VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

DIPLOMOVÁ PRÁCE

Brno, 2024 Bc. Julie Nejezchlebová

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

IDENTIFIKACE NEZNÁMÝCH BAKTERIÁLNÍCH GENOMŮ POMOCÍ ONLINE DATABÁZOVÉHO NÁSTROJE

IDENTIFICATION OF UNKNOWN BACTERIAL GENOMES USING AN ONLINE DATABASE TOOL

DIPLOMOVÁ PRÁCE

MASTER'S THESIS

AUTOR PRÁCE — Bc. Julie Nejezchlebová

AUTHOR

VEDOUCÍ PRÁCE — Ing. et Ing. Jana Schwarzerová, MSc

SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní program **Biomedicínské inženýrství a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Bc. Julie Nejezchlebová ***ID:*** 221525

Ročník: 2 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

Identifikace neznámých bakteriálních genomů pomocí online databázového nástroje

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s principy laboratorních technik zaměřené na bakteriální typizaci. 2) Prostudujte sbírku bakteriálních genomů poskytnuté Výzkumným ústavem veterinárního lékařství (VÚVEL). 3) Vytvořte automatický softwarový nástroj, který provede přímé porovnání bakteriálních sekvenačních dat. 4) Vytvořený nástroj implementujte do online databáze na Výzkumném ústavu veterinárního lékařství (VÚVEL). 5) Vámi vytvořený nástroj otestujte a odstraňte případné uživatelské nedostatky. 6) Celé Vámi navrhnuté řešení včetně výsledků diskutujte.

Práce bude prováděna na datech poskytnutých z Výzkumného ústavu veterinárního lékařství (VÚVeL).

DOPORUČENÁ LITERATURA:

[1] MEDVEDCKY, M., CEJKOVA, D., POLANSKY, O., KARASKOVA, D., KUBESOVA, T., CIZEK, A. and RYCHLIK, I., M, 2018. Whole genome sequencing and function prediction of 133 gut anaerobes isolated from chicken caecum in pure cultures. BMC genomics, 19(1), pp.1-15.

[2] PENG, C., LIN, Y., LUO, H., GAO, F., 2017. A comprehensive overview of online resources to identify and predict bacterial essential genes. Frontiers in microbiology, 8, p.2331.

Termín zadání: 5.2.2024 ***Termín odevzdání:*** 22.5.2024

Vedoucí práce: Ing. et Ing. Jana Schwarzerová, MSc

Konzultant: RNDr. Bohuslav Zmek

prof. Ing. Valentine Provazník, Ph.D.
předseda rady studijního programu

ABSTRAKT

Diplomová práce se zabývá vytvořením automatického softwarového nástroje Bacterial Explorer, který umožňuje odhalení nových bakterií za pomoci dostupných bioinformatických nástrojů. Nástroj je vytvořen v souladu s požadavky Výzkumného ústavu veterinárního lékařství (VÚVeL) a je testován na datech z poskytnuté VÚVeL databáze. Teoretická část je věnována popisu bakteriální typizace, metodám, které se používají pro genomickou analýzu a již dostupným nástrojům pro bakteriální typizaci. V praktické části se práce zaměřuje na implementaci automatického softwarového nástroje s názvem Bacterial Explorer zahrnující popis nástrojů na pozadí vytvořeného nástroje, uživatelské prostředí a implementaci do online databáze VÚVeL. V poslední části se praktická část zabývá testováním nástroje a diskusí výsledků.

KLÍČOVÁ SLOVA

Bakterie, Bioinformatika, Genomická analýza, VÚVeL (Výzkumný ústav veterinárního lékařství), Bacterial Explorer

ABSTRACT

This master's thesis deals with the develop of an automatic software tool - Bacterial Explorer which allows the discovery of unknown bacteria using available bioinformatics tools. The tool is developed in accordance with the requirements of the Veterinary Research Institute (VRI) and is tested on data from the provided VRI database. The theoretical part is dedicated to the description of bacterial typing, methods used for genomic analysis, and already available tools for bacterial typing. In the practical part, the thesis focuses on the implementation of the automatic software tool called Bacterial Explorer, including a description of the tools behind the created tool, the user interface and implementation in the online database of VRI . The final part of the practical section deals with tool testing and discussion of the results.

KEYWORDS

Bacteria, Bioinformatics, Genomic Analysis, VRI (Veterinary Research Institute), Bacterial Explorer

NEJEZCHLEBOVÁ, Julie. *Identifikace neznámých bakteriálních genomů pomocí online databázového nástroje*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2024, 98 s. Diplomová práce. Vedoucí práce: Ing. et Ing. Jana Schwarzerová, MSc

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Bc. Julie Nejezchlebová
VUT ID autora:	221525
Typ práce:	Diplomová práce
Akademický rok:	2023/24
Téma závěrečné práce:	Identifikace neznámých bakteriálních genomů pomocí online databázového nástroje

Prohlašuji, že svou závěrečnou práci jsem vypracovala samostatně pod vedením vedoucí/ho závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

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

Brno ....

podpis autorky*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Ráda bych poděkovala vedoucí diplomové paní Ing. et Ing. Janě Schwarzerové, MSc. za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci a panu RNDr. Bohuslavu Zmekovi za odborné konzultace. Panu doc. RNDr. Ivanu Rychlíkovi, Ph.D. a dalším pracovníkům zapojeným do projektu NaCeBiVet bych ráda poděkovala za odborné rady a možnost spolupráce na projektu NaCeBiVet.

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Bakteriální typizace a identifikace hrají klíčovou roli při odhalování nových bakterií. V současné době se Výzkumný ústav veterinárního lékařství (VÚVeL) snaží identifikovat nové bakterie v mikrobiomu zvířat pro účely inovativních probiotických metod. Bakteriální typizace obecně zahrnuje dva přístupy - fenotypové a genotypové metody. Fenotypové metody představují nemolekulární metody identifikace bakterií a jejich hlavní nevýhodou je zejména časová náročnost. Genotypové metody představují molekulární metody typizace a oproti fenotypovým metodám jsou rychlejší. Diplomová práce se zabývá vytvořením nástroje Bacterial Explorer, který umožňuje rychlé porovnání neznámé bakterie s referenčními genomy.

Diplomová práce se v první kapitole teoreticky věnuje bakteriální typizaci. Je zde nastíněna problematika fenotypových a genotypových metod, přičemž je kladen důraz zejména na genotypové metody. Jsou zde popsány jak nesekvenační genotypové metody, tak sekvenační genotypové metody.

Druhá kapitola popisuje metody, které se uplatňují při analýze genomických dat. Tato kapitola přináší přehled algoritmů, které jsou na pozadí bioinformatických nástrojů. V první části této kapitoli je popsán hladový algoritmus a heuristické vyhledávání podobných sekvencí. Druhá část této kapitoly se zabývá skrytými Markovovými modeli, algoritmem mapování bez zarovnání a algoritmem minMLST.

Ve třetí kapitole jsou popsána data poskytnutá z bakteriální databáze VÚVeLu. Na tuto třetí kapitolu navazuje čtvrtá kapitola, která se zabývá “state-of-the-art“ dostupnými nástroji, které jsou nyní nejčastěji uplatňovány při bakteriální typizaci. Jsou zde popsány nástroje BLAST, Cd-hit, Barrnap, FastANI a BLAT.

Poslední kapitola představuje nově vytvořený nástroj - Bacterial Explorer, který byl následně propojen s online databází VÚVeLu. Kapitola se zabývá backendem nástroje, frontendem nástroje a testováním nástroje. V rámci testování nástroje je testována uživatelská přívětivost, ale také funkčnost nástroje Bacterial Explorer. Testování je provedeno jednak na bakteriích z poskytnuté bakteriálí databáze VÚVeLu, ale také na bakteriích, které byly staženy z NCBI databáze. Konkrétně jsou použity bakterie *Treponema pallidum*, *Helicobacter pylori* a *Escherichia coli Nissle* 1917. Tyto bakterie jsou vhodné k testování, jelikož jsou laboratorně prozkoumány. V rámci testování je také provedena diskuse.

1 Bakteriální typizace

Bakteriální typizace hraje klíčovou roli při výrobě probiotik. Probiotika jsou živé mikroorganismy, které jsou při požití v dostatečném množství prospěšné pro zdraví hostitelského organismu [1]. Přesná identifikace bakteriálních kmenů je nezbytná k vyloučení patogenních a rezistentních druhů, což je důležité pro ochranu zdraví spotřebitelů. Každý bakteriální kmen má specifické vlastnosti, které přinášejí různé zdravotní přínosy. Bakteriální typizace umožňuje výběr těch, které jsou nejvhodnější pro dané terapeutické účely [1], [2].

Identifikace bakteriálních kmenů je také klíčovým krokem v prevenci a řízení infekčních onemocnění. Studium vnitrodruhové variability umožňuje lékařům a vědcům lépe porozumnět vlastnostem a chování těchto bakterií, což vede k efektivnějšímu boji proti rizikovým bakteriím. [3]

Vnitrodruhová variabilita je způsobena zejména horizontálním přenosem genů, ztrátou nebo akvizicí genů a rekombinací [4]. Bakteriální typizace je klíčovým nástrojem pro identifikaci různých kmenů v rámci jednoho druhu [5]. V závislosti na konkrétní situaci může být bakteriální typizace provedena na různých úrovních - lokálně, regionálně nebo mezinárodně. Lokální úroveň zahrnuje primární laboratoře, na regionální úrovni můžou být vzorky zaslány do referenční laboratoře a na mezinárodní úrovni dochází ke spolupráci v rámci mezinárodních sítí a organizací [6].

Bakteriální typizace obecně zahrnuje dva přístupy - fenotypové a genotypové metody [6]. V současnosti jsou tradiční fenotypové metody postupně nahrazovány modernějšími genotypovými metodami. Nicméně neexistuje univerzální typizační metoda, která by byla vhodná pro všechny situace. Proto je nezbytné vybírat techniky nebo kombinace technik, které nejlépe vyhovují konkrétní situaci. [6], [7]

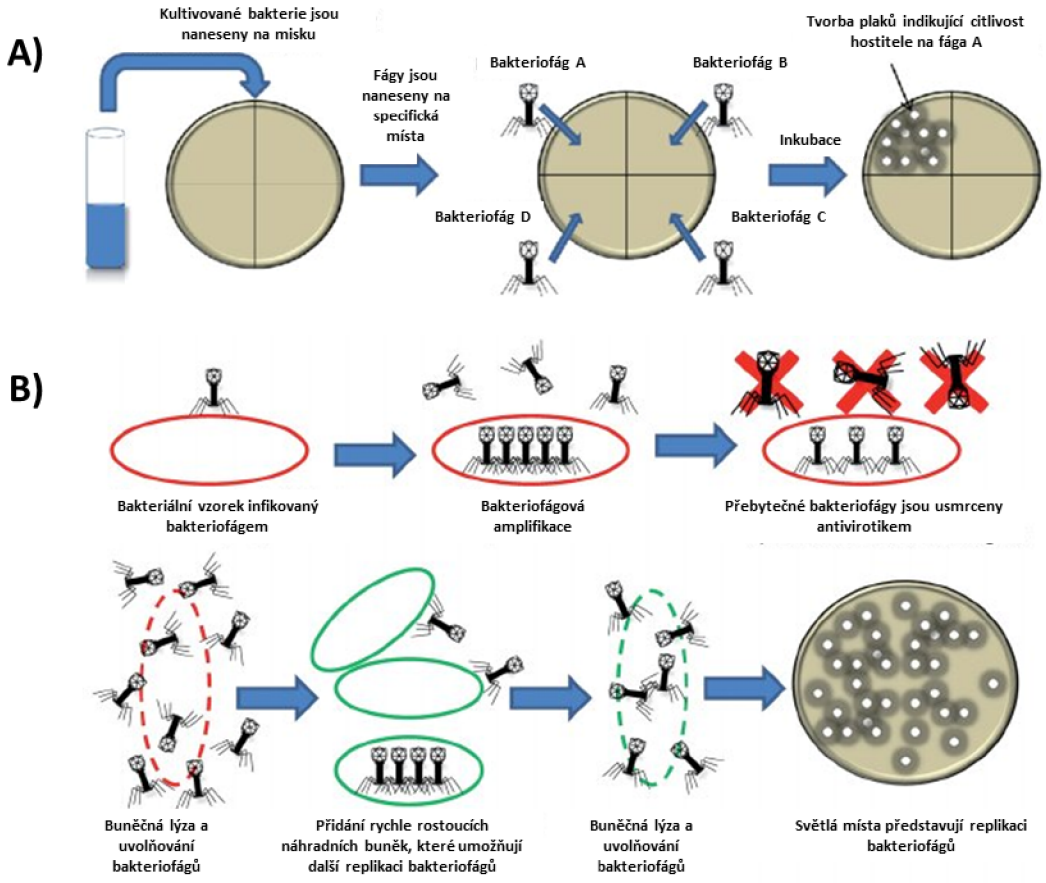
1.1 Fenotypové typizační metody

Fenotypové typizační metody představují nemolekulární metody identifikace bakterií na základě pozorování projevů genové exprese [8]. Projevy genové exprese umožňují sledovat změny chování bakterií v závislosti na okolních podmínkách [8]. Mezi příklady fenotypových parametrů, které jsou využívány při typizaci bakteriálních izolátů, patří biochemické reakce, analýza bakteriofágů, stanovení antimikrobiální rezistence a studium sérologických vlastností [9].

