinným nástrojem k analýze grafů je detekce hlavních komunit. Cílem detekce je najít v grafu skupiny uzlů, které jsou hustě propojeny hranami uvnitř skupiny, avšak řídce

propojeny hranami s ostatními skupinami [50]. Hustota nebo také denzita grafu je popsána jako:

$$\rho = \frac{2m}{n(n-1)/2}, \tag{3}$$

kde m je počet hran a n je počet uzlů v grafu. Rovnice 3 může být snadno modifikována k výpočtu hustoty komunity grafu a sloužit tak k posouzení kvality shluku porovnáním hustoty komunity s celkovou hustotou grafu.

V praxi se jako optimalizační kritérium pro hledání optimálních komunit využívá tzv. modularity. Jedná se o funkci měřící kvalitu rozdělení grafu do skupiny komunit porovnáním s náhodným grafem o stejném počtu uzlů a jejich stupňů, tedy hran vycházejících z jednotlivých uzlů.

Skóre modularity je dáno rovnicí 4: [51]

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j), = \sum_{u}(e_{uu} - a_u^2), \tag{4}$$

kde m je počet hran grafu, A je matice sousedností grafu, k_i a k_j značí počet stupňů uzlu i a j, c_i a c_j komunitu uzlu i a j, e_{uu} je zlomek hran uvnitř komunity u:

$$e_{uu} = \sum_{ij}\frac{A_{ij}}{2m}1_{i\epsilon c_u}1_{j\epsilon c_u} \tag{5}$$

A a_i je zlomek hran mezi komunitami:

$$a_u = \frac{k_u}{2m} = \sum_{v} e_{uv} \tag{6}$$

Modularita nabývá hodnot od -0,5 do 1, přičemž vyšší modularita značí vyšší kvalitu komunit ve smyslu hustého propojení hranami uvnitř shluků a řídkého propojení mezi shluky.

Ve vytvořeném balíčku bipartiteOTU je implementováno 6 detekčních algoritmů knihovny igraph. Jedná se o metody mezilehlost hran (z anglického edge betweenness), propagace značky (z angl. label propagation), rychlého rozvíjení komunit (z angl. fast community unfolding), metody bloudění (v angl. walktrap community), vlastního vektoru (z angl. leading eigenvector) a optimálního rozložení komunit. [65]

Mezilehlost hran

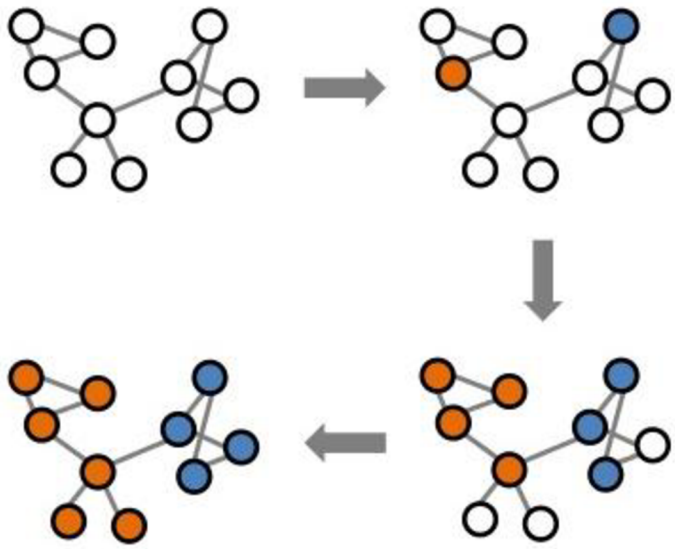
Metoda se zaměřuje na hrany grafu a vytváří komunity postupným odstraňováním hran propojujících vysoký počet relací. Je zavedena vlastnost centralita mezilehlosti $C_B(e)$ charakterizující počet nejkratších cest $\sigma_{st}(e)$ vedoucích skrz hranu e:

$$C_B(e) = \sum_{s \neq t}\frac{\sigma_{st}(e)}{\sigma_{st}} \tag{7}$$

Hrany s vysokým skóre $C_B(e)$ jsou odstraňovány jako první a celkový počet odstraněných hran, a tím i počet vzniklých komunit, je optimalizován funkcí modularity [52]. Metoda také umožňuje využití váhy hran pro důkladnější analýzu vazeb v grafu.

Propagace značky

S odlišným přístupem k detekci komunit přichází metoda propagace značky. Proces je zahájen přiřazením značky každému z uzlů grafu, přičemž na začátku je značka každého uzlu odlišná. Poté jsou náhodně vybírány uzly grafu a jejich značka je přepisována na značku vyskytující se mezi sousedními uzly (uzly spojenými hranou) nejvíce. V případě remízy je mezi nejvýše zastoupenými značkami vybráno náhodně. Tento krok rozšiřování značek se opakuje, dokud všechny uzly nemají značku, která je zároveň nejvíce zastoupená mezi sousedy daného uzlu (viz obr. 26). Algoritmus implementovaný v knihovně igraph na závěr vypočte skóre modularity pro porovnání dosažené kvality shluků. Původní metoda popsaná v [53] je zde rozšířena o možnost pracovat s váhováním hran a nastavením pevně daných značek.

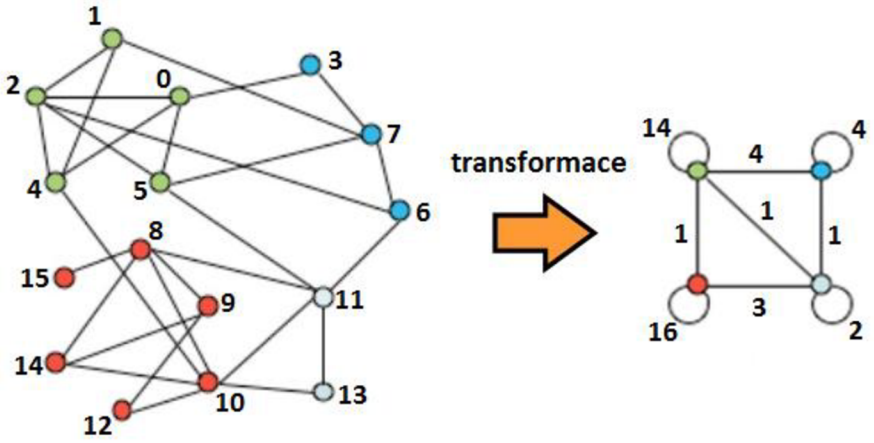


Obr. 26: Princip detekce komunit pomocí metody propagace značky. [63]

Rychlé rozvíjení komunit

Metoda rychlého rozvíjení funguje na podobném principu jako zmíněná metoda propagace značky. Stejným způsobem jako u předešlé metody i zde je každému shluku přiřazována charakteristická značka, avšak s tím rozdílem, že před každou změnou značky uzlu je nejprve spočítána modularita grafu. Uzel získává značku pouze v případech, které vedou k navýšení modularity. Tento proces se opakuje, dokud nejsou vyčerpány všechny kroky vedoucí ke zvýšení modularity. Všechny komunity jsou poté transformovány do tzv. superuzlů zastupujících dané komunity (viz obr. 27). Váhy hran mezi superuzly jsou dány počtem hran

mezi komunitami v původním grafu. V nově získaném grafu jsou uzly opět porovnávány se svými sousedy a v případě navýšení modularity sloučeny do jediné komunity. [52]



Obr. 27: Princip detekce komunit pomocí algoritmu rychlého rozvíjení. [64]

Metoda bloudění

Pascal Pons a Matthieu Latapy popsali v [54] další metodu detekce komunit, tentokrát založenou na náhodném pohybu po hranách grafu. Předpokladem metody bloudění je skutečnost, že náhodné krátké přesuny bodu po hranách grafu mají tendence zůstávat uvnitř komunit. Iteracemi je tak získána vzdálenost mezi jednotlivými vrcholy, která může být využita pro hierarchickou detekci komunit. Váhování hran může určovat pravděpodobnost, s jakou je daná hrana grafu vybrána pro přesun, a tím sloužit k zevrubnější analýze komunit.

Metoda vlastního vektoru

Na rozdíl od výše zmíněných heuristických postupů, hledání komunit pomocí metody vlastního vektoru přistupuje k problému přímou optimalizací matice modularity B výpočtem hlavního vlastního vektoru:

$$B_{i,j} = A_{i,j} - \frac{k_i k_j}{2m}, \tag{8}$$

kde A je matice sousedností a m počet hran grafu a k_i a k_j značí počet stupňů uzlu i a j.

Algoritmus pracuje ve třech krocích: [55]

1. Výpočet hlavního vlastního vektoru matice modularity.
2. Uzly, které odpovídají pozitivním hodnotám vlastního vektoru jsou přiřazeny do jedné komunity. Zbylé uzly jsou zařazeny do druhé komunity.
3. Krok dvě se opakuje pro každou komunitu znovu, dokud jsou ve vlastním vektoru přítomny pozitivní hodnoty.

Metoda optimálního rozložení komunit

Implementovaná metoda optimálních shluků vrací optimální rozložení komunit v závislosti na maximalizaci funkce modularity. Metoda zkouší všechny možné kombinace a vybírá z nich tu, která je modulární funkcí hodnocena jako nejkvalitnější. Výpočetní nároky této metody jsou vysoké, ale jak autoři igraph funkce uvádějí, pro grafy o velikosti do 50 uzlů by měl výpočet proběhnout rychle. [65]

Projekce bipartitního grafu

Úskalím výše zmíněných detekčních metod knihovny igraph je jejich zkreslená detekce při aplikaci na bipartitní grafy. Příčinu snížené funkčnosti můžeme hledat buď přímo v charakteru metod nebo také ve využití funkce modularity jako optimalizačního kritéria. Pokud chceme využívat metody detekce komunit knihovny igraph, je žádoucí převést bipartitní grafy do takzvaných projekcí bipartitních grafů. Guimera a kolektiv [66] nenašli žádný rozdíl mezi detekovanými komunitami při použití maximalizace modularity u projekce bipartijního grafu a maximalizace bimodularity (alternativy funkce modularity pro bipartitní grafy) u bipartitních grafů. Funkce bimodularity v současné době není u igraph funkcí dostupná.

Nazveme-li jednu partitu bipartitního grafu P a druhou S, pak projekce partity P grafu $G = (P \vee S, E)$ je graf $G_P = (P, E_P)$, kde dva uzly i a $j \in P$ jsou spojeny hranou E_p, pokud mají alespoň jeden společný sousedící uzel v S. Projekce mohou být váhované i neváhované, avšak váhované projekce jsou obvykle považovány za více reprezentativní [67]. V balíčku bipartiteOTU je využito jednoduchého váhování hran, kde váha hrany mezi dvěma uzly odpovídá počtu jejich společných sousedů.

Matice sousedností G_P pro daný graf G je definována jako:

$$A_{i,j} = \begin{cases} 1, \text{když uzel } i \text{ a } j \text{ mají společný sousední uzel} \\ 1, \text{když má uzel } i \text{ sousední uzel, který nemá žádné další sousedy v } P \\ 0 \text{ ve všech dalších případech} \end{cases}$$

Váha hran $W_{i,j}$ ve váhované projekci je pak dána vztahem 9, kde $\Gamma(i)$ značí množinu sousedů uzlu i:

$$W_{i,j} = |\Gamma(i) \cap \Gamma(j)|, \qquad i \neq j \tag{9}$$

Takto získané váhované projekce mohou zastoupit původní bipartitní grafy a být vhodným vstupem zmíněných algoritmů detekce komunit knihovny igraph.

6.3 Charakteristiky grafu

Vlastnost *community* definující příslušnost uzlů do konkrétních komunit není jedinou charakteristikou grafu, kterou můžeme získat pomocí navrženého balíčku bipartiteOTU.

Velmi důležitou vlastností zmíněnou již v předešlých kapitolách, která může do značné míry ovlivnit i detekované komunity, je váhování hran bipartitního grafu a jeho projekcí. Při tvorbě grafů nabízí balíček možnost volby, zda chceme využít váhování hran, či nikoliv. V případě nevyužití váhování hran *E* je kalkulováno s váhou definovanou jako:

$$E_{i,j} = \begin{cases} 1, \text{když uzel } i \text{ a } j \text{ mají společný sousední uzel} \\ 0 \text{ ve všech dalších případech} \end{cases}$$

Při využití váhování je zavedena vlastnost *weight* odvozená od počtu čtení v OTU tabulce P:

$$E_{i,j} = 10 \frac{P_{i,j}}{\max(P)}, \tag{10}$$

Váhování hran projekcí bipartitního grafu je pak dáno vztahem 9 popsaným v kapitole 6.2.

Balíček při tvorbě grafů rovněž automaticky zavádí vlastnost uzlů *type* popisující, ke které partitě bipartitního grafu uzel náleží, a dále vlastnost *degree* udávající počet hran vycházejících z uzlu. Vlastnost *degree* pak může být využita při vykreslování grafu pro optimalizaci velikosti uzlů.