Nevýhody fenotypových metod spočívají v tom, že jsou obvykle pracné a časově náročné, zejména kvůli nutnosti přípravy kultivačních médií a následné kultivaci bakterií. Také jsou příliš variabilní pro praktické využití v epidemiologických studiích [10]. Fenotypové metody jsou schopny rozlišit pouze druhy s výraznými rozdíly v expresi fenotypových genů [8]. Vzhledem k tomu, že většina infekčních patogenů

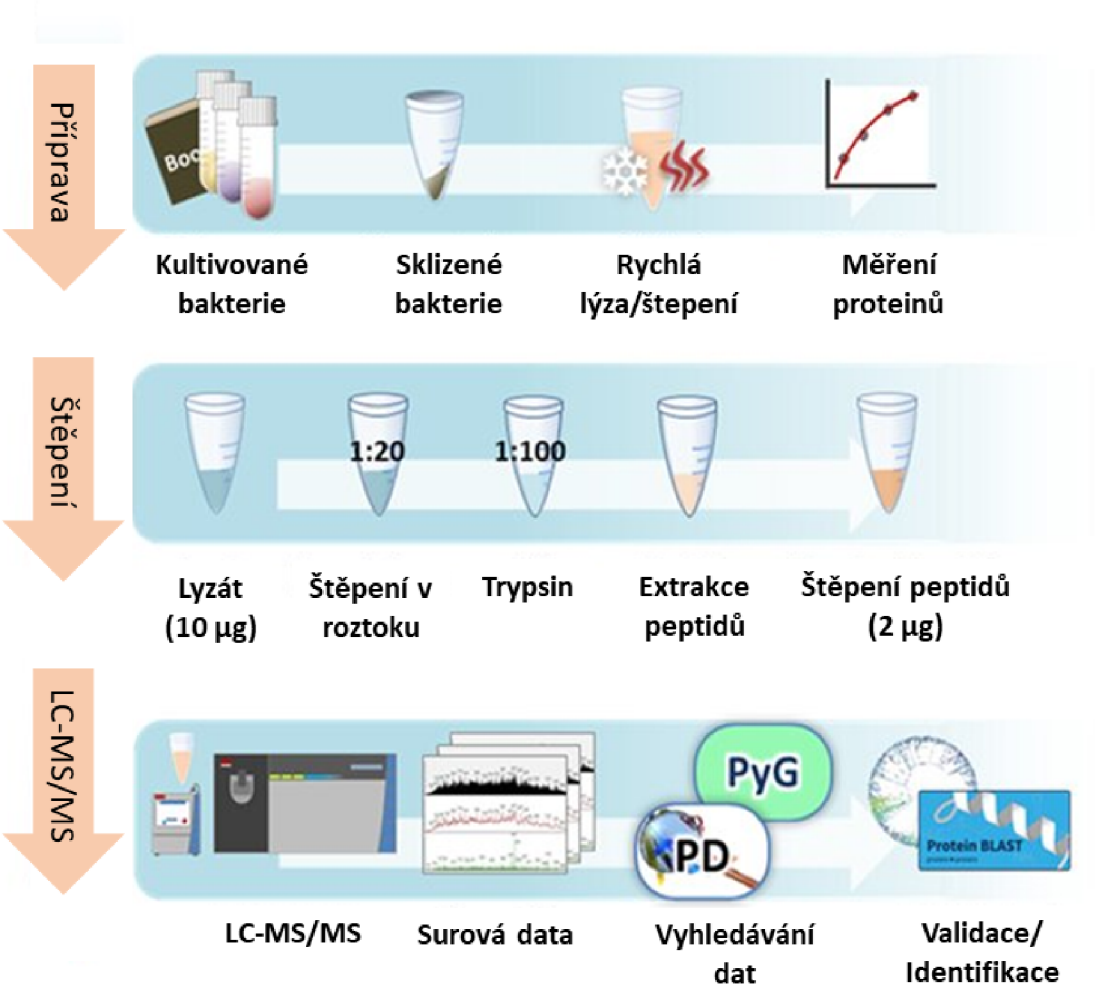
představuje menší podskupinu uvnitř kmenů jednoho druhu, mohou tyto metody nedostatečně zachytit jejich genetickou rozmanitost. Další nevýhodou fenotypových metod je, že často vyžadují předběžné znalosti o bakteriálních izolátech, které mají být typizovány [8].

Mezi nejběžněji používané typizační metody patří biotypizace, sérotypizace, typizace na základě antibiogramu, fágová typizace a proteomická typizace [6]. Biotypizace je metoda, která se opírá o kultivační a biochemické testy. Tuto typizaci lze provádět prostřednictvím testů prováděných na makrozkumavce. Sérotypizace slouží k identifikaci a rozlišení bakteriálních kmenů na základě odlišných antigenních determinant na povrchu buněk. Typizace na bázi antibiogramu se provádí buď difúzí léčiva na pevném agaru nebo ředěním léčiva v tekutém médiu. Hlavním cílem této metody je určování rezistence bakterií vůči antibiotikům. Fágová typizace využívá rozdílnou citlivost bakteriálních kmenů k určitým druhům bakteriofágů. Na obrázku 1.1 je zobrazen princip a postup této metody. [6]



Obr. 1.1: Fágová typizace. V části A) je znázorněn princip fágové typizace, v části B) je znázorněn princip replikačních testů bakteriofágů. Obrázek byl převzat a modifikován z [11].

Proteomické metody (viz obrázek 1.2) jsou založeny na výkonné hmotnostní spektrometrii, která umožňuje analýzu proteinů bakteriálních kmenů [6]. Každá z těchto metod má své výhody i nevýhody a může být vhodná pro konkrétní epidemiologické nebo vědecké účely. Avšak v současné době jsou fenotypové metody postupně nahrazovány genotypovými metodami [6].



Obr. 1.2: Princip proteomických metod založených na hmotostní spektrometrii. Kultivované bakterie jsou štěpeny pomocí enzymů (např. trypsinu) na peptidy. Následně jsou peptidy extrahovány a očištěny od přebytečných solí. Poté je provedena hmotnostní spektrometrie a následná analýza dat. Obrázek byl převzat a modifikován z [12].

1.2 Genotypové typizační metody

Genotypové metody představují molekulární metody typizace, které slouží především k analýze genetických rozdílů mezi bakteriálními izoláty [6]. Rozvoj molekulární genotypizace umožnil klasifikaci mikroorganismů na úrovni druhu a poddruhu. Do té doby byla jedinou metodou umožňující klasifikaci poddruhu sérotypizace, která rozlišuje bakteriální druhy na sérotypy. [6], [8]

Hlavními výhodami genotypových metod, oproti fenotypovým, jsou rychlost analýzy a schopnost poskytovat vysokou diskriminaci mezi různými izoláty [6]. Mnoho genotypových metod nabízí také přístup k rozsáhlým databázím s referenčními genotypy [13], [14], [15], [16], což zvyšuje reprodukovatelnost výsledků a umožňuje mezinárodní srovnání dat [6]. Obecně lze molekulární metody rozdělit na metody založené na sekvenování a metody, které nejsou založené na sekvenování tzv. nesekvenační genotypové metody [6].

1.2.1 Nesekvenační genotypové metody

Nesekvenační genotypové metody rozlišují zkoumané kmeny na základě rozdílů ve velikosti DNA fragmentů, které vznikají amplifikací genomové DNA nebo štěpením DNA pomocí restrikčních enzymů [3]. Mezi nesekvenační genotypové metody patří například pulzní gelová elektroforéza (PFGE), ribotypizace nebo analýza polymorfismu délky restrikčních enzymů (RFLPs) [17]. Tyto techniky jsou však postupně nahrazovány typizací založenou na celogenomovém sekvenování (WGS) [17].

Pulzní gelová elektroforéza

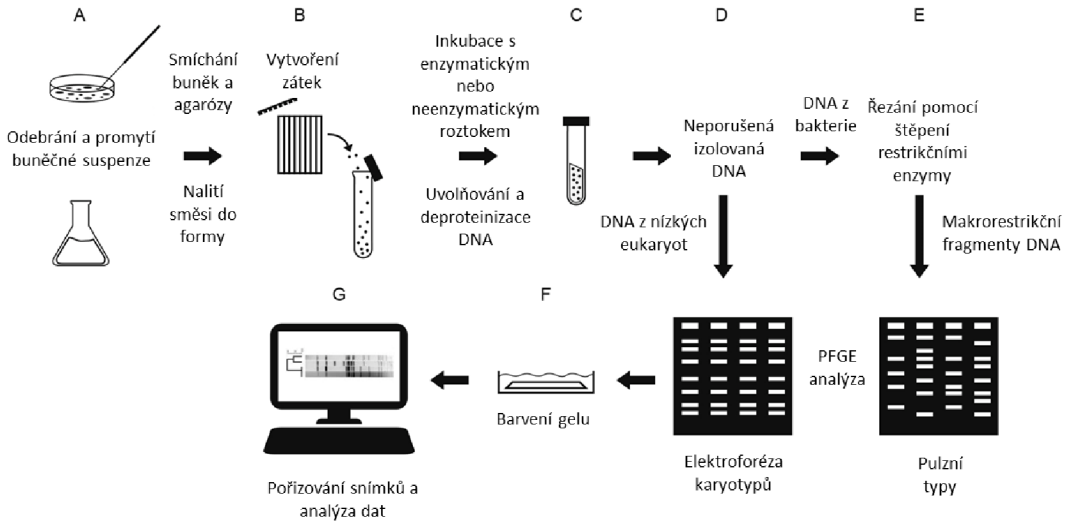
Pulzní gelová elektroforéza (PFGE) je technika, která je široce uznávaná jako "zlatý standard" pro bakteriální typizaci [6]. Její popularita vychází z vysoké diskriminační schopnosti, nízkých nákladů, vysoké epidemiologické relevantnosti a vnitrolaboratorní reprodukovatelnosti [5]. Tato metoda se výrazně uplatňuje při kontrole infekcí a sledování přenosu patogenů v nemocničním prostředí [18].

Porovnání výsledků PFGE z bakterií izolovaných od pacientů, zdravotnických pracovníků a prostředí nemocnice umožňuje identifikovat zdroje infekce, zastavit šíření patogenů a tím přispět k prevenci nemocí a snížení úmrtnosti [19]. Pravidelné odebírání vzorků od zdravotnických pracovíků a monitorování prostředí nemocnice spolu s PFGE typizací umožňuje identifikaci a sledování dominantních kmenů, které mohou přežívat v nemocničním prostředí [20] [21].

Základem PFGE je princip časově řízené změny orientace migrace DNA v závislosti na velikosti fragmentů [18]. Tato periodická změna polarity elektrického pole

umožňuje oddělit molekuly DNA různé délky [18]. Protokoly PFGE byly v průběhu posledních dvou desetiletí vyvinuty a standardizovány, což umožnilo efektivní mezilaboratorní srovnání v rámci sítí, jako je například PulsNet USA Network [6].

Obecně se jednotlivé protokoly liší především v metodách lýzy bakteriálních buněk, výběrem restrikčního enzymu, nastavením parametrů spínače a délkou elektroforézy. Většina těchto protokolů využívá pufr 50mM Tris [18]. Pro konkrétní typy bakterí jsou používány různé protokoly, které specifikují restrikční enzymy a optimální podmínky. Toto rozmanité použití protokolů umožňuje srovnání pulzních typů mezi různými laboratořemi a usnadňuje identifikaci bakteriálních kmenů spojených například s potravinovými onemocněními [18]. Obecný postup procesu přípravy DNA pro PFGE je popsán na obrázku 1.3.



Obr. 1.3: Obecné schéma Pulsní gelové elektroforézy. Obrázek byl převzat a modifikován z [22].

Jako alternativy k PFGE byly zavedeny typizační metody založené na polymerázové řetězové reakci (PCR). Tyto typizační metody jsou robustnější a snadněji proveditelné. Příkladem jsou metody ERIC PCR, rep-PCR a MLVA. Přesto se tyto metody v současnosti používají pouze pro výzkumné účely nebo jako doplňková technika k PFGE v referenčních centrech. [18]

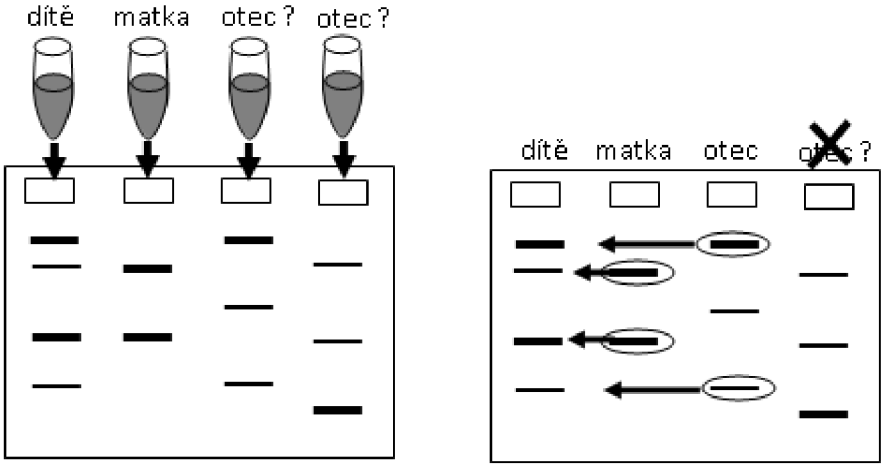
Polymorfismus délky restrikčních fragmentů

Polymorfismus délky restrikčních fragmentů (angl. Restriction Fragment Length Polymorphism = RFLP) se používá pro analýzu unikátních vzorů ve fragmentech DNA

za účelem genetické diferenciace mezi organismy [23]. RFLP využívá tyto rozdíly v DNA sekvencích pro rozpoznání a studium genetických variací jak uvnitř druhu, tak mezi druhy [24].

Restrikční endonukleázy štěpí DNA na krátké fragmenty [23]. Vzdálenost mezi místy štěpení konkrétní restrikční endonukleázy se může mezi jednotlivci lišit, což vede k různým délkám DNA fragmentů produkovaných touto endonukleázou jak mezi jednotlivými organismy, tak mezi různými druhy [25]. Prvním krokem analýzy RFLP je extrakce a čištění DNA. Vyčištěná DNA je poté štěpena pomocí restrikčních enzymů, které obvykle rozpoznávají určité sekvence délky 6-8 bp [23]. Fragmentace DNA je následně analyzována pomocí gelové elektroforézy, během které jsou fragmenty odděleny na základě jejich velikosti a náboje. Během elektroforézy migrují menší fragmenty rychleji než větší fragmenty směrem ke kladné elektrodě [23]. To vede k rozdělení vzorků DNA do různých pásů na gelu. Následně je gel ošetřen luminiscenčními barvivy, aby byly pásy DNA viditelné [23].

RFLP bylo dříve používáno v aplikacích, jako jsou testy otcovství (viz 1.4), forenzní analýza, medicína a genetické studie [24]. S nástupem PCR byla tato metoda postupně nahrazována metodami založenými na PCR [23]. Postup RFLP vyžaduje řadu kroků a trvá týdny, než jsou vidět výsledky, zatímco PCR poskytuje výsledky během několika hodin. Proto PCR výrazně nahradila RFLP ve většině aplikací vyžadujících sekvenování DNA, jako jsou testy otcovství nebo forenzní analýza vzorků [23].



Obr. 1.4: Určení otcovství na základě metody RFLP: Izolovaná DNA je použita k amplifikaci specifické sekvence. Produkt amplifikace (amplikon) je následně rozštěpen enzymem a výsledné fragmenty jsou separovány gelovou elektroforézou. Výsledné fragmenty lze využít k vyloučení nebo potvrzení otcovství. Obrázek byl převzat z [26].

1.2.2 Sekvenační genotypové metody

Sekvenování celých genomů (angl. Whole-Genome Sequencing = WGS) představuje moderní přístup v analýze sekvencí celých organismů[6]. Tato metoda se stala klíčovým nástrojem v obasti bakteriální genomiky a umožňuje detailní studium genetických vlastností bakteriálních kmenů[6].

Původní Sangerovo sekvenování [27] bylo velmi nákladné a časově náročné. To vedlo ke vzniku nových sekvenačních technologií, označovaných jako sekvenování nové generace (angl. Next Generation Sequencing = NGS), a sekvenování třetí generace (angl. Third Generation Sequencing = 3GS), které je založené na sekvenování jediné molekuly DNA [28], [29], [30]. Tyto technologie přinesly zásadní zlom ve WGS, jelikož systémy NGS jsou rychlejší, cenově dostupnější a umožňují sekvenování bakteriálních genomů během jediného dne [6].

Celogenomové sekvenování poskytuje obrovské množství dat. Díky tomu mají typizační metody založené na WGS vysokou diskriminační schopnost [31]. Metody typizace založené na WGS se staly žádanými technikami v mnoha aplikacích, včetně predikce fenotypu, genomických studií nebo vývoje vakcín [6].

WGS data tvoří technologický základ pro různé metody typizace. Příkladem je jednolokusová sekvenční typizace (angl. Sigle Locus Sequencs Typing = SLST), multilokusová sekvenční typizace (angl. Multi Locus Sequence Typing = MLST), MLST jádra genomu (angl. core genome = cgMLST), MLST celého genomu (angl. whole genome = wgMLST) a detekce jednonukleotidových polymorfismů (angl. Single Nucleotide Polymorphism = SNP), s různou mírou rozlišení [6].

Multilokusová sekvenční typizace

Multilokusová sekvenční typizace je jednou z nejpopulárnějších genotypových metod pro charakterizaci bakteriálních kmenů. Tato technika se zaměřuje na sekvenování souboru fragmentů DNA o délce přibližně 400 - 600 bp [6]. Počet zkoumaných lokusů závisí na konkrétním výzkumu, ale obvykle se používá sedm lokusů [6], [3].

Pro dosažení dostatečného rozlišení mezi blízce příbuznými bakteriemi byla vyvinuta schémata MLST [35], [36] pro bakterie patřící do stejného rodu (nebo dokonce i druhu) [37]. Schémata umožňují srovnání genotypů s referenčními databázemi, které jsou snadno dostupné prostřednictvím internetu, což umožňuje rychlé a efektivní srovnání alelických profilů mnoha kmenů [38] .

V rámci MLST je prvním krokem použití PCR k amplifikaci vnitřních segmentů genů z chromozomální DNA pomocí párů primerů uvedených na webových stránkách [6], [39]. Získané fragmenty jsou následně sekvenovány v obou směrech. Poté jsou sekvence porovnány s existujícími alelami na daném lokusu, a každému lokusu je přiřazeno číslo odpovídající specifické alele [6], [39].

Sekvenování více lokusů přináší MLST výhodu vyšší diskriminační schopnosti a poskytuje významný obraz ohledně skutečné evoluční historie kmene, čehož není možné dosáhnout při použití jediného lokusu [6], [40]. Nicméně, MLST má také několik nevýhod. Alelám jsou přiřazena čísla, která nemusí plně odpovídat skutečné genové sekvenci [3]. To činí fylogenetickou analýzu méně spolehlivou. Kromě toho používání vysoce konzervovaných genů v MLST někdy nepostačuje k detekci variability u blízce příbuzných kmenů. Navíc, sekvenování sedmi lokusů je časově náročné a finančně nákladné [3], [6].

Bakteriální typizace analýzou jednonukleotidových polymorfismů

Analýza jednonukleotidových polymorfismů rozlišuje bakteriální izoláty na základě mutací na konkrétních místech v jejich genomu. Analýza SNP identifikuje jednotlivé izoláty a určuje jejich vzájemnou příbuznost [6]. Výsledky analýzy závisí na typu a poloze změn nukleotidů v genomu. Mohou vznikat tzv. synonymním mutace, kdy dojde k mutaci, ale aminokyselina se nemění, nebo mohou vznikat nesynonymní mutace, při kterých se mění aminokyselinová sekvence [41]. Zejména synonymní mutace jsou využívány k určení evolučního původu mikroorganismů a k identifikaci izolátů, které jsou si vzájemné blízce příbuzné [41]. Příbuznost kmenů lze objasnit zkoumáním více SNP [41].

Analýza SNP porovnávává izoláty s jedním referenčním genomem, což umožňuje identifikaci jednonukleotidových mutací. V této analýze je důležité, aby byl referenční genom úzce spojen s genomy zkoumaných izolátů, což zajišťuje spolehlivost identifikace fylogeneticky relevantních SNP. Jinak by hrozilo podhodnocení výskytu SNP v důsledku rozdílů mezi izolátem a zvoleným referenčním genomem [42], [43]. Samotný proces identifikace SNP se skládá ze dvou hlavních kroků. Prvním krokem je mapování surových sekvencí na referenční genom. Následně je provedena identifikace SNP s využitím přísných kritérií, která minimalizují riziko chybných zarovnání nebo chyb v procesu sekvenování [6], [44], [45].

2 Metody analýzy genomických dat

Analýza genomických dat je stěžejní pro pochopení genetických mechanismů a funkčních a evolučních vztahů. S rostoucím množstvím dat byly postupně vyvinuty algoritmy, které umožňují rychlejší a efektivnější práci s objemnými genomickými daty. Kapitola přináší přehled algoritmů, které jsou na pozadí bioinformatických nástrojů pro účely bakteriální typizace.

2.1 Hladový algoritmus

Hladový algoritmus hraje v bioinformatice a genomice významnou roli, jelikož rychle vyhledává a analyzuje biologické sekvence s ohledem na jejich podobnost [46]. Hladový algoritmus postupně vybírá nejlepší řešení v daný okamžik, aniž by bral v úvahu budoucí důsledky [47]. V bioinformatice se tento přístup používá k rychlému vyhledávání podobných sekvencí v rozsáhlých genomických databázích [48].

Hladový algoritmus neprovádí komplexní zarovnání sekvencí, ale pracuje s krátkými úseky sekvencí nebo-li semínky. Semínka představují krátké úseky sekvencí, které slouží jako počáteční body pro vyhledávání podobností [49]. Tento přístup umožňuje analýzu velkých datových sad a identifikaci homologních sekvencí [48], což je důležité např. pro vyhledávání genů [50], identifikaci kódovaných proteinů [51] a nebo analýzu evolučních vztahů [46].

2.2 Heuristické vyhledávání podobných sekvencí

Heuristika hraje v bioinformatice klíčovou roli při analýze genomických dat, zejména při vyhledávání podobných sekvencí [52]. Heuristický přístup zahrnuje použití různých zjednodušených strategií a triků, které umožňují rychlejší a efektivnější analýzu biologických dat, aniž by byla prováděna výpočetně náročná zarovnání sekvencí [52].

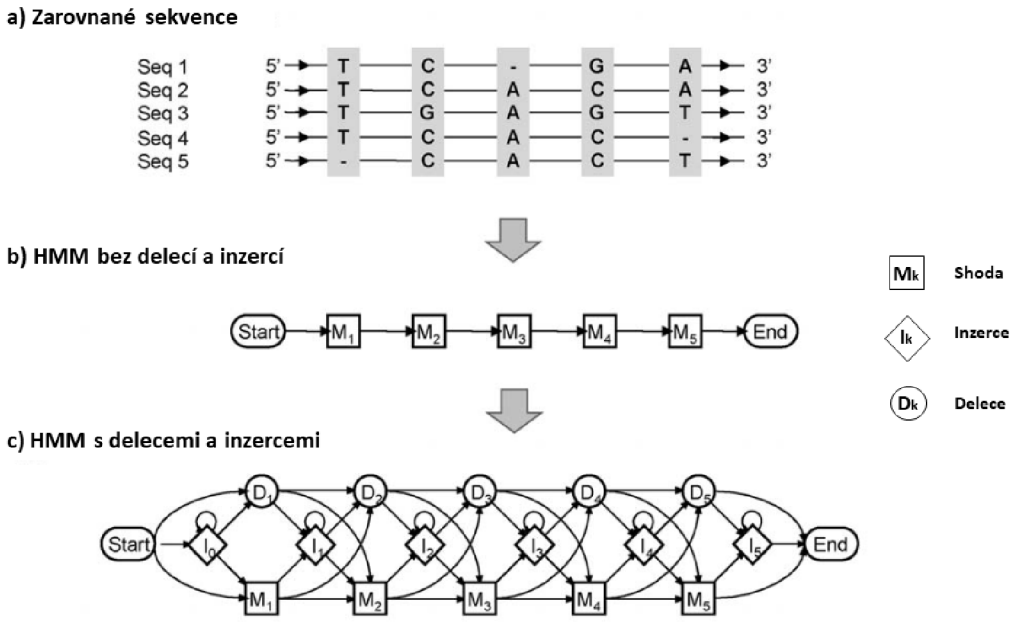
Heuristické přístupy často pracují např. s principy filtrace nízké podobnosti [53], použitím semínek [54] a hledáním lokálních shod [55]. Filtrace nízké podobnosti eliminuje sekvence s nízkou pravděpodobností relevantní podobnosti. Semínka představují krátké úseky sekvencí, které slouží jako počáteční body pro vyhlednávání shod [54]. Často se také heuristické přístupy zaměřují na hledání lokálních shod, což je důležité pro identifikaci funkcí nebo vzorů [55].