6.4 Doplňkové funkce pro práci s grafy

Balíček nabízí doplňkové funkce pro práci s grafy. Umožňuje jak vykreslení grafu přímo v R, tak také obsahuje funkci *save_as_gml,* která převede získaný graf do GML formátu. Tento formát je podporován řadou vizualizačních programů, do nichž tak může být snadno importován. Na obr. 28 je ukázka vizualizace pomocí Gephi softwaru.

Graf může také být pro další výpočty převeden do matice sousedností pomocí funkce *adjacency_matrix.* Pro úpravu automaticky vygenerovaných popisků softwarem QIIME slouží funkce *name_adjust.* Tato funkce odstraňuje z popisků návěstí a přejmenovává dlouhé názvy na kratší. Příklady takové transformace jsou uvedeny v tabulce 1.

Tabulka 1: Transformace popisků

Původní popisek	Nový popisek
p_Firmicutes	Firmicutes
c_Bacteroidia	Bacteroidia
New.ReferenceOTU15639	New15639
New.CleanUp.ReferenceOTU746	CleanUp746

Přejmenování popisků usnadňuje čitelnost výsledného grafu.

7 Analýza

V 6. kapitole byly popsány základní vlastnosti grafu a mechanismy implementované v balíčku bipartiteOTU. V této kapitole jsou zhodnoceny vlastnosti balíčku, především pak jeho redukční schopnosti a schopnosti detekovat hlavní komunity.

Ve druhé části této kapitoly jsou porovnány tři hlavní přístupy k procesu OTU picking a možnosti jejich nastavení.

7.1 Analýza funkcí balíčku bipartiteOTU

Jako vstupní data pro porovnání redukčních schopností balíčku byl použit dataset mikrobiomů vyizolovaných z vody využívané k přepravě ryb. Dataset byl publikován v [56]. Jednotlivá čtení byla získána pyrosekvenováním V3/V4 variabilních regionů 16S rRNA genu. Data byla dále prostřednictvím platformy QIIME a výpočetních serverů MetaCentra zpracována *de novo* OTU picking metodou. Jako shlukovací algoritmus byl volen uclust a prahem podobnosti pro přiřazení dvou OTU do shluku byla volena konvenční hranice 97 %. Výsledné shluky byly dodatečně porovnány s referenční databází Greengenes 13_8, aby bylo možno využít taxonomické informace k redukci dat. V případě srovnávání již vytvořených shluků s referenční databází za účelem taxonomického zařazení byla vždy vybrána jedna reprezentativní sekvence a ta byla porovnávána s databází na prahu podobnosti 90 %. Výsledné taxonomické určení je pak průnikem taxonomie tří OTU z referenční databáze, které vykazovaly nejvyšší podobnosti s porovnávanou sekvencí.

Redukce dat na základě taxonomické kategorie

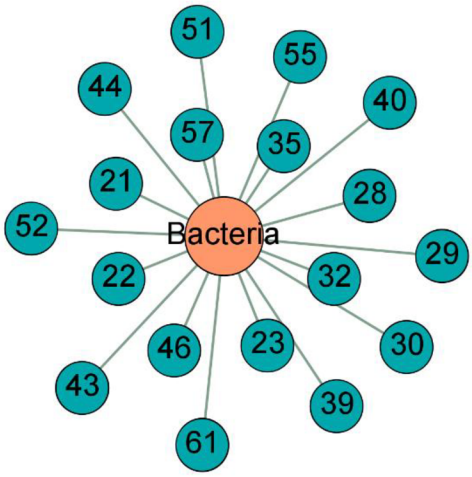
Tabulka 2 ukazuje výsledky redukce dat na základě taxonomického určení popsaného v kapitole 6.1. Pro její sestavení byla funkce *merge_taxonomy* spuštěna s nastavením NewOTU = FALSE, tedy všechny OTU, u kterých nebyla definována požadovaná taxonomie, byly z další analýzy vyřazeny. V tabulce 2 je počet takto vyřazených OTU značen jako nedefinované OTU.

Tabulka 2: Výsledky redukce dat na základě taxonomického určení.

	shluky	Taxonomická kategorie						
		říše	kmen	třída	řád	čeleď	rod	druh
Počet uzlů	4167	19	37	56	81	131	203	69
Počet hran	6453	18	102	239	454	670	798	172
Průměrný stupeň uzlů	3,097	1,895	5,514	8,536	11,210	10,229	7,862	4,986
Denzita grafu	0,001	0,211	0,306	0,310	0,280	0,157	0,078	0,147
Počet čtení	56661	54973	54973	54946	54792	52664	35625	5042
Nedefinované OTU	921	921	921	936	1011	1325	2600	3939

Je patrný trend nárůstu počtu vyřazených OTU a poklesu celkových zpracovaných čtení s volbou konkrétnější taxonomické kategorie. Při taxonomické kategorii druhů bylo do grafu zahrnuto pouze 5042 čtení z původních 56661. Výrazný pokles zpracovaných sekvencí se projevil také na počtu uzlů. Zde je dobré říci, že analýza byla provedena na celkem 18 vzorcích vody, kde je každý reprezentován jedním uzlem.

Na obr. 28, obr. 29 a obr. 30 jsou znázorněny bipartitní grafy popsané tabulkou 2. Konkrétně se jedná o bipartitní grafy říše, kmenů a druhů. Jednotlivé partity jsou odděleny barevně.

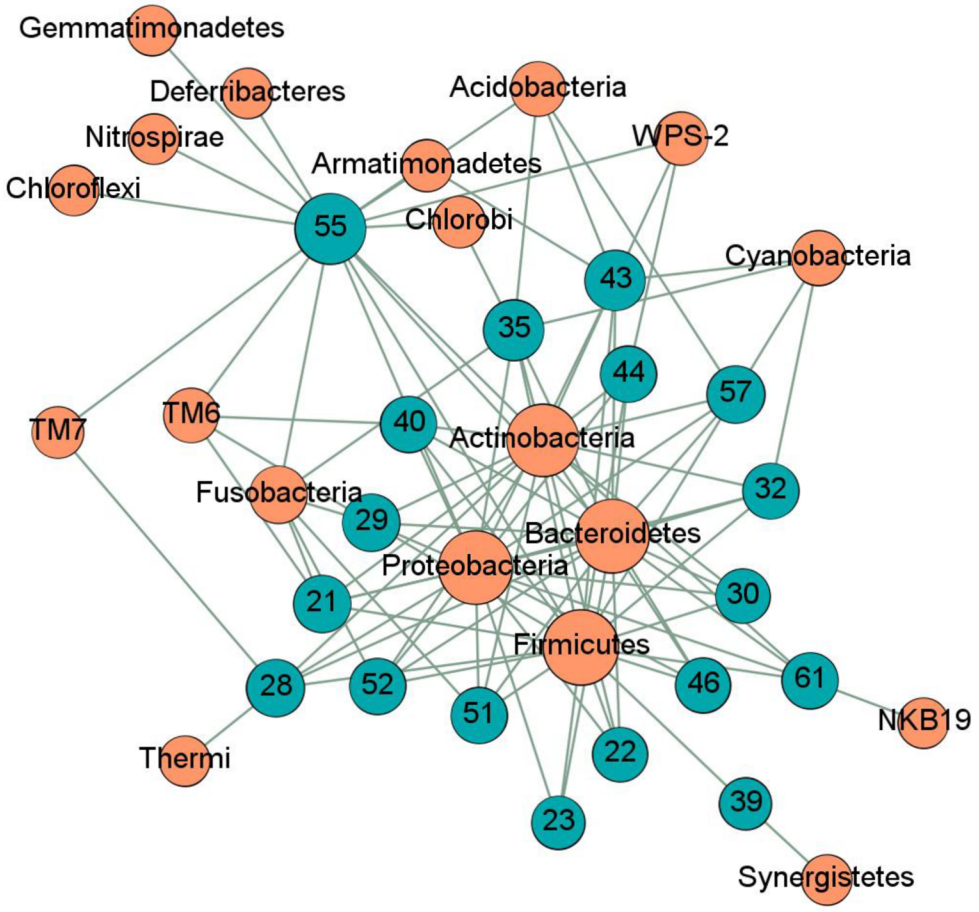


Obr. 28: Detekované taxonomické říše

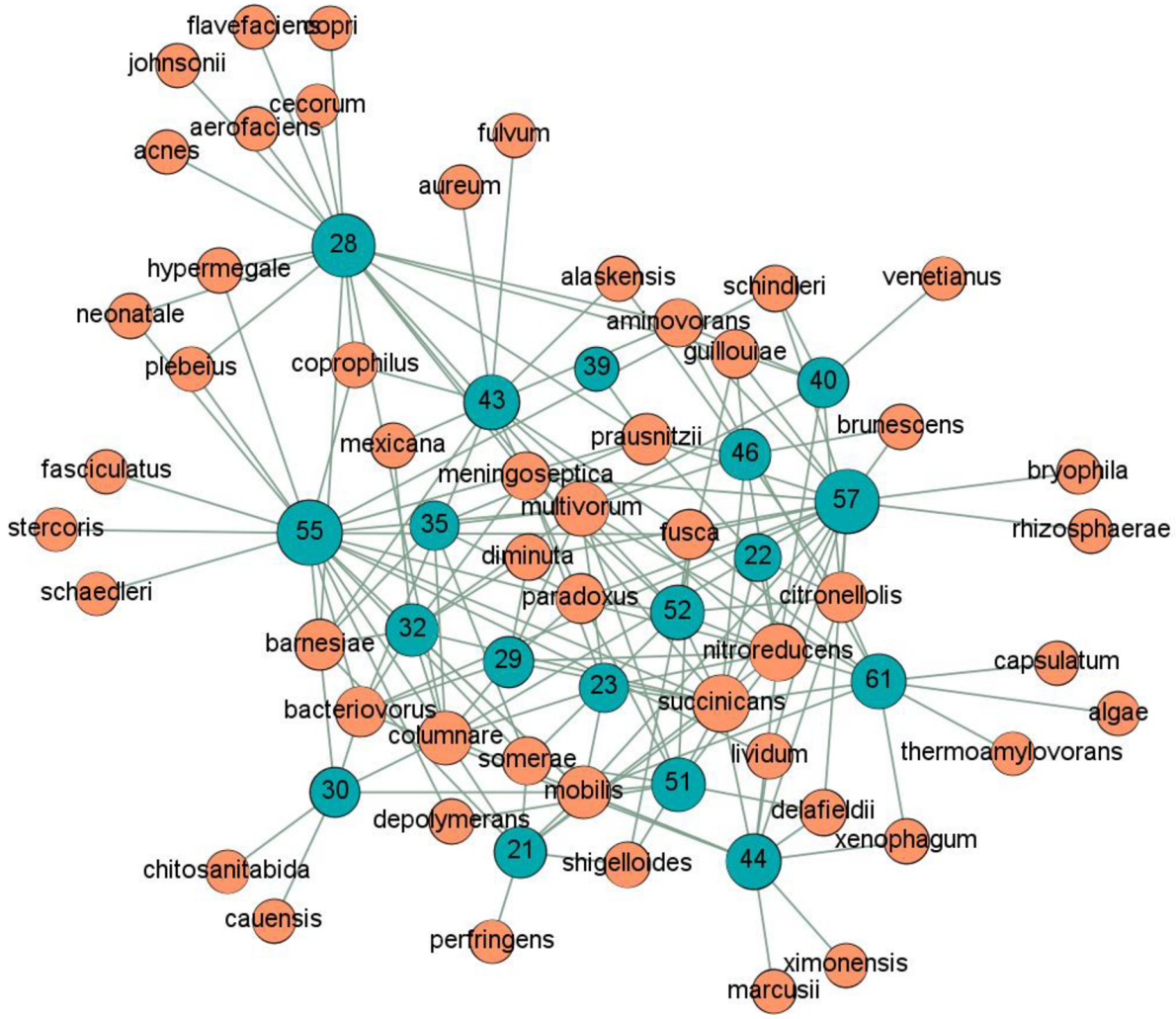
Při taxonomické redukci dat dle kategorie říše zůstal v datasetu pouze jediný uzel reprezentující OTU, a to sice uzel *Bacteria*. Žádné jiné taxonomické říše nebyly ve vzorku nalezeny. Tento jediný OTU uzel je propojen se všemi uzly z druhé partity (vzorky izolátů vody značené čísly), a tím zvyšuje průměrný stupeň uzlů a udává denzitu grafu. V grafu je patrné, že uzel je větší než ostatní. Zde bylo využito váhování uzlů – uzly s vyšším stupněm jsou větší.

Redukce podle kmenů je na obr. 29. Došlo k nárůstu uzlů reprezentujících OTU. Všechny vykreslené OTU samozřejmě spadají pod *Bacteria,* neboť to byl jediný uzel získaný zpracováním dat s redukcí podle říše. Zástupci kmenů *Bacteroidetes*, *Proteobacteria*, *Firmicutes* a *Actinobacteria* byli přítomni takřka ve všech vzorcích (kmen *Firmicutes* ve všech vzorcích, kmeny *Proteobacteria* a *Bacteroidetes* nebyly přítomny pouze ve vzorku 39 a zástupci kmene *Actinobacteria* nebyli nalezeni ve vzorcích 39 a 23). Zdá se, že tyto kmeny tvoří jakési jádro, a i v bipartitním grafu jsou znázorněny velkými uzly lokalizovanými zejména ve středu grafu. Obklopeny jsou pak uzly vzorků, ve kterých byly zaznamenány. Kmeny více specifické pro vzorky se nacházejí na okrajích grafu. Velikost jejich uzlu je menší. Váhování uzlů také napovídá, že největší kmenová diverzita byla zaznamenána u vzorku 55.

Pokud srovnáme grafy znázorňující taxonomické kmeny (obr. 29) a druhy (obr. 30), vidíme, že u grafu taxonomických druhů došlo k nárůstu počtu uzlů a hran. Ačkoliv je jeho denzita, průměrný stupeň uzlu a celkový počet zpracovaných čtení nižší, je graf právě kvůli velkému počtu uzlů a hran méně přehledný než graf taxonomických kmenů. V případě méně přehledných grafů může být výhodné použití interaktivních vizualizačních programů (např. Gephi) nebo kombinace většího počtu redukčních technik.



Obr. 29: Detekované taxonomické kmeny



Obr. 30: Detekované taxonomické druhy

Redukce dat pomocí prahové hodnoty

Porovnávání počtu pozorování (počtu čtení) každého organismu se zadaným prahem se ukazuje jako další efektivní metoda předzpracování dat (viz tabulka 3). Již při prahové hodnotě 5 byl celkový počet uzlů zredukován z 4167 na 538, a to při odstranění 8128 čtení. Průměrný počet pozorování u odstraněných OTU tak byl 2,2 čtení. Prudce klesl také počet OTU, u kterých nemohlo být definováno taxonomické zařazení. Konkrétně z 921 OTU na 25. Při nastavení prahové hodnoty na 100 čtení zbylo již jen poslední OTU s neznámou taxonomií, které se zařadilo mezi 52 nejčetněji zastoupených OTU. Jeho celkový počet čtení byl 107.

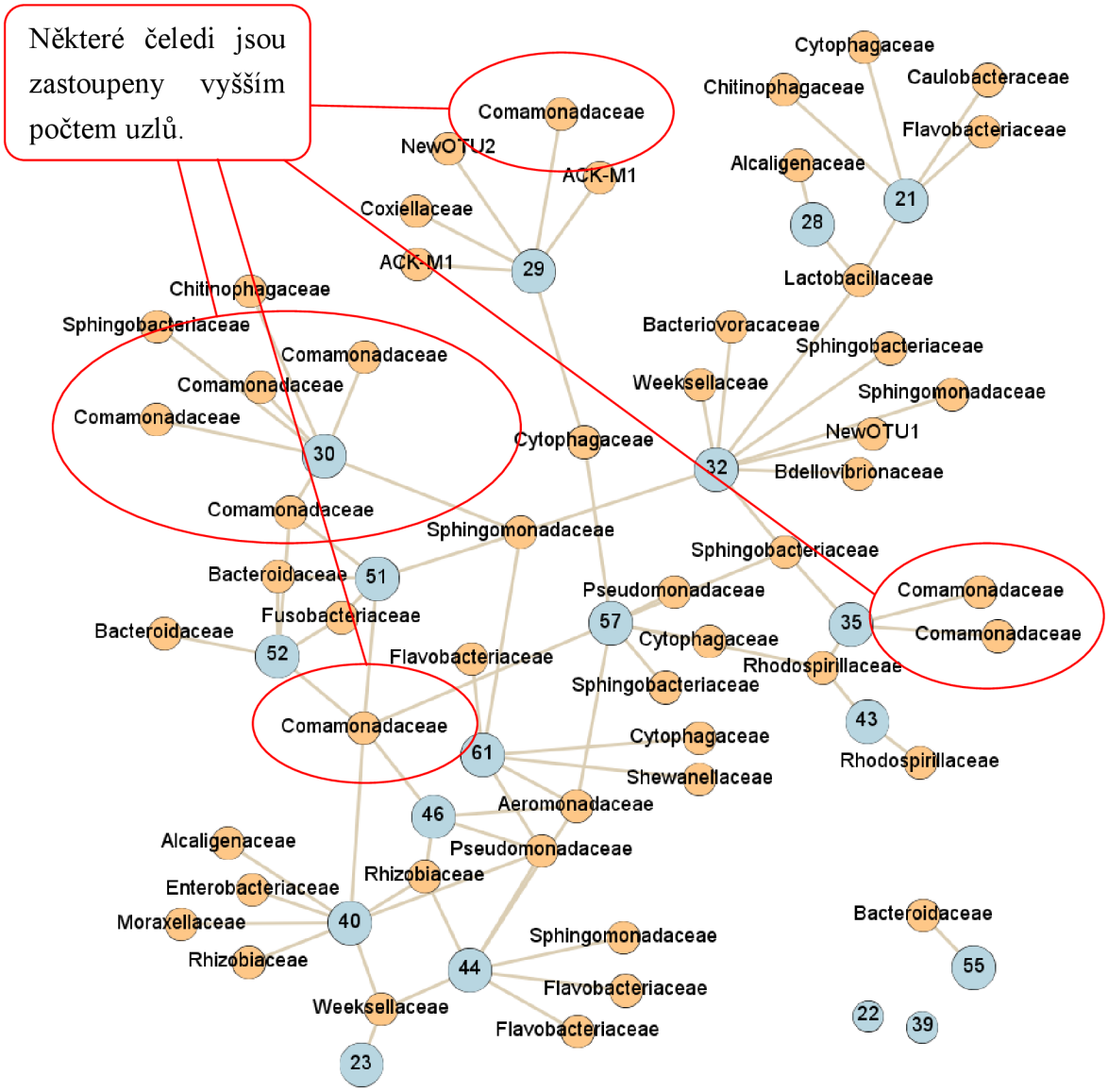
Tabulka 3: Výsledky redukce dat nastavením prahové hodnoty

Prahová hodnota	0	5	10	50	100
Počet uzlů	4167	538	312	106	70
Počet hran	6453	1068	620	160	80
Průměrný stupeň uzlů	3,097	3,970	3,974	3,019	2,286
Denzita grafu	0,001	0,015	0,026	0,058	0,066
Počet čtení	56661	48533	45650	35847	30024
Nedefinované OTU	921	25	12	1	1

Obr. 31 ukazuje data získaná redukcí původního datasetu porovnáním s prahovou hodnotou 100 čtení. Vykresleny jsou zastoupené čeledi. Červeně jsou zvýrazněny uzly reprezentující čeleď *Comamonadaceae*. Tato čeleď není jediná vyskytující se v grafu vícekrát, dále také *Sphingobacteriaceae*, *Cytophagaceae*, *Flavobacteriaceae* a další. Zpracováním takovýchto uzlů do jednoho se zabývá výše zmíněný přístup redukce na základě taxonomie, která tento problém dokáže odstranit.

Příčina výskytu uzlů se stejným názvem může být důsledkem jednoho z následujících:

- Uzly jsou totožné ve zvolené taxonomické kategorii (v tomto případě kategorii čeledí), avšak liší se v některé z podkategorií (v tomto případě kategorie rodu nebo druhu).
- Jedinému organismu může být v referenční databázi přiřazeno více sekvencí. Vysvětlení pro tento jev najdeme v historii taxonomického přiřazování odvíjející se nikoliv od studia genetické informace, ale od morfologických a chemických vlastností organismu. Hranice 97% podobnosti tak není ekvivalentem žádné specifické taxonomické kategorie.
- Gen 16S rRNA nemusí spolehlivě diferencovat všechny organismy. Porovnáváním 16S rRNA genů s referenční databází můžeme u více sekvencí v referenční databázi získat shodu 100 %. Sekvence mohou být shodné v 16S rRNA genu, ale mohou se lišit v jiných částech genetické informace. Přiřazená taxonomie pak odpovídá průniku taxonomického určení referenčních sekvencí. Taxonomické kategorie, ve kterých se referenční sekvence liší, zůstávají u analyzované sekvence nevyplněny.



Obr. 31: Detekované čeledi předzpracované porovnáním s prahovou hodnotou 100.

V pravém dolním rohu obrázku 31 jsou zaneseny vzorky 22 a 39, u kterých po zpracování nebyly zjištěny žádné organismy (všechny OTU byly pozorovány méně než stokrát). U vzorku 55 byla po zpracování objevena pouze jediná čeleď, a to *Bacteroidaceae*.

Se vzorky 29 a 32 jsou propojeny dvě nedefinované OTU nazvané jako *NewOTU1* a *NewOTU2*. Ačkoliv se dvě nedefinované OTU zdají v rozporu s tabulkou 3, jejich výskyt je oprávněný. Musíme si uvědomit, že v grafu jsou vykresleny všechny získané čeledi, a tak i OTU, u kterých jsou vyšší taxonomické kategorie známé, avšak chybí informace o příslušném zařazení do čeledí, a jsou tedy označeny jako NewOTU. Zatímco u NewOTU1 nemáme žádnou informaci o taxonomii (OTU nemohla být na hranici 90% podobnosti přiřazena k žádné referenční sekvenci), u NewOTU2 je znám rod *Bacteria*, kmen *Bacteroidetes*, třída

Sphingobacteriia a řád *Sphingobacteriales*. Tato informace je však převedením do grafu pozorovaných čeledí ztracena.

Redukce na základě informace uložené v metadatech

Doposud jsme redukovali pouze počet uzlů z partity zastupující OTU. Je nasnadě zaměřit se také na druhou partitu grafu. Dataset obsahoval následující metadata popsaná v tabulce 4:

Tabulka 4: Dostupná metadata

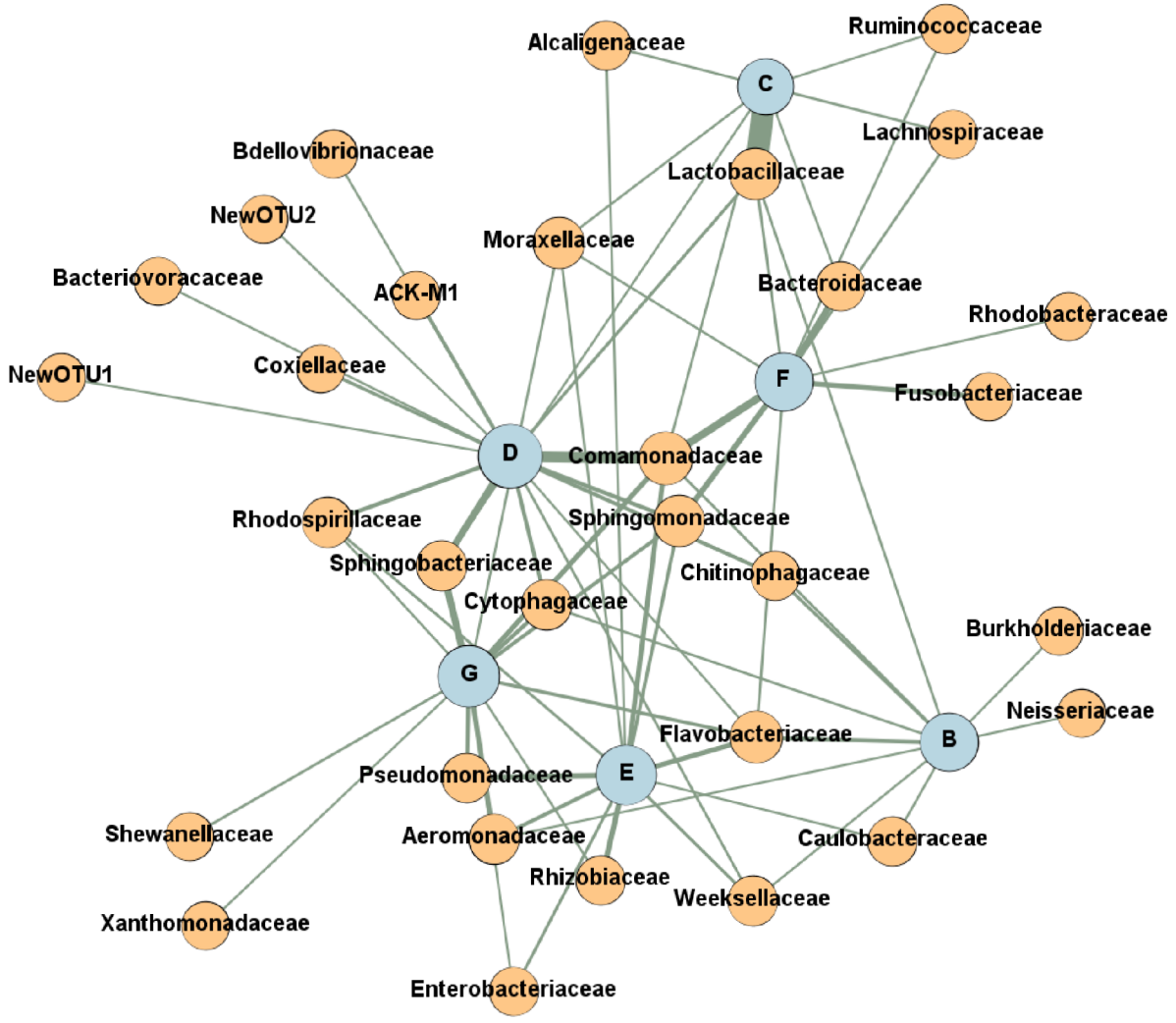
SampleID	Dodavatel	Země Původu	Description
21	B	Vietnam	Paracheirodon_axelrodi
22	B	Vietnam	Poecilia_sphenops
23	B	Vietnam	Hypostomus_plecostomus
29	D	Peru	Otocinclus_affinis
32	D	Peru	Otocinclus_affinis
44	E	Vietnam	Plecostomus_Gold
46	E	Vietnam	Gold_Black_Molla
55	F	Vietnam	V4
51	F	Singapur	Colisa_lalia_blood_red_K574,_samecci
52	F	Singapur	Colisa_lalia_blood_red_K574,_samicky
28	C	Peru	Panaqolus_changae
43	E	Thajsko	Poecilia_sphenops
30	D	Peru	Surubim_Lima
40	E	Vietnam	Sewelia
61	G	Vietnam	Xiphophorus_maculatus
35	D	Hong_Kong	Pseudotropheus_Cheni
39	E	Čína	Xenopus_laevis_Albin
57	G	Vietnam	Gyrinocheilus