2.3 Skrytý Markovovův model

Skrytý Markovův model (angl. Hidden Markov Model = HMM) představuje sofostikovaný matematický model, který má široké využití v oblasti analýzy genomických dat [56]. Jedná se o pravděpodobnostní model umožňující generování sekvencí a identifikaci vzorů v sekvencích. V bioinformatice zaujímá HMM zásadní postavení, zejména díky své schopnosti pracovat s pravděpodobnostními stavy [57].

HMM se skládá z konečného počtu stavů, kde každý stav reprezentuje určitý biologický kontext nebo vzor. Tyto stavy jsou spojeny přechody s pravděpodobnostmi, což určuje možný vývoj v analýze sekvencí [57]. Každý stav HMM může emitovat pozorování, což jsou konkrétní hodnoty nebo znaky v sekvenci. Tato emitovaná pozorování jsou spojena s pravděpodobnostmi pro každý stav. Ovšem některé stavy mohou být skryté, což odráží realitu v biologických datech, kde taky nejsou všechny informace zřejmé [56].

HMM nachází uplatnění např. při vícenásobném zarovnání sekvencí [57]. Oproti tradičním metodám vícenásobného zarovnání poskytují metody vícenásobného zarovnání založené na HMM mnoho výhod. Tradiční metody vícenásobného zarovnání vyžadují znalost skórovacích parametrů. HMM mohou být trénovány na nezarovnaných sekvencích a poté můžou být použity k vytvoření vícenásobného zarovnání, což vede k eliminaci potřeby obtížně volených skórovacích parametrů [56]. Princip vícenásobného zarovnání pomocí HMM je na obrázku 2.1.



Obr. 2.1: Princip vícenásobného zarovnání pomocí HMM. Obrázek byl převzat a modifikován z [58].

Další využití mají HMM například při rozpoznávání genů v DNA sekvencích [59]. Modely mohou identifikovat introny, exony nebo další funkční prvky v genomu. HMM lze také použít k provádění profilových analýz proteinových sekvencí nebo můžou predikovat sekundární struktury proteinů [60]. Také mohou být použity k analýze evolučních vztahů mezi různými druhy [57].

2.4 Algoritmus mapování bez zarovnání

Algoritmus mapování bez zarovnání (angl. Alignment-free mapping algorithm = AFMA) představuje moderní přístup v oblasti bioinformatické analýzy genomických dat [61]. Tento algoritmus se liší od tradičních metod zarovnání sekvencí tím, že nevyžaduje přesné zarovnání celých sekvencí, ale spoléhá se na jiné charakteristiky a vlastnosti dat [62]. Tento algoritmus umožňuje rychlé mapování sekvencí na referenční genomy [61].

AFMA používá k zarovnání k-mery, což jsou krátké úseky o pevné délce [63]. Tato metoda umožňuje rychle a efektivně určit podobnost mezi sekvencemi na základě výskytu a četnosti k-merů. Četnosti k-merů jsou pak porovnány mezi sekvencemi a referenčními genomy, což umožňuje určit, zda jsou sekvence podobné nebo odlišné [61].

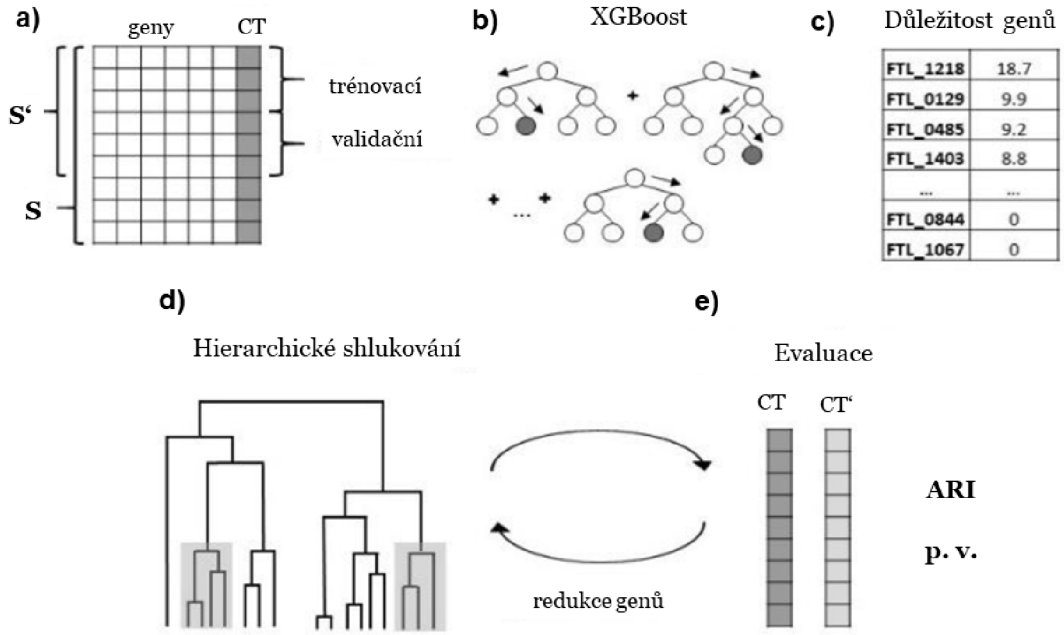
AFMA je důležitý nástroj v oblasti analýzy genomických dat a přispívá k rychlému a efektivnímu mapování sekvencí a analýze podobnosti mezi biologickými daty [61]. Jeho využití v bioinformatických aplikacích přináší nové možnosti pro studium genomických dat.

2.5 Algoritmus minMLST

Algoritmus minMLST, zobrazený na obrázku 2.2, představuje optimalizační algoritmus založený na strojovém učení [64]. Hlavním úkolem minMLST je minimalizace počtu genů ve schématech cgMLST pomocí identifikace informativních genů a analýzy kompromisu mezi redukcí počtu genů a výkonnostní typizací [64]. Schémata cgMLST představují specifické soubory genů, které jsou využívány pro bakteriální typizaci. Tyto soubory obsahují lokusy, které jsou považovány za jaderný genom dané skupiny mikroorganismů.

Identifikace informativních genů je založena na algoritmu XGBoost. XGBoost představuje soubor klasifikačních a regresních stromů, které postupně zlepšují chyby předchozích klasifikátorů. K vyhodnocení důležitosti genů se používají různé podobnostní metriky, např. SHAP hodnoty (angl. Shapley Additive explanations = SHAP), váha, pokrytí nebo zisk. Váha udává, kolikrát je gen použit k rozdělení

dat v průběhu tvorby rozhodovacích stromů. Zisk vyjadřuje průměrnou nebo celkovou účinnost daného genu při snižování chybovosti modelu. Pro porovnání hodnot důležitosti genů je pak využit Pearsonův a Spearmanův korelační koeficient a upravený náhodný index (angl. Adjusted Rand Index = ARI). MinMLST tedy kombinuje algoritmy strojového učení s podobnostními metrikami za účelem efektivní analýzy a optimalizace schémat cgMLST. [64]



Obr. 2.2: Postup metody minMLST. (a) Filtrování typů shluků s jedním izolátem z původního schématu cgMLST a následné rozdělení izolátů na trénovací a validační množinu. (b) Trénování klasifikátoru XGBoost, dokud není dosaženo minimální ztrátové funkce. (c) Kvantifikace důležitosti genu v natrénovaném modelu XGBoost pomocí zvolené míry (SHAP, váha, přírůstek). Iterativní opakování kroků (d) a (e) pro snížený počet nejdůležitějších genů: (d) provedení typizace kmenů všech izolátů ve schématu pomocí hierarchického shlukování založeného na vzdálenosti. (e) Vyhodnocení výkonnosti typizace použitím testu významnosti na upravený ARI, porovnání typů vyvolaných minMLST a základních referenčních typů shluků předdefinovaných v původním schématu cgMLST. Obrázek byl převzat a modifikován z [64].

3 Sbírka bakteriálních genomů

Testování vytvořeného nástroje, jako jednoho z hlavních cílů této diplomové práce, bylo provedeno na sbírce VÚVeL. Tato sbírka obsahuje celkem 452 bakteriálních izolátů [70], [71], [72], z nichž 398 pochází z *caeca* kuru domácího a 54 z prasečího trusu. Izoláty představují zástupce 8 různých kmenů, což nám poskytuje komplexní pohled na mikrobiom tohoto prostředí. Každý z těchto izolátů byl podroben sekvenaci a následně byla provedena analýza genomického obsahu a taxonomického zařazení [71].

3.1 Metodika přípravy a uchování bakteriálních izolátů

Přípravu 133 bakteriálních izolátů z kuřecích slepých střev popisuje studie Medveckého a kol. [71]. Stejným způsobem byly připraveny i zbývající izoláty, které byly doplněny do sbírky VÚVeLu později, jak popisují studie [70], [72]. Slepá střeva kuru domácího pocházela od náhodně vybraných kohoutů a slepic ve věku od 4 do 40 týdnů.

Vzorky byly sériově zředěny v předpřipraveném redukovaném PRAS dilučním roztoku a naneseny na Wilkins-Chalgrenův anaerobní agar (WCHA) doplněný 30 % bachorové tekutiny. Bachorová tekutina byla krávám odebrána orální sondou a následně byla přefiltrována přes bavlněnou tkaninu, odstředěna při 8000 g po dobu 30 minut a sterilizována filtrací přes filtr s velikostí pórů 0,22 μm. [71]

Po pětidenní inkubaci při 37 °C bylo z každé agarové plotny odebráno přibližně 10 dobře oddělených kolonií různé morfologie, které byly přečištěny subkultivací na WCHA [71]. Všechny izoláty byly uchovány při -80 °C v redukovaném PRAS dilučním roztoku obsahujícím glycerol o koncentraci 20 % a stejný objem ovčí krve [71].

Izolovaná DNA byla přečištěna pomocí DNeasy Blood & Tissue Kit (Quiagen). Sekvenační knihovna byla připravena z 1 ng genomové DNA bez RNA pomocí sady Nextera XT DNA Sample Preparation Kit (Illumina) [71]. Sekvenování celého genomu bylo provedeno pomocí platformy NextSeq 500 v režimu párových čtení (2 x 150 bp) za použití kitu NextSeq 500/550 High Output Kit v2 (Illumina) [71]. Surová sekvenační data byla ořezána nátrojem Trimmomatic [73] a ořezaná párová čtení byla sestavena pomocí *de novo* assembleru IDBA-UD v1.1.1 [74].

3.2 Předzpracovaná taxonomická identifikace a funkční analýza bakteriálních izolátů

Bakteriální druhy byly definovány na základě porovnání celých sekvencí 16S rRNA pomocí nástroje BLAST se záznamy uloženými v databázi sekvencí 16S rRNA NCBI. Jednotlivé druhy byly identifikovány na základě nejnižší e-hodnoty [71] . Ze 133 izolátů byla pro 15 izolátů podobnost s jinou sekvencí nižší než 94 %, což může naznačovat možnost nových druhů [71]. Všechny sekvence 16S rRNA byly porovnány také s databází RDP SeqMatch, což umožnilo alternativní taxonomii včetně zařazení jednotlivých izolátů do vyšších taxonomických jednotek [71].

Kromě toho byly k ověření taxonomické klasifikace použity databáze multilokusové sekvenční typizace ribozomálních proteinů (rMLST) [75] a databáze GTDB pro identifikaci organismů [71] . Predikce genů a funkční anotace byly provedeny pomocí nástroje RAST [76]. Sestavené a anotované genomy i surová sekvenační data jsou uložena na NCBI pod přístupovým číslem PRJNA377666 [71].