V prvním sloupečku vidíme identifikátor vzorku, ve druhém typ dodavatele, ve třetím zemi původu a ve čtvrtém druh přepravované ryby. Z tabulky byla vypuštěna informace o barcodové sekvenci, sekvenci primeru a sekvenci reversního primeru. V metadatech byla samozřejmě přítomna tak, jak ukazuje vzor na obr. 13.

Obrázek 32 porovnává jednotlivé dodavatele ryb mezi sebou. Data byla zpracována do čeledí a dále redukována prahovou hodnotou 100 čtení. Do jednoho uzlu nazvaného B byly dále dle tabulky 4 shluknuty vzorky 21, 22 a 23. Podobným způsobem byly vytvořeny i uzly C, D, E, F a G.

Graf se stal přehlednějším než graf uvedený na obr. 31. Zároveň, ačkoliv bylo použito více redukčních technik, zůstal vyšší počet čtení (viz tabulka 5). Shlukováním uzlů na základě taxonomie totiž stoupla abundance uzlů a redukce pomocí prahové hodnoty neodstranila tolik

čtení. Opačná volba pořadí těchto dvou technik by vedla k nižšímu počtu čtení (konkrétně 30024, jak bylo uvedeno v tabulce 3). Vzrostla také denzita grafu a průměrný stupeň uzlu.



Obr. 32: Bipartitní graf porovnávající dodavatele. Data byla zpracována redukcí podle abundance čeledí 100 a vyšší.

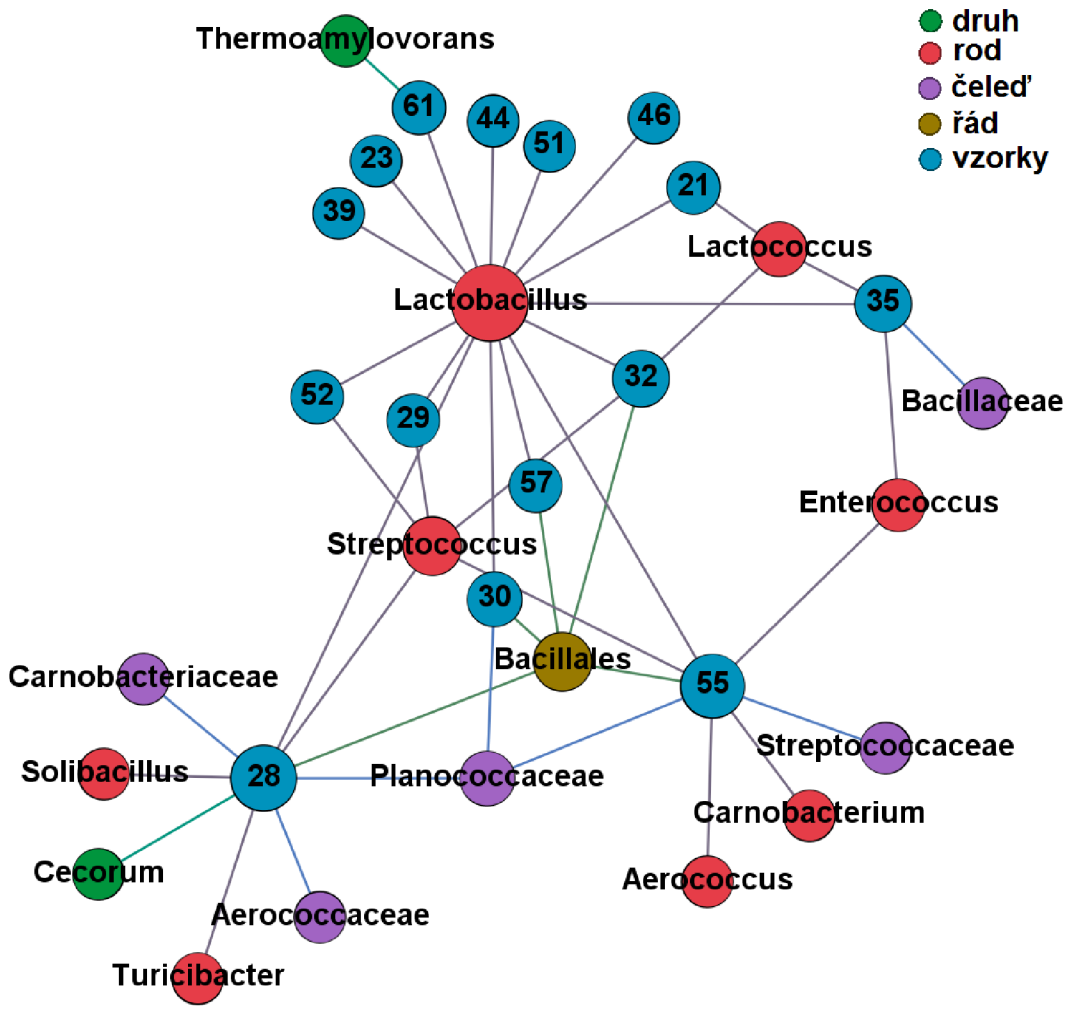
U grafu na obr. 32 bylo také použito váhování hran. Vidíme, že ačkoliv byla čeleď *Lactobacillaceae* přítomna ve vodě dodavatelů B, C, D i F, pro dodavatele C byla nejcharakterističtější. Četnost organismu *Lactobacillaceae* ve vodě dodavatele C byla mnohem vyšší než u všech vzorků dodavatele D, F nebo B dohromady. Vzorky dodavatele D byly lépe charakterizovány čeledí *Comamonadaceae* a *Sphingobacteriaceae*. Voda od dodavatele D obsahovala také mnoho specifických čeledí, které nebyly nalezeny u žádných jiných dodavatelů, např.: *Coxiellaceae*.

Tabulka 5: Parametry grafu získaného souborem tří redukčních funkcí

Počet uzlů	37
Počet hran	64
Průměrný stupeň uzlů	3,460
Denzita grafu	0,192
Počet čtení	42146
Nedefinované OTU	2

Zaměření pouze na specifické organismy

Další možností k získání přehlednějších bipartitních grafů je zaměření se pouze na určité organismy pomocí funkce *taxonomy_focus*. Na obrázku 33 je graf, ve kterém jsou uvedeny všechny organismy spadající do třídy *Bacilli*. Pro vykreslení do grafu byla volena nejspecifičtější známá taxonomická kategorie dané OTU. V grafu jsou tak zastoupeny druhy, rody, čeledi i řády. Některé vzorky neobsahovaly žádné organismy třídy *Bacilli* a nejsou tak v grafu vůbec zaznačeny. Parametry grafu jsou uvedeny v tabulce 6.



Obr. 33: Detekované organismy spadající do třídy *Bacilli*

Tabulka 6: Parametry grafu zaměřeného na třídu Bacilli

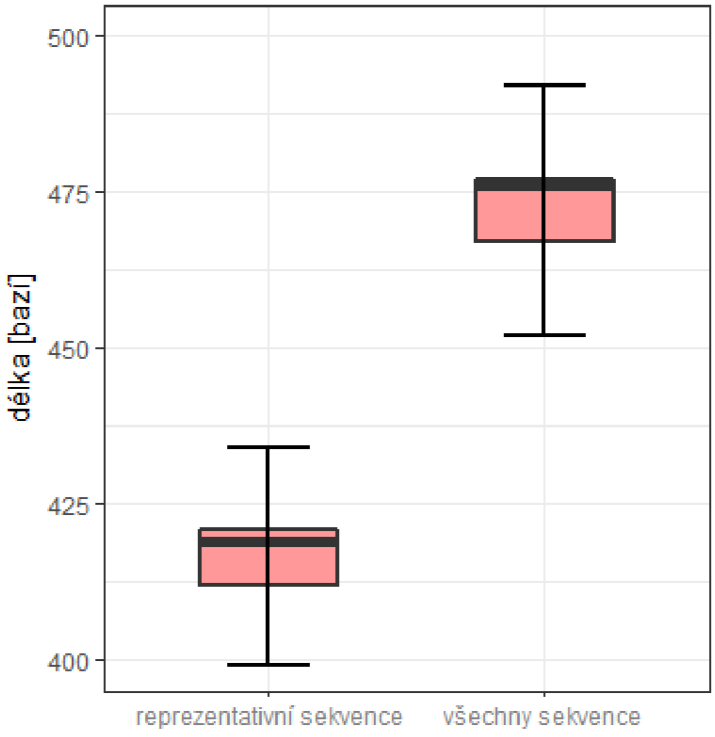
Počet uzlů	31
Počet hran	43
Průměrný stupeň uzlů	2,774
Denzita grafu	0,185
Počet čtení	8555
Nedefinované OTU	0

Práh podobnosti

V úvodu této kapitoly jsme si zmínili práh podobnosti 97 % pro přiřazení dvou OTU do jednoho shluku a 90 % pro přiřazení taxonomie. Pro lepší představu kolik je to vlastně bází, se podívejme na tabulku 7 a obr. 34 charakterizující délku vstupních sekvencí.

Tabulka 7: Rozložení délky analyzovaných sekvencí

	Minimum	1. kvartil	Median	Průměr	3. kvartil	Maximum
Všechny sekvence	43	467	476	467,7	477	998
Reprezentativní sekvence	189	412	419	418,2	421	559



Obr. 34: Krabicový graf rozložení délky sekvencí

Uvažujme medián všech sekvencí jako délku sekvencí, které chceme sloučit do jednoho OTU. Použijeme-li práh podobnosti 97 %, zjistíme, že sekvence se od sebe mohou lišit maximálně o 14 bází, jinak nebudou zařazeny do jednoho OTU. Podobně uvažujeme-li medián délky

reprezentativních sekvencí jako délku sekvence, u které chceme určit její taxonomii při prahu podobnosti 90 %, zjistíme, že se sekvence musejí shodovat minimálně v 378 bázích z 419.

7.2 Analýza algoritmů detekce komunit

V této kapitole se zaměříme na porovnání algoritmů pro detekci komunit zmíněných v kapitole 6.2. Data pro všechny analýzy v této kapitole byla stejná jako v kapitole 7.1, avšak byla již předzpracována redukcí podle čeledí a následným porovnáním s prahem 100 čtení.

Nejprve byly komunity hledány v partitě zastupující OTU. Tabulka 8 uvádí získané výsledky.

Tabulka 8: Výsledky detekce komunit mezi OTU

	Váhovaná projekce		Neváhovaná projekce	
Detekční metoda	Modularita	Počet komunit	Modularita	Počet komunit
Metoda bloudění	0,156	4	0,150	4
Hlavní vlastní vektor	0,155	3	0,159	6
Propagace značky	0	1	0	1
Mezilehlost hran	0,038	12	0,027	18
Rychlé rozvíjení	0,176	3	0,198	4
Optimální rozložení	0,178	4	0,198	4

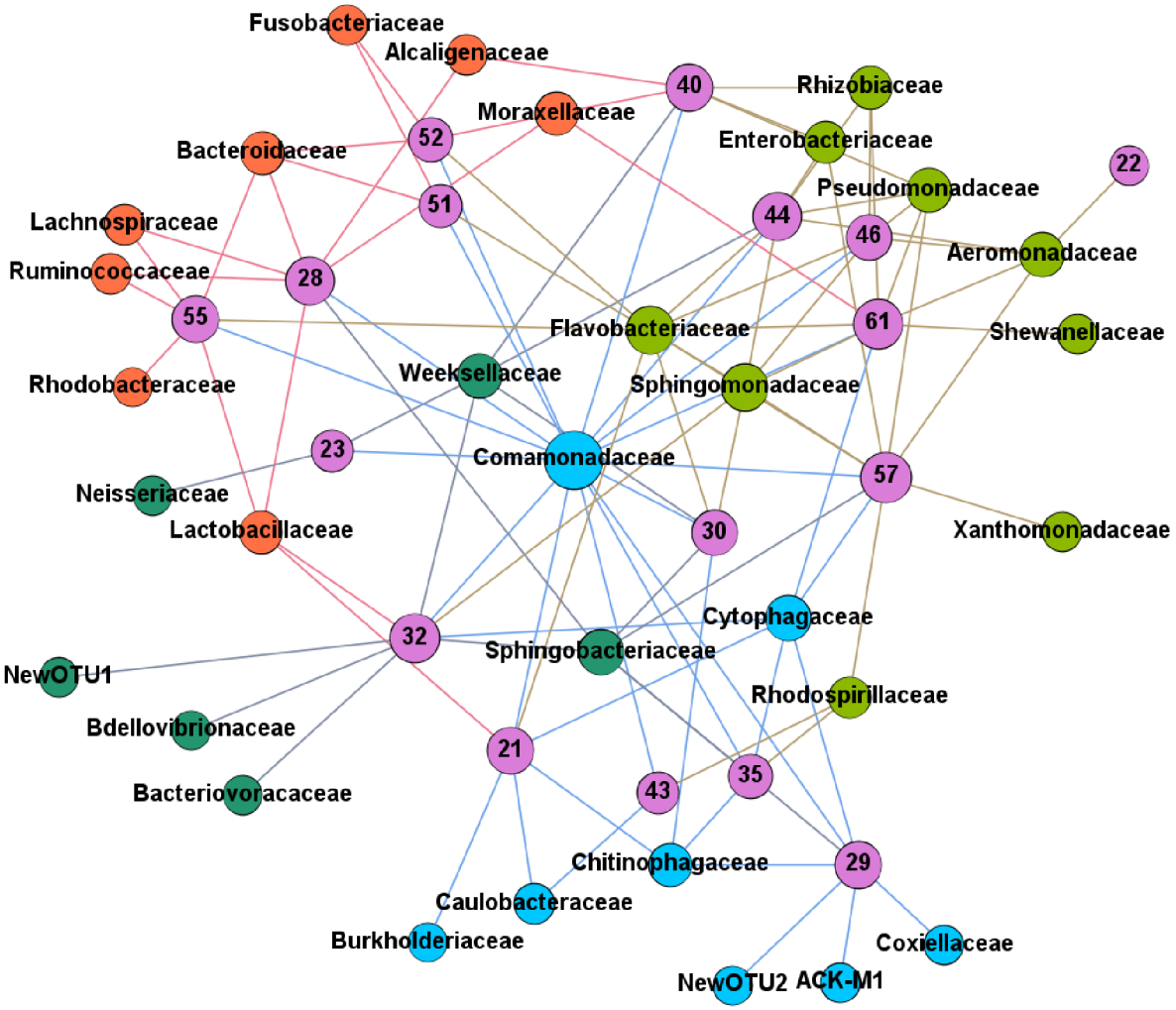
Vzhledem k nízkému počtu uzlů v této partitě mohlo být vypočteno také optimální rozdělení komunit ve smyslu maximalizace modulární funkce. Nejvyšší modulární funkce bylo dosaženo použitím neváhované projekce metodou rychlého rozvíjení stejně jako optimálního rozložení. Avšak odstraněním informace uložené v hranách grafu při použití neváhované projekce můžeme ztratit důležité vazby v grafu.

Váhované optimální rozložení se lišilo od neváhovaného v zařazení čeledi *Rhodospiracalleae* a *Sphingobacteriaceae*. Na obr. 35 je vykreslen graf získaný optimální neváhovanou detekcí komunit. Komunity jsou rozlišeny barevně. U váhované optimální detekce komunit byly čeledi *Rhodospiracalleae* i *Sphingobacteriaceae* přiřazeny do světle modré komunity.

Optimálnímu rozložení komunit se nejvíce přibližovala metoda rychlého rozvíjení. V případě neváhované projekce bipartitního grafu docílila stejných výsledků jako metoda optimálního rozdělení. V případě váhované projekce bylo nalezeno o jednu komunitu méně. Zatímco metoda optimálního rozdělení komunit měla tendence shlukovat vysoce specifické OTU (OTU nalezené pouze v jediném vzorku) do jedné skupiny, metoda rychlého rozvíjení je shlukovala do větších celků.

Podobných výsledků bylo dosaženo také metodou bloudění. Lišila se v odlišném zařazení čeledí *Commamonadaceae*, *Flavobactericiae*, *Neisseriaceae* a *Sphingomonadaceae*.

Naopak nízké modularity dosáhla metoda propagace značky, která všechny OTU zahrnula do jedné komunity, a také metoda mezilehlosti hran, která detekovala 12 komunit při váhované projekci a 18 při neváhované projekci. Nutno podotknout, že komunity působily značně chaotickým dojmem, a bylo detekováno mnoho komunit tvořených pouze jediným uzlem.



Obr. 35: Výsledek neváhované detekce komunit metodou optimálního rozložení.

Kromě OTU byly přiřazeny do komunit také jednotlivé odebrané vzorky. Výsledky jsou shrnuty v tabulce 9.

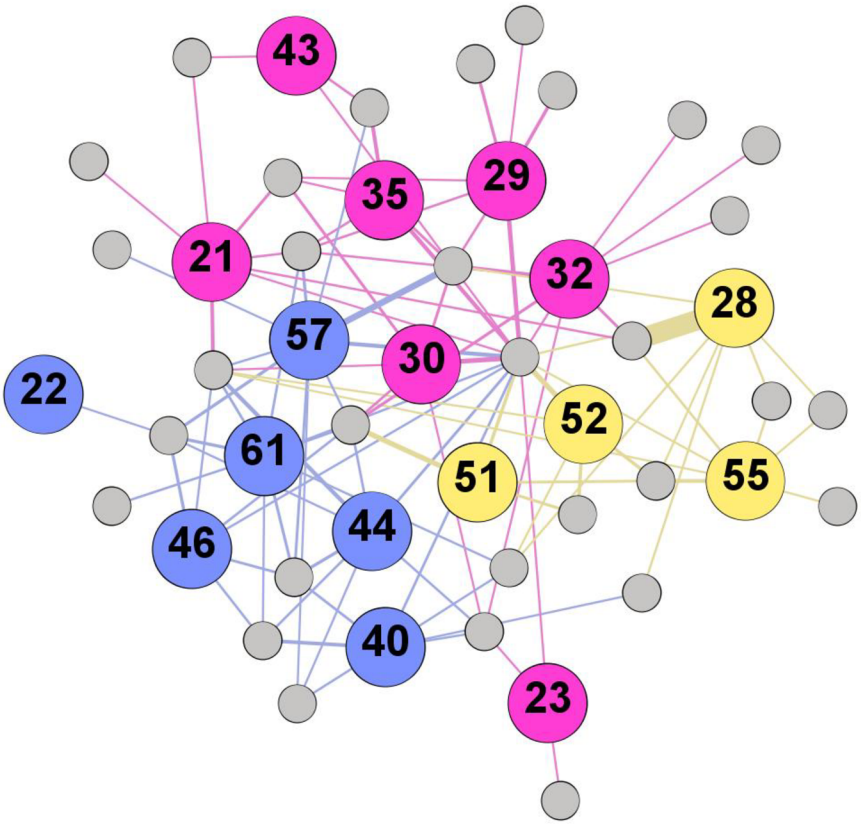
Získané výsledky skóre modularity zůstaly o poznání nižší než při detekci komunit u OTU. I zde dosáhly metody propagace značky a mezilehlost hran velmi nízkého skóre. Podobně na tom byla také metoda bloudění, která rovněž dosáhla nulové maximální modularity, ale na rozdíl od předešlých metod nezahrnula všechny shluky do jednoho, naopak

každému vzorku byla přiřazena vlastní komunita. Metoda vlastního vektoru a rychlého rozvíjení dosáhly jak u váhované, tak neváhované projekce stejných výsledků jako metoda optimálního rozložení.

Tabulka 9: Výsledky detekce komunit mezi vzorky

	Váhovaná bipartitní projekce		Neváhovaná bipartitní projekce	
Detekční metoda	Modularita	Počet komunit	Modularita	Počet komunit
Metoda bloudění	0,068	3	0,000	17
Hlavní vlastní vektor	0,086	3	0,011	2
Propagace značky	0,000	1	0,000	1
Mezilehlost hran	0,000	1	0,000	1
Rychlé rozvíjení	0,086	3	0,011	2
Optimální rozložení	0,086	3	0,011	2

Na obr. 36 je graf váhované detekce komunit optimálním rozložením. Na rozdíl od grafu na obr. 35 je zde možnost využití váhování hran. U předešlého grafu bylo váhování hran potlačeno. Byla volena také odlišná vizualizační metoda. Komunity jsou opět rozlišeny barevně, ale OTU byly vykresleny pouze menším typem uzlů. Uzly vzorků jsou všechny vykresleny ve stejné velikosti. U předešlého typu grafů byla velikost uzlů závislá na jeho stupni.



Obr. 36: Výsledek váhované detekce komunit optimálním rozložením.

Srovnáme-li výsledky z obrázku 36 s informacemi, které o vzorcích máme (uvedeny v tabulce 4), zjistíme, že všechny vzorky dodavatele F a C jsou přiřazeny do žlutě značeného shluku. Stejně tak všechny vzorky od dodavatele E pocházející z Vietnamu a vzorky od dodavatele G jsou v modrém shluku. V růžovém shluku jsou zařazeny vzorky od dodavatele D a vzorek E z Thajska. Vzorky od dodavatele B byly zařazeny do růžového i modrého shluku bez zjevných asociací.

7.3 Analýza procesu OTU picking

Navržený balíček bipartiteOTU byl využit pro analýzu hlavních parametrů procesu OTU picking. Byly porovnávány tři hlavní OTU picking metody, a to především z hlediska detekované diverzity. Dále byla testována závislost výstupních dat na nastavení hranice prahu podobnosti a volbě referenční databáze.

OTU picking přístupy

Byl použit dataset blíže popsaný v kapitole 7.1. Data byla prostřednictvím platformy QIIME a výpočetních serverů MetaCentra zpracována zvlášť closed reference, open reference a *de novo* metodou, u které byly výsledné shluky dodatečně porovnány s databází, aby bylo možno získaná data porovnat mezi sebou.

Analýzou byly zaznamenány určité rozdíly v celkové detekované mikrobiální diverzitě vzorků. Zatímco pomocí open reference OTU picking bylo nalezeno 16 mikrobiálních kmenů, přístup closed reference a *de novo* picking jich detekoval 19. Jak je patrné v tabulce 10, rozdíly ve výsledné diverzitě narůstaly s konkrétnějším taxonomickým zařazením.