3.3 Vlastní fylogenetická analýza

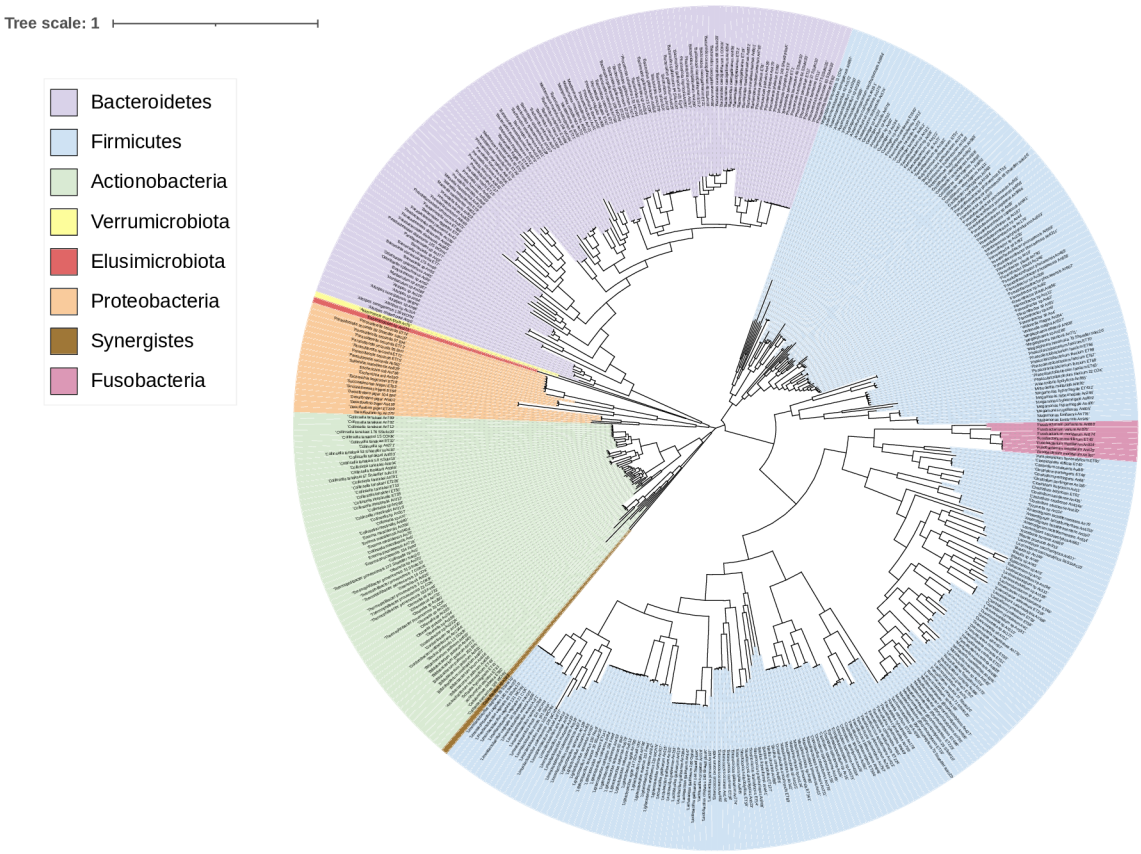
Za účelem hlubšího porozumnění rozmanitosti testované databáze byla provedena fylogenetická analýza. Pomocí nástroje UBCG v3.0 [77] byl vytvořen fylogenetický strom všech 452 bakteriálních izolátů. Fylogenetický strom byl následně graficky upraven pomocí interaktivního nástroje iTOL (angl. interactive tree of life), s cílem diferencovat bakterie na základě jejich příslušnosti k danému kmenu. Výsledný fylogenetický strom je zobrazen na obrázku 3.1.

Na základě fylogenetické analýzy bylo ověřeno, že testovací dataset obsahuje bakterie z 8 různých kmenů - *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, *Proteobacteria*, *Verrucomicrobia*, *Elusimicrobia*, *Synergistetes* a *Fusobacteria*. *Firmicutes* a *Bacteroides* představují dominantní kmeny mikrobiomu [79], [80]. *Firmicutes* patří mezi gram pozitivní bakterie [81], zatím co *Bacteroides* se řadí mezi gram negativní bakterie [82]. Oba kmeny se podílí na štěpení polysacharidů a regulaci energetické bilance.

Actinobacteria patří mezi gram pozitivní bakterie [83]. Tyto bakterie jsou významným producentem antibiotik, enzymů, signálních molekul a imunomolekul [84]. Mohou tak ovlivňovat celkovou imunitní odpověď hostitele.

Proteobacteria, *Verrumicrobiota*, *Elusimicrobiota*, *Synergistetes* a *Fusobacteria* jsou gram negativní bakterie [82], [85]. *Proteobacteria* reagují na různé stresové podněty a často bývají spojovány se zánětlivými onemocněními [87]. *Verrumicrobiota* se podílí na udržení slizniční integrity [88]. *Elusimicrobiota* se pravděpodobně podílí na fermentaci a rozkladu organických látek [89]. *Synergistetes* působí protizánětlivě

a udržují homeostázu a rovnováhu mikrobiomu [90]. Zvýšený výskyt *Fusobacterií* indikuje žaludeční onemocnění (např. žaludeční vředy) [91].



Obr. 3.1: Fylogenetický strom 452 sekvenovaných genomů získaných ze slepých střev kuru domácího a prasečího trusu. Modrou barvou jsou znázorněny bakterie z rodu *Firmicutes* (245 genomů), fialovou barvou jsou bakterie z rodu *Bacteroidetes* (113 genomů), zelenou barvou jsou bakterie z kmene *Actinobacteria* (65 genomů), žlutou barvou je kmen *Verrumicrobiota* (1 genom), červenou barvou *Elusimicrobiota* (1 genom), oranžovou jsou bakterie z kmene *Proteobacteria* (19 genomů), hnědou barvou je znázorněn kmen *Synergistes* (1 genom) a růžovou barvou je kmen *Fusobacteria* (7 genomů). Z fylogenetického stromu můžeme pozorovat, že *Firmicutes* a *Bacteroides* jsou dominantními kmeny mikrobiomu.

4 Dostupné nástroje pro bakteriální typizaci

Softwarové nástroje hrají v bakteriální typizaci klíčovou roli, jelikož umožňují identifikaci, charakterizaci a srovnání bakteriálních organismů na genetické úrovni. Tato kapitola představuje několik významných nástrojů, které jsou široce využívány v rámci bakteriální typizace. Mezi tyto nástroje patří BLAST, Cd-hit, Barrnap, FastANI a BLAT. Všechny nástroje, byly následně použity při tvorbě nového nástroje Bacterial Explorer.

4.1 BLAST

BLAST (Basic Local Alignment Search Tool) [92] slouží k vyhledávání lokálních podobností mezi proteinovými nebo nukleotidovými sekvencemi [93]. Tento nástroj je "vládním dílem Spojených států amerických" a na jeho používání se nevztahují žádná omezení [94]. Od svého vzniku prošel BLAST velkým vývojem. Zpočátku byl určený pro analýzu proteinových sekvencí, ale později byl rozšířen o možnost porovnávání nukleotidových sekvencí [92].

Hlavním úkolem BLASTu je porovnání vstupní sekvence (nukleotidů nebo proteinů) s referenční databází. V našem případě se jedná o poskytnutou databázi VÚVeLu. Toto porovnání je založeno na heuristickém porovnání podobných sekvencí (více výše v kapitole 2.2). [92]

Pro nukleotidové sekvence vyžaduje BLAST přesnou shodu mezi úseky, zatímco pro proteinové sekvence používá skóre shody, které je stanoveno pomocí substituční matice [93]. K identifikaci shod používá BLAST také definovanou prahovou hodnotu [93], [95]. Jestliže je nalezena shoda, pak BLAST prodlužuje zarovnání v obou směrech od nalezené shody, pokud se skóre nadále zvyšuje, nebo dokud neklesne o kritickou hodnotu kvůli negativnímu skóre způsobenému neshodami [93], [95], [96].

Výstupem BLASTu jsou zarovnané sekvence, které ukazují míru podobnosti mezi vstupními a databázovými sekvencemi. Kromě zarovnání poskytuje BLAST také e-hodnotu, která udává, kolikrát bychom náhodně očekávali podobné zarovnání vzhledem k velikosti databáze [93]. E-hodnota je vypočítána podle vzorce:

$$E = (n * m)/(2^{S'}), \tag{4.1}$$

kde n vyjadřuje celkový počet zbytků (aminokyselin nebo nukleových kyselin), m je délka detekované sekvence a S' je bitové skóre. Nižší e-hodnota značí větší významnost nálezu. [93]

V kontextu bakteriální typizace se nástroj BLAST používá k identifikaci nebo charakterizaci konkrétních genů nebo sekvencí v bakteriálním genomu [97]. Například

může být použit k porovnání určitého genomového fragmentu bakterie s databází genomových sekvencí, což umožňuje určit, zda daný fragment patří k určitému druhu nebo kmenu bakterie [97]. BLAST může být také využit jako součást analýzy MLST dat [98].

4.2 Cd-hit

Nástroj Cd-hit je distribuován pod licencí GNU General Public License v2.0[1]. Cd-hit slouží ke shlukování a porovnání proteinových nebo nukleotidových sekvencí. Díky filtrování krátkých slov dokáže rychle a efektivně pracovat s rozsáhlými databázemi. Původní verze Cd-hitu byla představena v roce 2001 [100]. Od té doby prošel nástroj několika transformacemi. Dnes existuje několik variant tohoto nástroje, např. Cd-hit-est, Cd-hit-2D a Cd-hit-2D-est [101].

Původní verze Cd-hitu umožňuje shlukování proteinových databází, zatímco Cd-hit-est je navržen pro shlukování DNA/RNA databází [101]. Obě verze pracují na principu hladové algoritmu (více výše v kapitole 2.1) přírůstkového shlukování. Průběh shlukování začíná seřazením sekvencí sestupně podle délky. Nejdelší sekvence se stává zástupcem prvního shluku. Následně jsou zbývající sekvence porovnány s existujícími zástupci shluků. Pokud jakákoli sekvence prokáže dostatečnou podobnost s některým z existujících zástupců, pak je začleněna do tohoto shluku [101]. Pokud podobnost nepřesáhne stanovenou prahovou hodnotu, vytvoří se nový shluk s touto sekvencí jako zástupcem. Během těchto porovnání sekvencí je také použito filtrování krátkých slov, pomocí kterého je ověřeno, zda podobnost dosahuje požadované hodnoty pro shlukování [101]. Pokud toto není splněno, dochází k provedení skutečného zarovnání sekvencí.

V rámci bakteriální typizace je nástroj Cd-hit často používán pro analýzu bakteriálních genomů. Pomáhá seskupit podobné sekvence do kmenů nebo druhů, což umožňuje rozdělit neznámý dataset do relevantních skupin. Následně lze identifikovat podobné druhy nebo kmeny bakterií v souboru sekvencí a provádět analýzy zaměřené na genetickou rozmanitost mezi různými kmeny. Cd-hit používá např. nástroj Mge-cluster [102].

[1]GNU General Public License v2.0 umožňuje uživatelům svobodně používat, měnit a distribuovat software za podmínky, že všechny odvozené práce zůstanou také pod touto licencí. [99]

4.3 Barrnap

Nástroj Barrnap je distribuován pod licencí GNU General Public License v3.0 [2]. Barrnap (Bacterial ribosomal rRNA predictor) je bioinformatický nástroj používaný pro detekci a anotaci ribozomálních RNA (rRNA) genů v genomických sekvencích bakterií [104]. Tento nástroj je založen na efektivním využití skrytých HMM (více výše v kapitole 2.3), které byly trénovány na známých sekvencích rRNA [104].

Rozpoznání rRNA genů v genomických datech hraje klíčovou roli v bakteriální typizaci a mikrobiologickém výzkumu. Bakterie jsou charakterizovány specifickými sekvencemi rRNA, které mohou různými způsoby rozlišovat jednotlivé bakteriální kmeny. Barrnap umožňuje rychlou a přesnou identifikace těchto genů v různých typech genomických dat. [105], [106], [107].

Při bakteriální typizaci se tento nástroj používá také pro ověření kvality genomických dat [108]. Výskyt rRNA značí dobrou kvalitu dat. Ověření přítomnosti rRNA se proto používá ke stanovení kvality genomických dat [109].

4.4 FastANI

Nástroj FastANI je distribuován pod licencí Apache License 2.0[3]. FastANI je bioinformatický nástroj, který slouží k porovnání a srovnání bakteriálních genomů na základě průměrné nukleotidové identity (ANI) [111]. Nukleotidová identita představuje klíčový parametr pro hodnocení podobnosti mezi bakteriálními genomy a hraje významnou roli v analýze bakteriální diverzity, klasifikaci a taxonimii [111], [112]. Průměrná nukleotidová identita je popsána vzorcem:

$$ANI_{G1 \to G2} = \frac{\sum_{bbh}(procento\ identity * délka\ zarovnání)}{délka\ BBH\ genů}, \tag{4.2}$$

kde BBH označuje obousměrně nejlepší shody (ang. bidirectional best hits = BBH).

Základní postup algoritmu FastANI se řídí postupem, který popsali Goris et al [113]. Pro odhad identity zarovnání používá FastANI Mishmap, který je založený na MinHash [111]. Přesnost tohoto nástroje je na stejné úrovni jako výpočet ANI pomocí nástroje BLAST [111], ale oproti BLASTu dosahuje FastANI dvou až třířádového zrychlení díky algoritmu AFMA (více výše v kapitole 2.4). Proto je tento nástroj ideální pro analýzu velkých souborů bakteriálních genomů.

[2] GNU General Public License v3.0 umožňuje uživatelům svobodně používat, měnit a distribuovat software, zatímco chrání komunitu před právními problémy souvisejícími s patenty a uzavřenými systémy. [103]

[3] Apache License 2.0 umožňuje uživatelům svobodně používat, měnit a distribuovat software. Tato licence nevyžaduje, aby odvozené práce byly distribuovány pod stejnou licencí. [110]

FastANI hraje v bakteriální typizaci klíčovou roli, jelikož umí identifikovat bakteriální druhy, kmeny a izoláty. Tento nástroj dokáže určit, zda dvě nukleotidové sekvence patří ke stejnému druhu nebo kmenu a usnadňuje zkoumání vztahů mezi různými bakteriálmními organismy [114], [115].

4.5 BLAT

Nástroj BLAT (BLAST-Like Alignment Tool) umožňuje zarovnání mRNA/DNA a proteinů napříč druhy [117]. Tento nástroj je dostupný pro osobní, akademické a neziskové použití pod specifickou licencí[4].

V porovnání s ostatními nástroji by měl být BLAT přesnější a až 500 krát rychlejší než stávající nástroje pro zarovnání mRNA/DNA. Rychlost BLATU je dána indexem všech nepřekrývajících se k-merů v genomu. [117]

Nástroj BLAT zahrnuje čtyři hlavní fáze. V první fázi používá index k vyhledávání oblastí genomu, které jsou pravděpodobně homologní s dotazovanou sekvencí. Ve druhé fázi dochází k zarovnání homologních sekvencí. Zarovnané oblasti (často exony) jsou potom spojeny do větších celků (obvykle genů). Nakonec BLAT projde znovu malé vnitřní exony, které byly v první fázi přehlédnuty a pokud je to možné, upraví hranice velkých mezer, které mají kanonická místa sesřihu. [117]

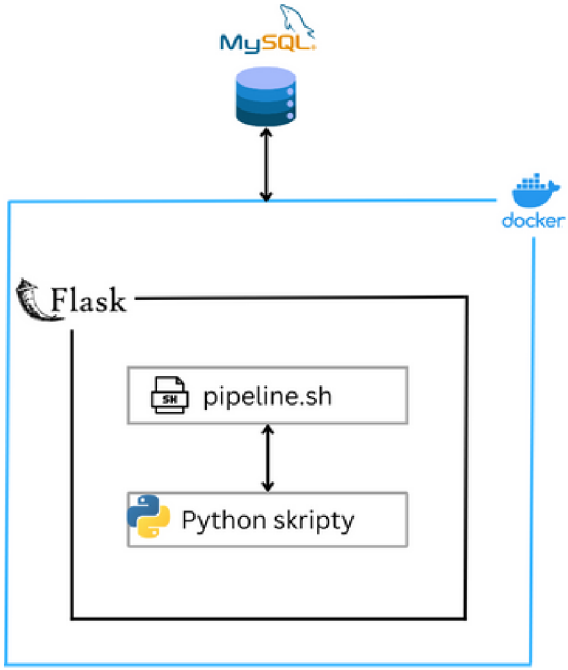
[4]Licence BLATu umožňuje osobní, akademické a neziskové použití zdarma. Komerční uživatelé mají licenci pouze pro obsah adresářů lib a inc. Pro přístup dalších modlů musí kontaktovat jim_kent@pacbell.net. [116]

5 Bacterial Explorer

Bacterial Explorer je automatický softwarový nástroj, který slouží k detekci nových, dosud neobjevených bakterií. Nástroj je propojený s online databází VÚVeLu a je dostupný na adrese `blast.vuvel.eu:5000/login`. Celý vývoj nástroje byl konzultován s odborníky z VÚVeLu. Nástroj byl vytvořen v souladu s požadavky budoucích uživatelů. V rámci této kapitoli je popsán backend, frontend a testování nástroje.

5.1 Backend aplikace

Backend nástroje Bacterial Explorer představuje neviditelnou, avšak zásadní část nástroje, která poskytuje základní podporu a funkcionalitu. Hlavní úlohou backendu je zpracování dat, komunikace s databází a zajištění běhu nástroje jako celku. Struktura backendu nástroje Bacterial Explorer je popsána na obrázku 5.1.



Obr. 5.1: Znázornění struktury backendu nástroje Bacterial Explorer. Celý nástroj, vytvořený pomocí frameworku Flask, je uložen v docker kontejneru, který je propojený s MySQL databází uživatelů. MySQL databáze je na stejném servru, jako je Docker kontejner.

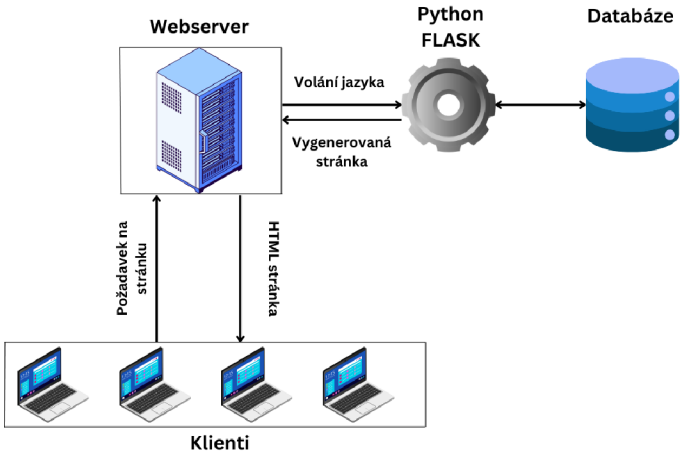
Pro vytvoření backendu nástroje Bacterial Explorer byl použit framework Flask [118], který umožňuje rychlý a efektivní vývoj online webových nástrojů v Pythonu. Nejpodstatnější část backendu představuje bash skript “pipeline.sh”, který řídí celý proces zpracování vstupních dat pomocí bioinformatických nástrojů Barrnap [104],

Cd-hit [100], FastANI [111], BLAT [117] a BLAST [92]. Celý nástroj je zabalen do Docker kontejneru, což zajišťuje snadné a konzistentní spouštění nástroje v různých prostředích. Díky tomu je Bacterial Explorer snadno škálovatelný a přenosný. Přes Docker kontejner probíhá také komunikace s MySQL databází pomocí rozhraní MySQLdb pro Python. Jednotlivé komponenty backendu jsou popsány v následujících podkapitolách.

5.1.1 Implementace nástroje pomocí Flask

Backend nástroje je vytvořen pomocí frameworku Flask [118]. Flask je mikroframework vytvořený v Pythonu, který umožňuje dynamické generování HTML stránek. Tento framework obsahuje pouze základní funkce, což umožňuje vývojářům přidávat funkce podle potřeby v průběhu implementace. Základem Flasku je odlehčená verze WSGI (Web Server Gateway Interface) [119] a dvě knihovny - Werkzeug a Jinja. Pro lokální vývoj a simulaci HTTP požadavků poskytuje Flask vláknový WSGI server. Kromě backendu lze Flask použít také pro frontend. [120]

Flask neobsahuje žádnou abstrakční vrstvu pro databázi, ani žádné druhy validace či zabezpečení, což umožňuje implementátorovi plnou flexibilitu při doplňování požadavků. [120] Proces generování HTML stránky pomocí frameworku Flask, je znázorněn na obrázku 5.2.



Obr. 5.2: Znázornění procesu generování HTML stránky pomocí frameworku Flask. Uživatel zadá do vyhledávače HTML adresu, čímž pošle požadavek na server. Server zavolá Flask, který zpracuje požadavek, připojí se k databázi a na základě dat vygeneruje novou HTML stránku. Hotová stránka je pak zaslána klientovi, kterému se zobrazí statická HTML stránka. Obrázek byl převzat a modifikován z [121].

5.1.2 Nasazení nástroje pomocí Docker kontejneru

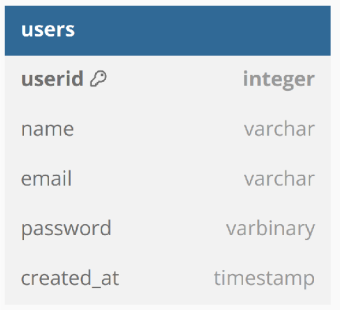
Pro snadnější nasazení na server byl Bacterial Explorer zabalen do Docker kontejneru. Docker [122] je virtualizační nástroj, který usnadňuje vytváření, nasazování a spouštění aplikací. Kontejnerizace je provedena prostřednictvím Dockerfiles, což jsou skripty obsahující veškeré informace potřebné k vyvtvoření a spuštění kontejneru. Během procesu kontejnerizace jsou do kontejneru nainstalovány a překopírovány všechny potřebné knihovny, nástroje a soubory, které aplikace vyžaduje. [123]

Oproti jiným formám virtualizace je Docker kontejner rychlejší a kompaktnější, jelikož obsahuje pouze data, která aplikace skutečně potřebuje. Kontejnery využívají jako hostitelské prostředí již běžící operační systém. Pouze se spouštějí v prostorech, které jsou izolovány od sebe navzájem a od určitých oblastí hostitelského operačního systému. Díky tomu není potřeba spouštět a vypínat celý operační systém, ale stačí pouze ukončit procesy běžící uvnitř kontejneru. [122]

5.1.3 Databáze uživatelů

Než se uživatel dostane k samotnému nástroji Bacterial Explorer, musí se zaregistrovat a následně přihlásit. Implementace databáze uživatelů byla provedena na základě bezpečnostních směrnic IT oddělení VÚVeLu, aby bylo jasné, kdo aplikaci používá.

Databáze uživatelů Bacterial Exploreru je realizována prostřednictvím databáze MySQL [124]. Údaje o uživatelích (jméno, e-mail, heslo) jsou uloženy v tabulce "Users". Každému novému uživateli je taky automaticky přiřazeno identifikační číslo a čas, kdy byla registrace provedena. Schéma tabulky "Users" je znázorněno na obrázku 5.3.



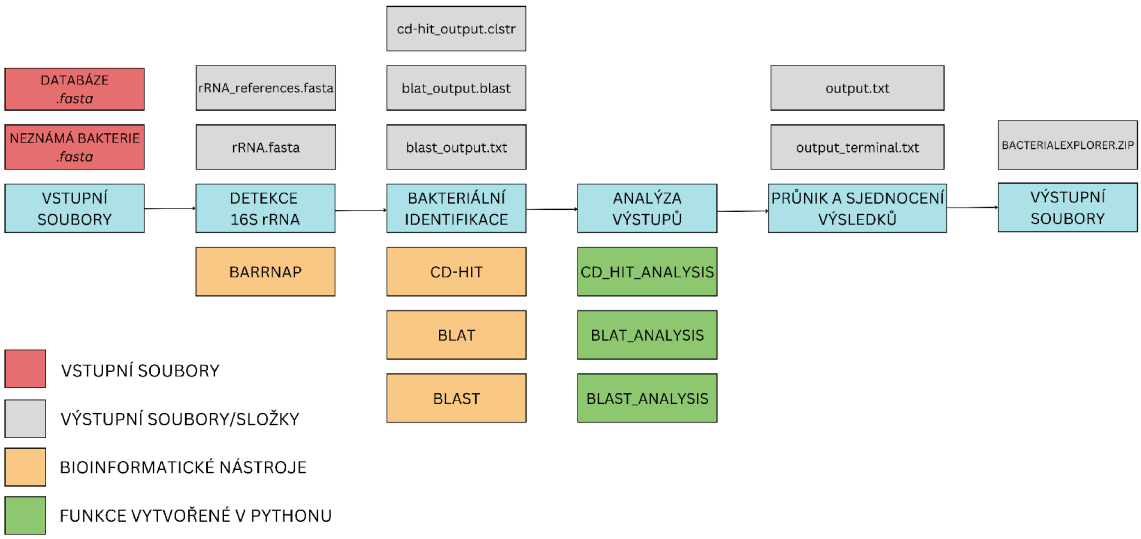
Obr. 5.3: Schéma tabulky "Users" z databáze nástroje Bacterial Explorer.