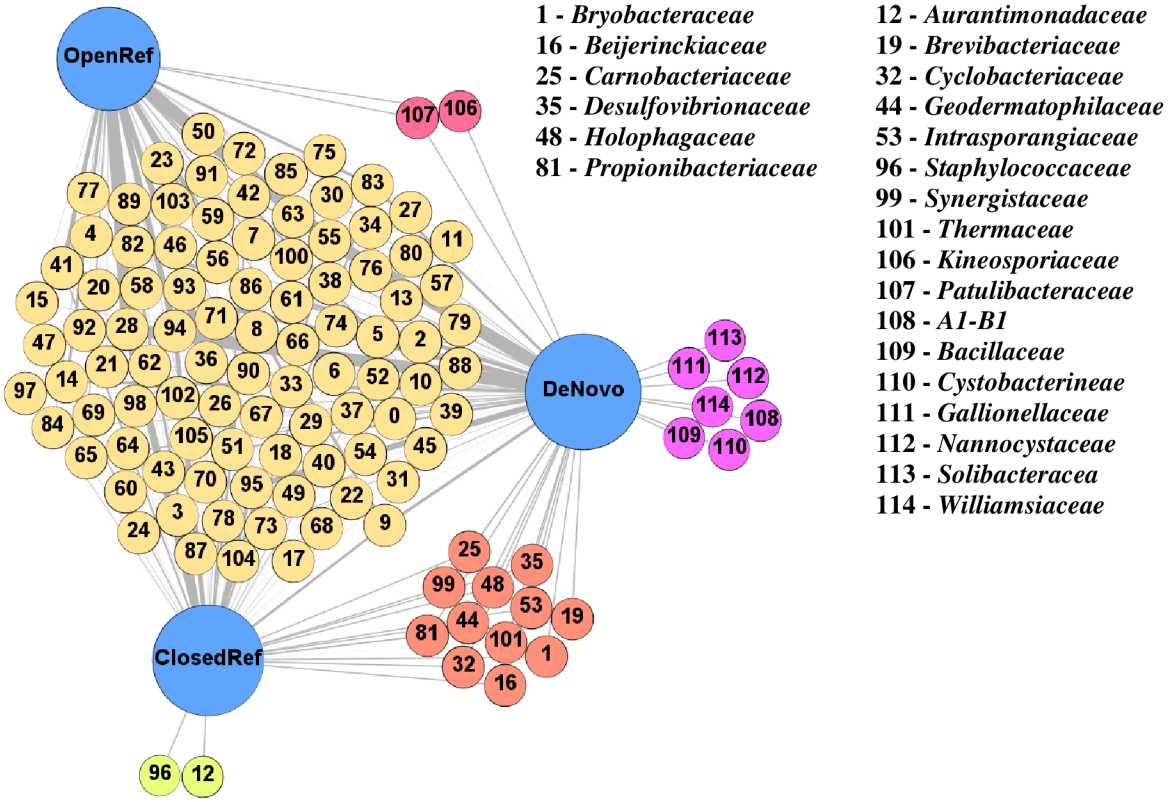
Tabulka 10: Srovnání výsledků získaných *de novo*, open reference a closed reference OTU picking metodami.

OTU picking metoda	shluky	Nová OTU	Celkový počet čtení	Nalezené taxonomické kategorie						
				říše	kmen	třída	řád	čeleď	rod	druh
De novo	4149	921	56661	1	19	38	63	113	185	51
Open reference	2596	276	53740	1	16	31	56	94	134	45
Closed reference	1410	0	30345	1	19	37	62	106	194	89

Obrázek 37 ukazuje odlišnosti v dosažené diverzitě na úrovni čeledi. OTU jsou rozděleny do pěti barevně odlišených komunit na základě příslušnosti k jednotlivým OTU picking metodám. Získaná nižší diverzita open reference metody je zapříčiněna odlišnými defaultními parametry. Open reference totiž využívá prahování a odstraňuje OTU tvořená jediným čtením. Na úrovni taxonomických čeledí se toto nastavení projevilo například potlačením čeledi *Staphylococcaceae* s počtem 11 čtení nalezených closed reference metodou.

Metody se také lišily v počtu nově detekovaných OTU. Z podstaty closed reference přístupu vyplývá, že nemůže detekovat žádné nové OTU, a rovněž celkový počet zpracovaných čtení je nižší, neboť čtení, která nemohou být spárována s referencí, jsou

z datasetu vyřazena. Naproti tomu open reference přístup vytvořil 276 nových OTU, a *de novo* metoda dokonce 921.



Obr. 37: Taxonomické čeledi nalezené closed reference, open reference a *de novo* OTU picking metodami.

Práh podobnosti

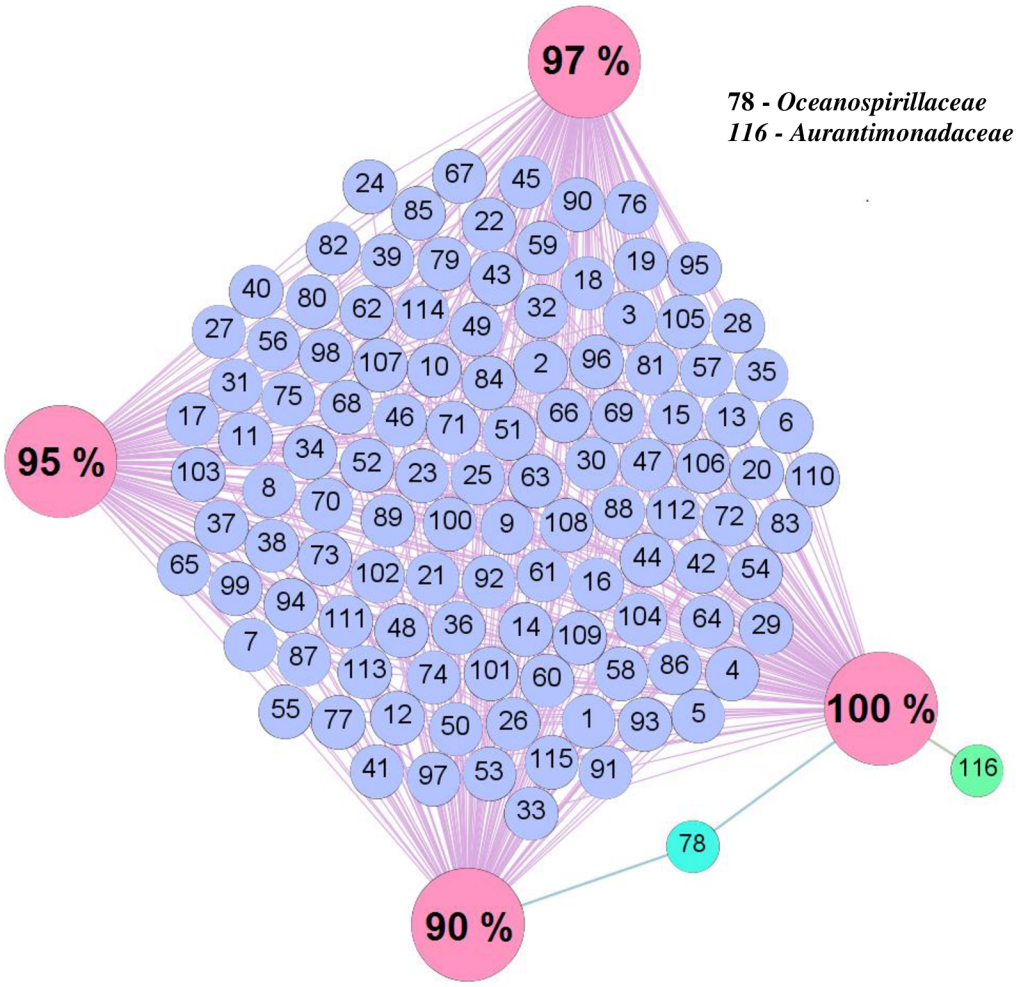
Pro sledování změn způsobených nastavením parametru podobnosti byla data zpracována open reference OTU picking jakožto obecně doporučovanou a pravděpodobně nejpoužívanější OTU picking metodou. Dříve popsané prahování open reference OTU picking metody, které odstraňuje OTU tvořená jediným čtením, bylo potlačeno.

Byl volen práh podobnosti 90 %, 95 %, 97 % a 100 %. V tabulce 11 je uveden maximální počet rozdílných bází pro přiřazení sekvencí do shluků, uvažujeme-li délku všech sekvencí 476 bází podle jejich mediánu (viz tabulka 7 v kapitole 7.1).

Tabulka 11: Maximální počet odlišných bází pro přiřazení do shluku

Práh podobnosti	Všechny sekvence
90 %	47 bazí
95 %	23 bazí
97 %	14 bazí
100 %	0 bazí

Rozdíly v dosažených výsledcích nebyly při změnách prahu podobnosti tak markantní jako při změnách OTU picking metody. Dle předpokladů při nastavení prahu podobnosti na 90 % bylo nalezeno nejméně OTU nezařazených žádné taxonomické čeledi (celkem 794). Následovalo 95 % nastavení (1170 nezařazených OTU), 97 % (1460 nezařazených OTU) a 100 % s 2933 nezařazenými OTU.



Obr. 38: Taxonomické čeledi nalezené open reference OTU picking při prahu podobnosti 90 %, 95 %, 97 % a 100 %.

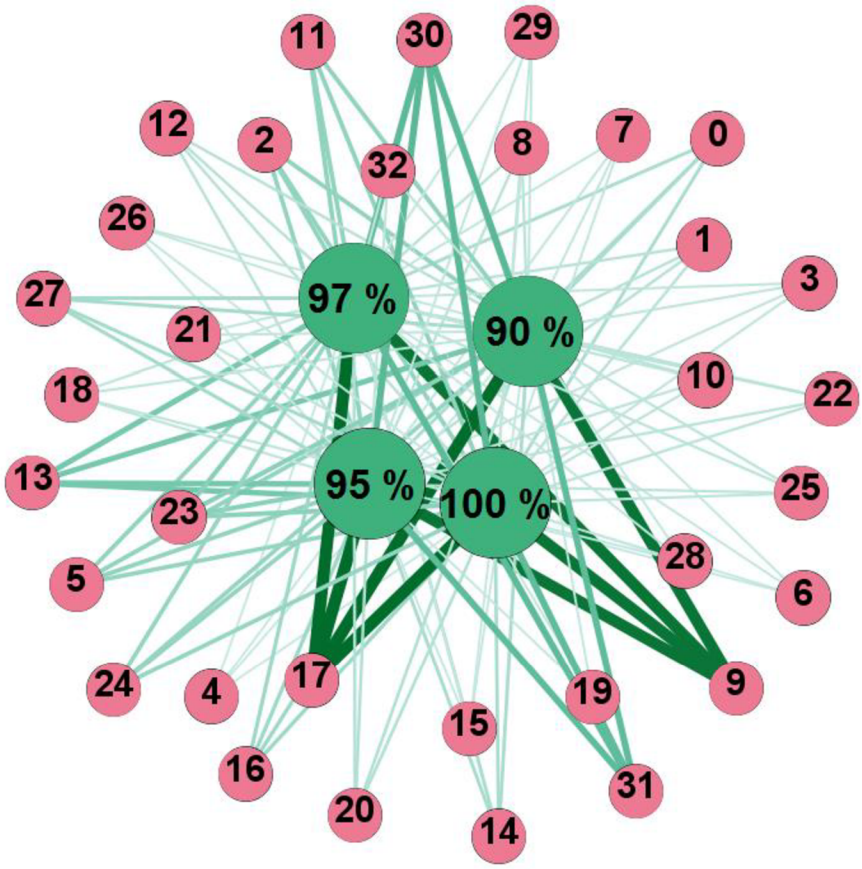
Obr. 38 shrnuje výsledky dosažené diverzity v taxonomických čeledích. 95% a 97% práh detekovaly stejnou diverzitu, nastavením 90% prahu byla navíc detekována čeleď *Oceanospirillaceae* a nastavením 100% byla kromě čeledi *Oceanospirillaceae* nalezena ještě čeleď *Aurantimonadacae*. Jedině prahovou hodnotou 100 % jsme se tak dorovnali diverzitě dosažené closed reference OTU picking přístupem při 97% prahu na obrázku 37. Čeleď *Aurantimonadacae* nebyla zastoupena velkým počtem čtení (9 čtení u open reference a 11 čtení u closed reference OTU picking metody). Zdá se, že schopnost její detekce byla silně závislá na postupu přiřazování taxonomické kategorie. Tvůrci QIIME přisuzují větší důvěryhodnost taxonomickému zařazení closed reference metodou, protože je použita přímo taxonomie referenční sekvence, ke které byla námi analyzovaná sekvence přiřazena. U open reference jsou nejdříve reprezentativní sekvence z closed i *de novo* kroku uloženy do jednoho

textového souboru, a poté je u nich hledáno taxonomické zařazení opakovaným porovnáváním s referenční databází.

Ani rozdíly v detekované abundanci druhů nebyly velké. V tabulce 12 a na obr. 39 jsou uvedeny čeledi, jejichž abundance byla 100 a vyšší. Čísla čeledí z tabulky odpovídají číslům čeledí z grafu. Porovnáme-li intervaly, v jakých se vyskytovaly pozorované četnosti organismů při změnách prahu podobnosti (90 % , 95 %, 97 % a 100 %), jejich variační rozpětí, tloušťky hran a jejich barevný odstín vycházející z jednotlivých OTU uzlů v grafu, zjistíme, že rozdíly jsou poměrně malé.

Tabulka 12: Rozdíly v zaznamenané četnosti jednotlivých organismů. V kolonce interval je zanesen pozorovaný interval napříč analýzami s 90%, 95%, 97% a 100% prahem podobnosti. Dosažené variační rozpětí je uvedeno v třetím a šestém sloupci tabulky.