5.1.4 Pipeline Bacterial Exploreru

Jádro nástroje Bacterial Explorer představuje bash skript, který je pojmenovaný "pipeline.sh" a dostupný na `https://github.com/JulNejez/Bacterial-Explorer`. V rámci tohoto bash skriptu jsou implementovány dva bioinformatické postupy pro odhalování nových bakterií. Obě metody jsou podrobněji popsány níže.

Metoda 1

První metoda je založena na porovnání 16S rRNA. Schéma metody 1 je znázorněno na obrázku 5.4.



Obr. 5.4: Schéma Bacterial Exploreru při volbě metody založené na 166S rRNA.

Vstupem této metody je *fasta* soubor s neznámou bakterií a databáze referečních bakterií. Referenční bakterie jsou uloženy ve *fasta* formátu, přičemž jeden *fasta* soubor představuje právě jednu bakterii. Název *fasta* souboru s referenční bakterií musí obsahovat stejné označení, jako obsahuje hlavička *fasta* souboru. Například je-li hlavička *fasta* souboru ">ET39_scaffold_1", pak je název souboru "Absiella_dolichum_ET39.fasta"

Tuto metodu lze rozdělit do čtyř hlavních fází. První fází je detekce 16S rRNA jak v neznámé bakterii, tak i v referenčních bakteriích. Ve druhé fázi je provedena bakteriální identifikace pomocí bioinformatických nástrojů. Následně jsou analyzovány výstupní soubory z těchto nástrojů a na závěr je proveden průnik a sjednocení výsledků.

Detekce 16S rRNA je provedena pomocí nástroje Barrnap [104]. Podrobné nastavení jednotlivých parametrů Barrnapu je popsáno v tabulce 5.1. Pro Barrnap jsou klíčové parametry "kingdom" "evalue" a "lencutoff". Parametr "kingdom" definuje

skupinu organismů, která bude podrobena analýze. V tomto případě je skupina nastavena na bakterie. Parametr "evalue" určuje minimální e-hodotu, kterou musí nalezené shody překročit, aby byly považovány za statisticky významné. Parametr "reject" určuje prah pro zamítnutí rRNA (tzn. shody rRNA, nižší než stanovená hodnota, nebudou zahrnuty do výsledků).

Tab. 5.1: Nastavení parametrů nástroje Barrnap.

Název parametru	**Popis parametru**	**Defaultní nastavení**
kingdom	skupina organismů, která bude analyzována	bac (=bakterie)
evalue	limit pro e-hodnotu	$1e^{-6}$
lencutoff	poměr minimální délky predikované délky rRNA ku standartní délce rRNA	0.8 (=80%)
reject	hranice pro zamítnutí rRNA	0.25 (=25%)
outseq	název výstupního souboru, kam budou uloženy predikované sekvence	rRNA_all.fasta/rRNA_references_all.fasta
threads	počet použitých jader	6

Výstupem Barrnapu jsou dva soubory, jeden soubor s predikovanou rRNA pro neznámou bakterii a druhý soubor s predikovanou rRNA pro referenční bakterie. Jelikož soubory neobsahují pouze 16S rRNA, ale veškerou rRNA, je pomocí příkazu v bash skriptu provedena filtrace výstupních souborů a do nových souborů je uložena pouze 16S rRNA. Soubory s 16S rRNA jsou pak vstupními soubory do nástrojů Cd-hit [100], BLAT [117] a BLAST [92].

Pomocí nástrojů Cd-hit, BLAST a BLAT je pak porovnána 16S rRNA neznámé bakterie s 16S rRNA referenčních bakterií. Porovnání je provedeno jedním z těchto nástrojů nebo libovolnou kombinací těchto nástrojů podle toho, co si uživatel zvolí. V následujících tabulkách 5.2, 5.3, 5.4 jsou uvedeny nastavené parametry pro nástroje Cd-hit, BLAT a BLAST.

U nástroje Cd-hit je klíčový parametr "c", který určuje prahovou hodnotu pro seskupení sekvencí na základě jejich podobnosti. Tato hodnota je nastavena podle toho, jakou prahovou hodnotu uživatel zvolí. Dalším důležitým parametrem je parametr "M" který stanovuje maximální využitou paměť nástroje Cd-hit. Tento parametr byl nastaven na hodnotu 0, aby paměť nástroje nebyla nijak omezena. Výstupem Cd-hitu jsou klastry s podobnými bakteriemi.

Před spuštěním nástroje BLAT je vytvořena indexovaná databáze, která následně urychluje proces BLATu. Indexovaná databáze je vytvořená již pomocí nástroje BLAST příkazem *makeblastdb*. Tato databáze je následně pomocí příkazu *blastdbcmd* konvertována do formátu *.fa*. Takto konvertovaná databáze je vstupem do nástroje BLAT. Důležitými parametry BLATu jsou parametry "out" a "minScore" . Pomocí parametru "out" se nastavuje formát výstupního souboru, který byl v Bacterial Exploreru nastaven na formát blast8 (viz obrázek 5.5). Parametr

“minMatch” stanovuje minimální délku shody neznámé bakterie s referenční bakterií. Minimální shoda byla nastaveno na hodnotu 50. Výstupem BLATU jsou pak všechna možná zarovnání neznámé bakterie s referenčními bakteriemi, která mají délku alespoň 50 nukleotidů.

Tab. 5.2: Nastavení parametrů nástroje Cd-hit pro metodu 1.

Název parametru	**Popis parametru**	**Defaultní nastavení**
i	název vstupního souboru (neznámého genomu)	rRNA.fasta
i2	název vstupního multifasta souboru s referenčními genomy	$\text{rRNA}_{references.fasta}$
o	název výstupního souboru	Cd-hit_output
T	počet použitých jader	6
M	maximální paměť, kterou může Cd-hit použít (v Mb)	0 (=paměť není nijak omezena)
c	prahová hodnota pro seskupení sekvencí na základě jejich podobnosti	$threshold

Tab. 5.3: Nastavení parametrů nástroje BLAT pro metodu 1.

Název parametru	**Popis parametru**	**Defaultní nastavení**
database	název referenční databáze	converted_database.fa
query	název vstupního souboru	rRNA.fasta
output	název výstupního souboru	output.blast
out	formát výstupu	blast8
minMatch	minimální délka shody	50

```
An90_scaffold_1 An78_scaffold_10        100.00  603750  0       0       1       603750  3936    607685  0.0e+00 1190712.0
An90_scaffold_1 ET44_scaffold_10        99.97   3965    1       0       1       3965    32552   36516   0.0e+00 7742.0
An90_scaffold_1 An22_scaffold_18        100.00  3872    0       0       94      3965    39870   43741   0.0e+00 7567.0
An90_scaffold_1 An161_scaffold_1        99.97   3872    1       0       94      3965    150889  154760  0.0e+00 7564.0
An90_scaffold_2 An90_scaffold_2 100.00  484750  0       0       1       484750  1       484750  0.0e+00 957458.0
An90_scaffold_3 An90_scaffold_3 100.00  480500  0       0       1       480500  1       480500  0.0e+00 946529.0
An90_scaffold_3 An189_scaffold_7        99.98   25362   4       0       306389  331750  56253   81614   0.0e+00 49547.0
An90_scaffold_3 An161_scaffold_1        99.98   21500   4       0       312280  333779  125276  103777  0.0e+00 42086.0
```

Obr. 5.5: Ukázka výstupního formátu blast8. V prvním a ve druhém sloupci je název dotazované a referenční sekvence. Ve třetím sloupci je procentuelní identita porovnávaných úseků. Ve čtvrtém sloupci je délka zarovnání. V pátém a šestém sloupci je počet neshodných znaků a počet inzercí a delecí. V sedmém a osmém sloupci je počáteční a konečná pozice zarovnání v dotazované sekvenci. V devátém a desátém sloupci je počáteční a konečná pozice zarovnání v referenční sekvenci. Jedenáctý sloupec vyjadřuje e-hodnotu a poslední sloupec bit skóre.

Nejdůležitějším parametrem nástroje BLAST je parametr “perc_identity”, který určuje minimální procentuální shodu mezi úseky neznámé a referenční bakterie. Tento parametr je úměrný prahu, který zvolí uživatel. Parametr “outfm” určuje formát výstupního souboru. Výstupní soubor BLASTu je v Bacterial Exploreru nastaven na formát “6 qseqid sseqid pident length evalue bitscore”, což znamená, že výstupní soubor obsahuje název dotazované neznámé sekvence, název podobné referenční sekvence, procentuální identitu mezi dotazovanou sekvencí a referenční sekvencí, délku zarovnání, e-hodnotu a bitskóre, které kvantifikuje kvalitu zarovnání mezi dotazovanou a referenční sekvencí.

Tab. 5.4: Nastavení parametrů nástroje BLAST pro metodu 1.

Název parametru	**Popis parametru**	**Defaultní nastavení**
query	název souboru s neznámým genomem	rRNA.fasta
db	název databáze, která bude použita k hledání homologií	blast_database
out	cesta a název výstupního souboru	$working_directory/blast.txt
outfmt	formát výstupníhou souboru (viz obrázek 5.6)	6 qseqid sseqid pident length evalue bitscore
num_threads	počet použitých jader	6
perc_identity	procentuální shoda	$threshold

```
An175_scaffold_1        An174_scaffold_105      100.000 46      1.45e-12        86.1
An175_scaffold_1        An174_scaffold_105      100.000 46      1.45e-12        86.1
An175_scaffold_1        An174_scaffold_105      91.667  48      5.27e-07        67.6
An175_scaffold_1        An174_scaffold_105      91.667  48      5.27e-07        67.6
An175_scaffold_1        An174_scaffold_105      100.000 30      0.001           56.5
```

Obr. 5.6: Ukázka výstupního souboru ve formátu “6 qseqid sseqid pident length evalue bitscore” z nástroje BLAST. V prvním sloupci je název dotazované neznámé sekvence, ve druhém sloupci je název nalezené podobné referenční sekvence, ve třetím sloupci je procentuální shoda mezi dotazovanou sekvecní a nalezenou podobnou referenční sekvencí, ve čtvrtém sloupci je délka zarovnání, v pátém sloupci je e-hodnota a v šestém sloupci je bitskóre.

Po proběhnutí Cd-hitu, BLATu a BLASTu je provedena analýza výstupních souborů. Jelikož se formáty výstupních souborů jednotlivých nástrojů liší, jsou v Pythonu vytvořeny tři různé funkce pro analýzu výstupů Cd-hitu, BLATu a BLASTu. Přestože jsou vytvořeny tři různé funkce pro analýzu výstupních souborů, princip všech funkcí je stejný. Cílem je projít výstupní soubor z daného nástroje a uložit identifikátory bakterií, které byly identifikovány jako podobné. Následně jsou v databázi dohledány celé názvy bakterií, které jsou uloženy do nového souboru, který

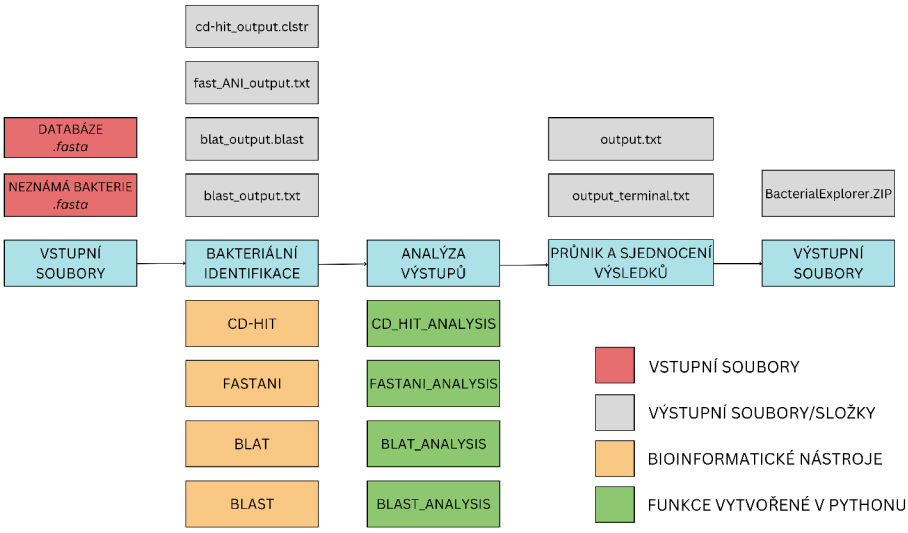
je výstupem funkce.