čeleď	interval	variační rozpětí	čeleď	interval	variační rozpětí
0-Weeksellaceae	1140 - 1153	13	17-Lactobacillaceae	8480 - 8489	9
1-ACK-M1	689 - 698	9	18-Lachnospiraceae	458 - 469	11
2-Aeromonadaceae	1970 - 1996	26	19-Microbacteriaceae	260 - 270	10
3-Alcaligenaceae	405 - 408	3	20-Moraxellaceae	893 - 906	13
4-Bacteriovoracaceae	192 - 202	10	21-Neisseriaceae	400 - 409	9
5-Bacteroidaceae	1954 - 1993	39	22-Oxalobacteraceae	574 - 584	10
6-Bdellovibrionaceae	255 - 256	1	23-Pseudomonadaceae	2267 - 2298	31
7-Burkholderiaceae	308 - 312	4	24-Rhizobiaceae	1880 - 1918	38
8-Caulobacteraceae	601 - 609	8	25-Rhodobacteraceae	536 - 540	4
9-Comamonadaceae	7979 - 8019	40	26-Rhodocyclaceae	268 - 346	78
10-Coxiellaceae	495 - 504	9	27-Rhodospirillaceae	1350 - 1352	2
11-Cytophagaceae	2037 - 2068	31	28-Ruminococcaceae	281 - 308	27
12-Enterobacteriaceae	670 - 733	63	29-Shewanellaceae	289 - 374	85
13-Flavobacteriaceae	2995 - 3034	39	30-Sphingobacteriaceae	4072 - 4114	42
14-Fusobacteriaceae	1194 - 1195	1	31-Sphingomonadaceae	3918 - 3953	35
15-Hyphomicrobiaceae	210 - 212	2	32-Xanthomonadaceae	456 - 460	4
16-Chitinophagaceae	1572 - 1574	2			

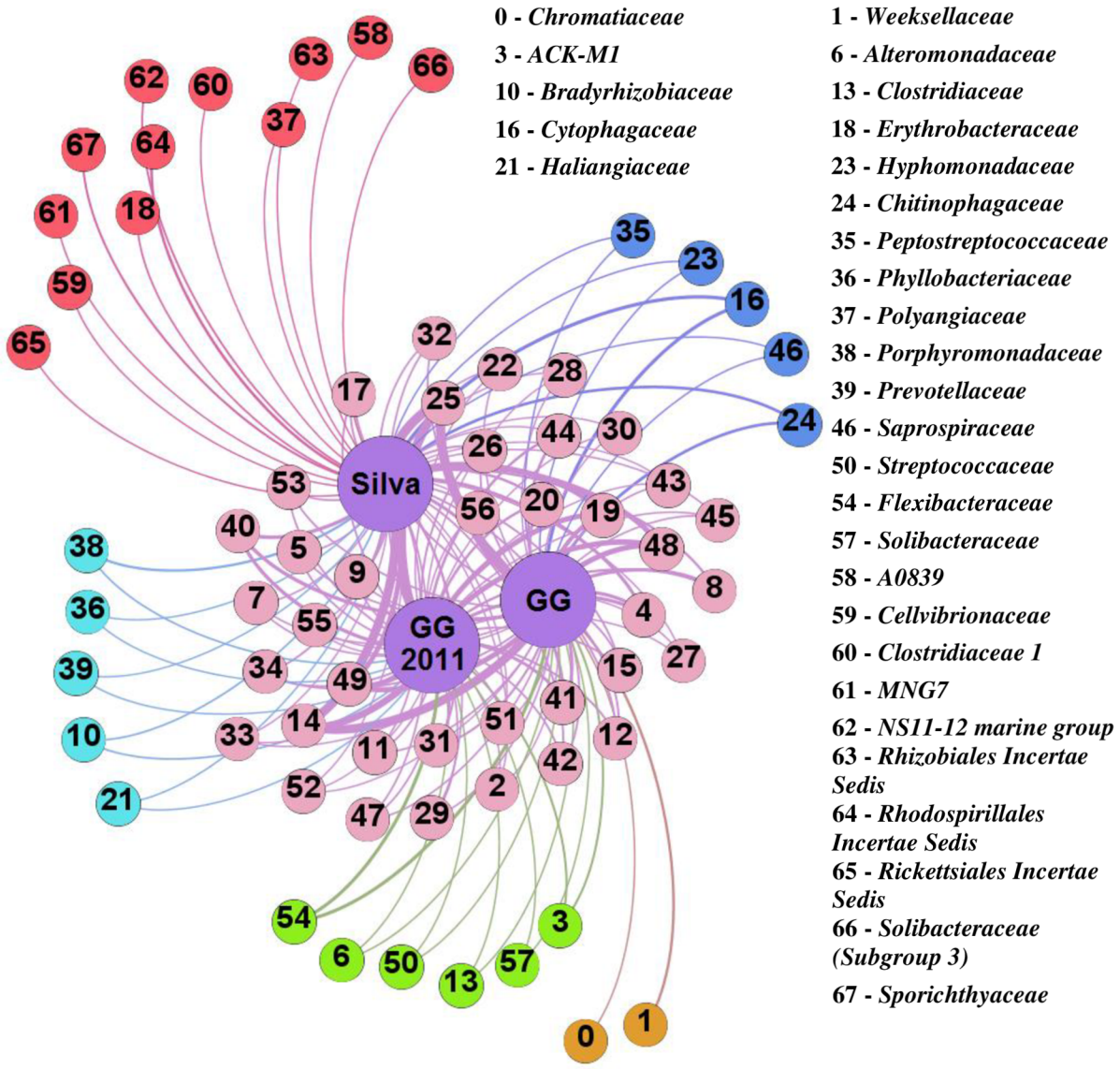


Obr. 39: Graf srovnávající abundanci OTU při různém nastavení prahové hodnoty.

Volba referenční databáze

Posledním zkoumaným faktorem ovlivňujícím dosažené výsledky byla volba referenční databáze. Data byla zpracovávána open reference metodou při 97% prahu podobnosti. Jako shlukovací algoritmus byl volen uclust a porovnávány byly tři referenční databáze – nejaktuálnější QIIME kompatibilní databáze Silva 128 z 29. září 2016, nejaktuálnější databáze Greengenes 13_8 ze srpna 2013 a dále Greengenes databáze ze 4. února 2011. Nastavení open reference přístupu pro odstraňování shluků s jediným pozorováním bylo ponecháno.

Výsledky analýzy jsou zachyceny na obr. 40. Z důvodu velkého počtu uzlů byla data redukována pouze na čeledi, jejichž abundance byla 20 a vyšší. Při využití databáze Silva byl zaznamenán velký počet OTU, které nebyly detekovány ani novější, ani starší databází Greengenes (v grafu značeny jako GG a GG 2011). Podíváme-li se však podrobněji na uzly, v jejichž detekci se databáze lišily, zjistíme, že se může jednat také o nejednotnost názvosloví. Například uzly 57 a 66 oba dva zastupují čeleď *Solibacteraceae*. Rozdíl je ale ve značení této čeledi – zatímco obě Greengenes databáze ji značí jako *„Solibacteraceae“*, v databázi Silva je pojmenována jako *„Solibacteraceae (subgroup 3)“*, ačkoliv se v databázi Silva žádná další podtřída této čeledi nenachází, jak by její název mohl napovídat (viz [68]).



Obr. 40: Taxonomické čeledi nalezené použitím tří odlišných databází.

V grafu se také vyskytly uzly 63, 64 a 65 s taxonomickou kategorií *Incertae Sedis*, tedy nejasného zařazení. Jedná se o taxony, u nichž nejsou známy vyšší taxonomické skupiny. Například bakterie druhu *Reynarella massiliensis* spadá do řádu *Rhizobiales*, avšak v současné době není zařazena do žádné čeledi. V Silva databázi je tak její čeleď vyplněna jako *Rhizobiales Incertae Sedis*.

Zdá se tedy, že se databáze Silva snaží o přesnější určení OTU, avšak vykazuje nejednotnosti, které počítačové zpracování datasetu mohou komplikovat. Například neznámá taxonomická kategorie bakterie může být značena uncultured bacterium, uncultured, uncultured organism, unassigned, unidentified a další.

Přiblížit odlišnosti se také pokouší tabulka 13. Do této tabulky byly zahrnuty pouze ty OTU, jejichž abundance byla 100 a vyšší.

Tabulka 13: Rozdíly v zaznamenané četnosti jednotlivých organismů. V kolonce interval je zanesen pozorovaný interval napříč analýzami s použitím databáze Silva, Greengenes z roku 2013 a Greengenes z roku 2011. Dosažené variační rozpětí je uvedeno v třetím a šestém sloupci tabulky.

čeleď	interval	variační rozpětí	čeleď	interval	variační rozpětí
Weeksellaceae	0 - 1106	1106	Lachnospiraceae	415 - 614	199
ACK-M1	0 - 696	696	Microbacteriaceae	252 - 258	6
Aeromonadaceae	1902 - 1959	57	Moraxellaceae	798 - 843	45
Alcaligenaceae	383 - 389	6	Neisseriaceae	360 - 377	17
Bacteriovoracaceae	196 - 198	2	Oxalobacteraceae	442 - 513	71
Bacteroidaceae	986 - 1955	969	Porphyromonadaceae	0 - 1007	1007
Bdellovibrionaceae	223 - 238	15	Pseudomonadaceae	2169 - 2229	60
Bradyrhizobiaceae	0 - 305	305	Rhizobiaceae	1637 - 1869	232
Burkholderiaceae	302 - 356	54	Rhodobacteraceae	491 - 513	22
Caulobacteraceae	566 - 569	3	Rhodocyclaceae	0 - 243	243
Comamonadaceae	5751 - 7896	2145	Rhodospirillaceae	0 - 1327	1327
Coxiellaceae	497 - 497	0	Ruminococcaceae	189 - 252	63
Cytophagaceae	0 - 2013	2013	Shewanellaceae	270 - 390	120
Enterobacteriaceae	537 - 613	76	Sphingobacteriaceae	3994 - 4044	50
Erythrobacteraceae	0 - 282	282	Sphingomonadaceae	3643 - 3860	217
Flavobacteriaceae	2888 - 3988	1100	Xanthomonadaceae	416 - 430	14
Fusobacteriaceae	1176 - 1199	23	Flexibacteraceae	0 - 1929	1929
Hyphomicrobiaceae	187 - 193	6	NS11-12 marine group	0 - 406	406
Chitinophagaceae	0 - 1477	1477	Rhodospirillales Incertae Sedis	0 - 1251	1251
Lactobacillaceae	8441 - 8454	13	Sporichthyaceae	0 - 695	695

Velké rozdíly jsou vidět u čeledi *Weeksellaceae*, která byla detekována pouze Greengenes databází z roku 2013. A také například u četně zastoupené čeledi *Cytophagaceae*, která nebyla s využitím Greengenes databáze z roku 2011 vůbec detekována.

Co se týče nalezených nových OTU, nejvíce jich bylo detekováno starší Greengenes databází (737). Novější Greengenes verze ji následovala (501), a naopak nejméně nezařazených OTU (277) nalezla databáze Silva, která je z analyzovaných databází nejnovější.

Závěr

Výstupem diplomové práce je především vypracovaný R balíček pro snadnou tvorbu bipartitních grafů z mikrobiálních dat. Součástí práce bylo také vypracování literární rešerše zabývající se problematikou vizualizace tohoto typu dat a ukázalo se, že současné metody nesplňují požadavky kladené na studium mikrobiomů. Nedostatky vidím především v nízké provázanosti studovaných společenstev a jejich propojení, složitosti detekce důležitých vzorů a komunit, a také možnosti vykreslení konkrétních organismů. Zdá se, že vizualizace pomocí bipartitních grafů by mohla mnohé z těchto problémů vyřešit.

Součástí balíčku je také několik funkcí umožňujících zpracování dat vedoucí k výsledné větší přehlednosti finálního grafu. Jedná se konkrétně o metody redukce na základě taxonomického shlukování, odstranění OTU s nízkou abundancí, shlukování na základě aditivních informací o vzorcích a metody zaměřující se na konkrétní taxonomické větve. Analýza těchto funkcí ukázala, že vhodnou kombinací zmíněných metod můžeme docílit přehledných grafů za současného zachování velkého množství původních sekvencí.

Do balíčku je rovněž implementováno 7 funkcí pro detekci komunit. Díky převedení bipartitního grafu do jeho projekce je tak možno využít detekce komunit pomocí metody bloudění, vlastního vektoru, propagace značky, mezilehlosti hran, rychlého rozvíjení a při nízkém počtu uzlů také optimálního rozložení. Algoritmy byly testovány na vzorcích vody sloužící k přepravě ryb a obzvláště dobrých výsledků bylo dosaženo pomocí algoritmu rychlého rozvíjení. Jeho výpočetní náročnost nebyla tak vysoká jako při výpočtu optimálního rozložení a dosažené výsledky byly totožné. Navíc se zdá, že výsledky korespondují s informacemi o datasetu, totiž se zemí původu a s konkrétními dodavateli.

V diplomové práci jsem se dále věnovala popisu zpracování surových dat získaných ze sekvenátoru, přičemž jsem se zaměřila na ty kroky zpracování, které mohou výsledná data nejvíce ovlivnit. Některé z těchto kroků byly poté analyzovány v praktické části práce. Velký vliv na výsledná data měla volba referenční databáze. Naopak nastavení 90 %, 95 %, 97 % a 100 % prahu podobnosti vedlo na taxonomické úrovni čeledí k takřka nepozměněným výsledkům. Dále se ukázalo, že ačkoliv se open reference OTU picking považuje za zlatou střední cestu mezi *de novo* a closed reference OTU picking metodami, je velmi složité přiblížit se open reference přístupem výsledkům těchto dvou metod. Zejména pak kvůli odlišnému stylu přiřazování taxonomie, a také vlivem dodatečných filtračních kroků.

Pro potřeby této diplomové práce bylo využito platformy QIIME a výpočetních serverů MetaCentra. Navržený balíček je volně dostupný na www.github.com/safma/bipartiteOTU.

Literatura

[1] WETTERSTRAND, K. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)* [online]. [cit. 2016-10-16].

[2] HEATHER, J. M., CHAIN, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* [online]. 2016, roč. 107, č. 1, s. 1–8. doi: 10.1016/j.ygeno.2015.11.003

[3] HUGENHOLTZ, P., GOEBEL, B. M., PACE, N. R. Impact of culture independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology*. 1998, roč. v, č. 18, s. 180p4765–4774. doi: 0021-9193/98/$04.00+0

[4] HANDELSMAN, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews* [online]. 2004, roč. 68, č. 4, s. 669–685. doi: 10.1128/MMBR.68.4.669-685.2004

[5] DASH, S., et al. The gut microbiome and diet in psychiatry: focus on depression. *Current opinion in psychiatry* [online]. 2015, roč. 28, č. 1, s. 1–6. doi: 10.1097/YCO.0000000000000117

[6] TILG, H., MOSCHEN, A. R. Microbiota and diabetes: an evolving relationship. 2014, s. 1–9. doi: 10.1136/gutjnl-2014-306928

[7] HU, Y., et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nature communications* [online]. 2013, roč. 4, s. 2151. doi: 10.1038/ncomms3151

[8] CABRERA-RUBIO, R., et al. Microbiome diversity in the bronchial tracts of patients with chronic obstructive pulmonary disease. *Journal of Clinical Microbiology*. 2012, roč. 50, č. 11, s. 3562–3568. doi: 10.1128/JCM.00767-12

[9] HU, Y., et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nature communications*, 2013, 4.

[10] UHLÍK, O., et al. Identifikace a charakterizace bakterií s bioremediačním potenciálem - Od kultivace k metagenomice. *Chemicke Listy*. 2013, roč. 107, č. 8, s. 614–622.

[11] KNIERIM, E., et al. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS ONE*. 2011, roč. 6, č. 11, doi: 10.1371/journal.pone.0028240

[12] SANGER, F., NICKLEN, S., COULSON, R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* [online]. 1977, roč. 74, č. 12, s. 5463–7. doi: 10.1073/pnas.74.12.5463

[13] WOESE, C. R., FOX, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*. 1977, roč. 74, č. 11, s. 5088–5090. doi: 10.1073/pnas.74.11.5088

[14] HAMADY, M., et al. Error-correcting barcoded primers allow hundreds of samples to be pyrosequenced in multiplex. 2012, roč. 5, č. 3, s. 235–237. doi: 10.1038/nmeth.1184

[15] CAPORASO, J. G., et al. QIIME allows analysis of high-throughput community sequencing data [online]. [cit. 2016-10-16].

[16] HSIAO, William. Module 2: Marker Gene Based Analysis [online]. [cit. 2016-10-16].

[17] MAXAM, a M, GILBERT, W. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* [online]. 1977, roč. 74, č. 2, s. 560–4. doi: 10.1073/pnas.74.2.560

[18] 454 Life Sciences: How is genome sequencing done? *www.454.com* [online]. [cit. 2016-11-01].

[19] SIQUEIRA, J. F., A. F. FOUAD a I. N. RÔÇAS. Pyrosequencing as a tool for better understanding of human microbiomes. *Journal of Oral Microbiology* [online]. 2012-1-23, 4, - [cit. 2016-11-01]. doi: 10.3402/jom.v4i0.10743. ISSN 2000-2297.

[20] LIU, L., Yinhu L., Siliang L., et al. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* [online]. 2012, 2012, 1-11 [cit. 2016-11-01]. doi: 10.1155/2012/251364. ISSN 1110-7243.

[21] Advance Genomics with single molecule: Real-time (SMRT) sequencing. *www.pacb.com* [online]. 2016 [cit. 2016-11-01].

[22] TURNER, S. DNA Sequencing/genomics:: Toward personalized medicine: 3G DNA sequencing. *BioOptics World* [online]. 2010, 2010(3) [cit. 2016-11-01].

[23] EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010, 26(19), 2460-2461. doi: 10.1093/bioinformatics/btq461. ISSN 1367-4803.

[24] LI, W., GODZIK, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006, 22(13), 1658-1659. doi: 10.1093/bioinformatics/btl158. ISSN 1367-4803.

[25] SCHLOSS, P. D.,WESTCOTT, S. L., RYABIN T., et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*. 2009, 75(23), 7537-7541. doi: 10.1128/AEM.01541-09. ISSN 0099-2240.

[26] LOZUPONE, C., KNIGHT, R., RYABIN, T., et al. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*. 2005, 71(12), 8228-8235. doi: 10.1128/AEM.71.12.8228-8235.2005. ISSN 0099-2240.

[27] SEDLAR, K., SKUTKOVA, H., VIDENSKA, P., et al. Bipartite graphs for metagenomic data analysis and visualization: a New Phylogenetic Method for Comparing Microbial Communities. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2015, 71(12), 1123-1128. doi: 10.1109/BIBM.2015.7359839. ISBN 978-1-4673-6799-8. ISSN 0099-2240.

[28] GOWER, J. C., Principal Coordinates Analysis: a New Phylogenetic Method for Comparing Microbial Communities. *Wiley StatsRef: Statistics Reference Online*. Chichester, UK: John Wiley, 2014, 71(12), 1. doi: 10.1002/9781118445112.stat05670.pub2. ISBN 9781118445112. ISSN 0099-2240.

[29] BARBERÁN, A., BATES, S. T., CASAMAYOR, E., et al. Using network analysis to explore co-occurrence patterns in soil microbial communities: a New Phylogenetic Method for Comparing Microbial Communities. *The ISME Journal*. Chichester, UK: John Wiley, 2011, 6(2), 343-351. doi: 10.1038/ismej.2011.119. ISBN 9781118445112. ISSN 1751-7362.

[30] DAMETTO, P. Sequencing technologies - the next generation. *Presentation*. [online]. 2011. [cit. 2016-12-12]

[31] Applied Biological Materials (ABM) Inc. Next Generation Sequencing (NGS) - An Introduction [online]. 2015. [cit. 2016-12-12]

[32] DALLMAN, T. Applications of Whole Genome Sequencing to Food Safety - Perspective from a reference laboratory. *Presentation.* [online]. 2015. [cit. 2016-12-12]

[33] Center for Genomics and Transcriptomics, *Services – Next-Generation Sequencing.* [online]. [cit. 2016-12-12]

[34] MASOUDI-NEJAD, A., NARIMANI, Z., HOSSEINKHAN, N.. Next Generation Sequencing and Sequence Assembly: Methodologies and Algorithms. *Springer New York*, 2013. 86 s. doi: 10.1007/978-1-4614-7726-6. ISBN 9781461477259. ISSN 2193-4746.

[35] Alimetrics, DNA Sequence analysis – the only species-specific approach for novel bacteria. [online]. [cit. 2016-12-12]

[36] CAPORASO, J. G., et al. QIIME: Input Files [online]. 2015. [cit. 2016-12-12]

[37] CAPORASO, J. G., et al. QIIME: OTU picking [online]. 2015. [cit. 2016-12-12]

[38] YATSUNENKO, T., et al. Human gut microbiome viewed across age and geography. 2012, roč. 486, č. 7402, s. 222–227. doi: 10.1038/nature11053.Human

[39] JARKOVSKÝ, J., LITTNEROVÁ, S. Vícerozměrné statistické metody: Shluková analýza. *Prezentace.* In: IBA MU [online]. Brno, 2015 [cit. 2016-05-12].

[40] HU, M., et al. The diversity and abundance of As (III) oxidizers on root iron plaque is critical for arsenic bioavailability to rice. *Nature Publishing Group* [online]. 2015, č. July, s. 1–10. doi: 10.1038/srep13611

[41] MAATEN, L., HINTON, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, s. 2579-2605.

[42] HINTON, G., ROWEIS, E. Stochastic neighbor embedding. *Advances in neural information processing systems*. 2002. s. 833-840.

[43] WANG, Y., et al. Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. *Applied and environmental microbiology*, 2012. s. 8264-8271.

[44] HUMAN MICROBIOME PROJECT CONSORTIUM, et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 2012. s.207-214. doi: 10.1038/nature11234

[45] LETUNIC, I., BORK, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, 2016. doi: 10.1093/nar/gkw290

[46] HUERTA-CEPAS, J., SERRA, F., BORK, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 2016, s. 1635-1638. doi: 10.1093/molbev/msw046
[47] ASNICAR, F., et al. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, 2015, doi: 10.7717/peerj.1029
[48] CHERVEN, K. Mastering Gephi network visualization. *Packt Publishing Ltd*, 2015. ISBN: 978-1-78398-734-4
[49] KARLSSON, F., et al. Assessing the human gut microbiota in metabolic diseases. *Diabetes*, 2013, 62.10: 3341-3349. doi: https://doi.org/10.2337/db13-0844
[50] GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 2002, 99.12: 7821-7826.
[51] SHEN, H.W. Community structure of complex networks. *Springer Science & Business Media*, 2013.
[52] NEWMAN, M., GIRVAN, M. Finding and evaluating community structure in networks. *Physical review E*, 2004, 69.2: 026113.
[53] RAGHAVAN, U. N., ALBERT, R., KUMARA, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 2007, 76.3: 036106.
[54] PONS, P., LATAPY, M. Computing communities in large networks using random walks. *International Symposium on Computer and Information Sciences.* Springer Berlin Heidelberg, 2005. p. 284-293.
[55] HOBBS, W. My favorite application using eigenvalues: partitioning and community detection in social networks. 2013.
[56] GERZOVA, L., et al. Characterization of microbiota composition and presence of selected antibiotic resistance genes in carriage water of ornamental fish. *PloS one*, 2014, 9.8: e103865.
[57] TURNER, S. Pacific Biosciences Sequencing Technology. 2016.
[58] YARZA, P., et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, 2014, 12.9: 635-645.
[59] MAHÉ, F., et al. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2014, 2: e593.
[60] HERNANDEZ, M. E., et al. Oxygen availability is a major factor in determining the composition of microbial communities involved in methane oxidation. *PeerJ*, 2015, 3: e801.
[61] GISBRECHT, A., HAMMER, B., MOKBEL B., SCZYRBA, A.. Nonlinear Dimensionality Reduction for Cluster Identification in Metagenomic Samples. *Information Visualisation (IV)*, 2013 17th International Conference [online]. IEEE, 1307, s. 174-179. ISSN 15506037.
[62] ROSSANT C., An illustrated introduction to the t-SNE algorithm, 2015. *www.oreilly.com* [online].
[63] MEJOVA, Y. Language of Politics on Twitter. *Presentation*. [online]. 2015. [cit. 2017-4-28]
[64] PAPADOPOULOS, S. Community Detection in Social Media. *Presentation.* [online]. 2011. [cit. 2017-4-28]
[65] CSARDI, G., NEPUSZ T., The igraph software package for complex network research, *InterJournal, Complex Systems* 1695. 2006. *http://igraph.org*
[66] GUIMERÀ, R., SALES-PARDO, M., AMARAL, L., Nunes A. Module identification in bipartite and directed networks. *Physical Review E*, 2007, 76.3: 036102.

[67] ALZAHRANI, T., HORADAM, K. J. Community detection in bipartite networks: Algorithms and case studies. *Complex Systems and Networks*. Springer Berlin Heidelberg, 2016. p. 25-50.
[68] Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *www.arb-silva.de*

Seznam zkratek

GML	Graph Modeling Language
NGS	Next Generation Sequencing
OTU	Operational Taxonomic Unit
PCA	Principal components analysis
PCoA	Principal coordinate analysis
QIIME	Quantitative Insights Into Microbial Ecology
SMRT	Single molecule real time sekvenování
SMS	Single molecule sequencing
SOLiD	Sequencing by Oligo Ligation and Detection
t-SNE	t-distributed stochastic neighbor embedding
UniFrac	Unique Fraction metrika
ZMWs	Zero-mode waveguides

Seznam příloh

A. Obsah přiloženého CD

Praktické zpracování R/Bioconductor balíčku je součástí přiloženého CD. Funkce byly napsány v programovacím prostředí R verze 3.3.0 (2016-05-03). Přiložené CD obsahuje kromě souborů balíčku také diplomovou práci ve formátu pdf.