Poslední fáze nástroje Bacterial Explorer se liší podle toho, zda uživatel zvolil jeden nebo více nástrojů pro bakteriální identifikaci. Pokud uživatel zvolil pouze jeden nástroj, pak je výstup Bacterial Exploreru stejný jako výstup z Python funkce pro analýzu výstupů. Pokud uživatel zvolí více nástrojů, pak jsou vytvořeny dva výstupní soubory. První soubor představuje průnik výsledků všech nástrojů a vypisuje se do výstupního okna nástroje. Druhý soubor pak představuje sjednocení výsledků všech nástrojů a je součástí výstupní složky aplikace Bacterial Explorer. Průnik i sjednocení je provedeno pomocí základních příkazů v bash skriptu.

Na závěr jsou do složky "BacterialExplorer.zip" uloženy přímo výstupy z nástrojů Cd-hit, BLAT a BLAST (podle toho, jaké nástroje byly uživatelem zvoleny) a soubor "output.txt", který obsahuje sjednocení výsledků. Do složky je taky uložen univerzální soubor "README.txt", kde jsou popsány výstupní soubory Bacterial Exploreru. Celá tato složka je zazipována a při dokončení běhu Bacterial Exploreru je dostupná ke stažení.

Metoda 2

Oproti metodě 1, metoda 2 nepracuje pouze s 16S rRNA, ale s celými genomy. Schéma metody 2 je znázorněno na obrázku 5.7.



Obr. 5.7: Schéma Bacterial Exploreru v případě, že je zvolena metoda 2, která jako vstup používá celý genom.

Stejně jako u metody 1 je vstupem *fasta* soubor s neznámou bakterií a databáze referenčích bakterií, kde jeden *fasta* soubor představuje jednu bakterii. Oproti metodě 1 se metoda 2 skládá pouze ze tří fází, jelikož je přeskočena detekce 16S rRNA.

V první fázi dochází k porovnání neznámé bakterie s referenčními bakteriemi pomocí nástrojů Cd-hit, FastANI, BLAT a BLAST. Ve druhé fázi je opět provedena analýza výstupních souborů jednotlivých nástrojů. V poslední fázi je provedeno sjednocení a průnik výsledků, čímž je získán finální výsledek Bacterial Exploreru.

Nastavení Cd-hitu je stejné jako v metodě 1. Jediným rozdílem je, že do Cd-hitu nevstupují 16S rRNA, ale původní vstupní soubory s celými bakteriálními genomy. Výstupem jsou, stejně jako v metodě 1, kontigy (angl. contigs), které obsahují podobné bakterie.

U metody 2 je navíc oproti metodě 1 použit nástroj FastANI. FastANI nebylo použito u metody 1, jelikož neumožňuje zpracování 16S rRNA. Podrobné nastavení FastANI v Bacterial Exploreru je popsáno v tabulce 5.5.

U nástroje FastANI není nastavený žádný specifický parametr. Pouze jsou nastaveny cesty k souboru s neznámou bakterií a k souborům s referenčními bakteriemi, název výstupního souboru a počet použitých jader. Filtrace výstupu na základě minimální procentuální shody je provedena pomocí funkce vytvořené v Pythonu 3.

Tab. 5.5: Nastavení parametrů nástroje FastANI.

Název parametru	**Popis parametru**	**Defaultní nastavení**
q	cesta k souboru s neznámou sekvencí	$input_file
rl	cesta k seznamu referenčních genomů	$reference_path/genomes_list.txt
o	cesta a název výstupního souboru, kam budou uloženy výsledky	$working_directory/fast_ANI_results.txt
t	počet použitých jader	6

Nástroj BLAT je rovněž nastaven stejně jako v metodě 1, akorát parametr ”minMatch” je nastaven na hodnotu 300. Na rozdíl od metody 1 není vstupem nástroje 16S rRNA, ale celý genom. Stejně jako v metodě 1 je vytvořena indexovaná referenční databáze, která je následně konvertována. Všechny nastavené parametry jsou uvedeny v tabulce 5.6.

Tab. 5.6: Nastavení parametrů nástroje BLAT pro metodu 2.

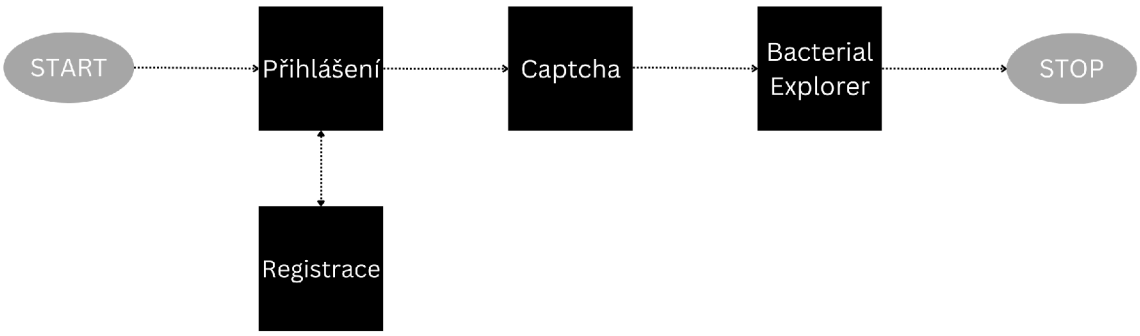
Název parametru	**Popis parametru**	**Defaultní nastavení**
database	název referenční databáze	converted_database.fa
query	název vstupního souboru	$input_file
output	název výstupního souboru	output.blast
out	formát výstupu	blast8
minMatch	minimální délka shody	300

Nástroj BLAST je nastaven úplně stejně jako v metodě 1. Rozdílem jsou jen vstupní soubory, kdy místo 16S rRNA vstupují do BLASTu celé genomy. Výstupní formát je opět ve formátu "6 qsesid sseqid pident length evalue bitscore".

Analýza výstupních souborů jednotlivých nástrojů je stejná jako v metodě 1. Pro analýzu výstupních souborů jsou použity čtyři Python funkce, jejichž cílem je vypsat názvy referenčních bakterií, které jsou dle zadaných požadavků podobné neznámé bakterii. Poslední fáze je rovněž stejná jako v metodě 1, kdy jsou na závěr vytvořeny 2 výstupní soubory, kdy jeden výstupní soubor představuje průnik výsledků jednotlivých nástrojů a druhý soubor představuje sjednocení výsledků. Průnik výsledků je vypsán ve výstupním okně nástroje Bacterial Explorer, soubor se sjednocenými výsledky je uložen do výstupní složky, kterou si po doběhnutí aplikace může uživatel stáhnout.

5.2 Frontend aplikace

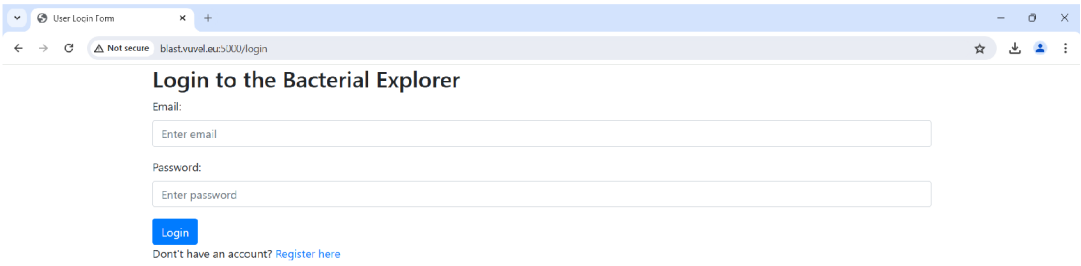
Uživatelské rozhraní je ve Flasku vytvářeno pomocí šablon a statických souborů. Šablony představují HTML soubory, ve kterých je definovaná struktura a obsah uživatelského rozhraní. Statické soubory pak představují CSS styly, JavaScript soubory nebo obrázky. Uživatelské rozhraní nástroje Bacterial Explorer zahrnuje celkem 4 šablony (viz obrázek 5.8), které představují jednotlivé HTML soubory, které se uživateli postupně načítají po spuštění nástroje.



Obr. 5.8: Propojení jednotlivých html šablon, které tvoří uživatelské rozhraní Bacterial Exploreru.

Po spuštění nástroje se uživateli nejprve objeví okno s přihlášením do Bacterial Exploreru (viz obrázek 5.9). Tato šablona byla implementována pomocí externího CSS souboru *bootstrap.min.css* verze 4.6.1 knihovny Bootstrap. Pokud je uživatel zaregistrován a správně vyplní pole "E-mail" a "Password", tak je po stisknutí tlačítka "Login" přesměrován na stránku "captcha", která ověří, že uživatelem není robot. Jestliže uživatel zadá údaje špatně, pak se objeví errorová hláška, která ho

informuje o nesprávnosti zadaných údajů. Po kliknutí na text “Register here ” je uživatel přesměrována na stránku s registrací.



Obr. 5.9: Přihlášení uživatele do nástroje Bacterial Explorer.

Stránka s registrací (viz obrázek 5.10) slouží pro registraci nových uživatelů nástroje Bacterial Explorer. Stejně jako stránka pro přihlášení byla i tato stránka implementována pomocí externího CSS souboru *bootstrap.min.css* verze 4.6.1 knihovny Bootstrap.

Obr. 5.10: Registrace uživatele do nástroje Bacterial Explorer.

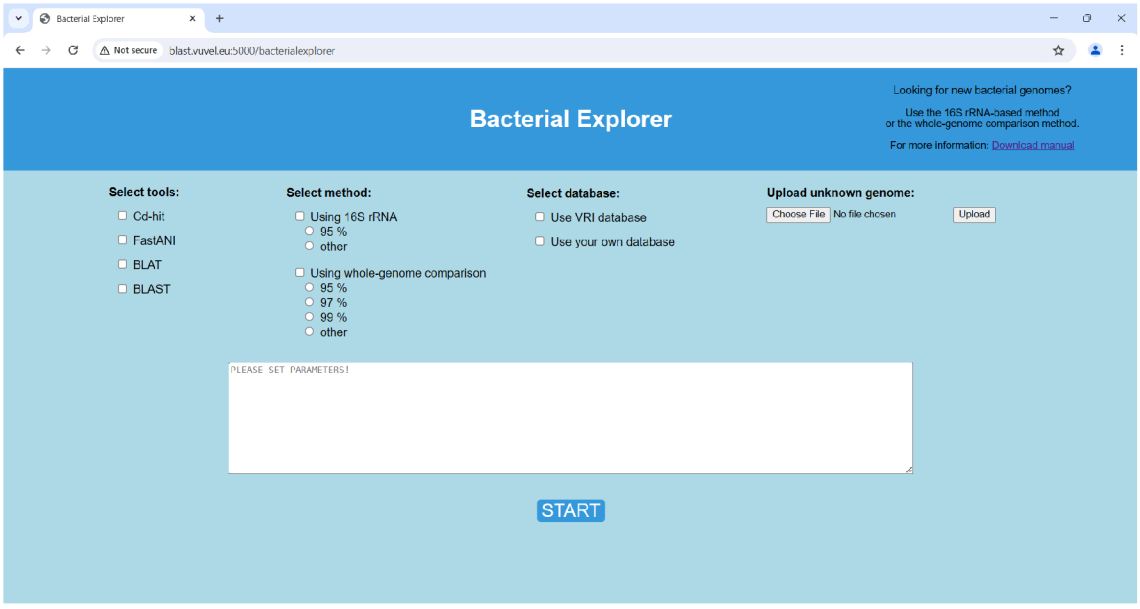
Registrační stránka obsahuje celkem tři pole k vyplnění - jméno, e-mail a heslo. Po správném vyplnění všech údajů může uživatel zmáčknout tlačítko “Register”. Tím se registruje do Bacterial Exploreru a jeho údaje jsou uloženy do databáze uživatelů nástroje Bacterial Explorer. Následně je uživatel po kliknutí na “Login here” opět přesměrován na stránku přihlášení, odkud je dále přesměrován na stránku “captcha”.

Stránka “captcha” (viz obrázek 5.11) obsahuje CAPTCHA (Completely automated public Turing test to tell computers and humans apart) [125] test. Jedná se o Turingův test, jehož účelem je rozlišení počítačů od lidí [126]. V rámci webové aplikace byla použita obrazová CAPTCHA, která spočívá v zobrazení zdeformovaného

textu, který má uživatel správně přepsat. Pokud uživatel text opíše správně, pak je přesměrován na stránku “bacterialexplorer”, jinak se generuje nová CAPTCHA. Pro generování CAPTCHA textu byla využita knihovna Pythonu captcha verze 0.5.0 [127].

Obr. 5.11: Ukázka stránky s CAPTCHou.

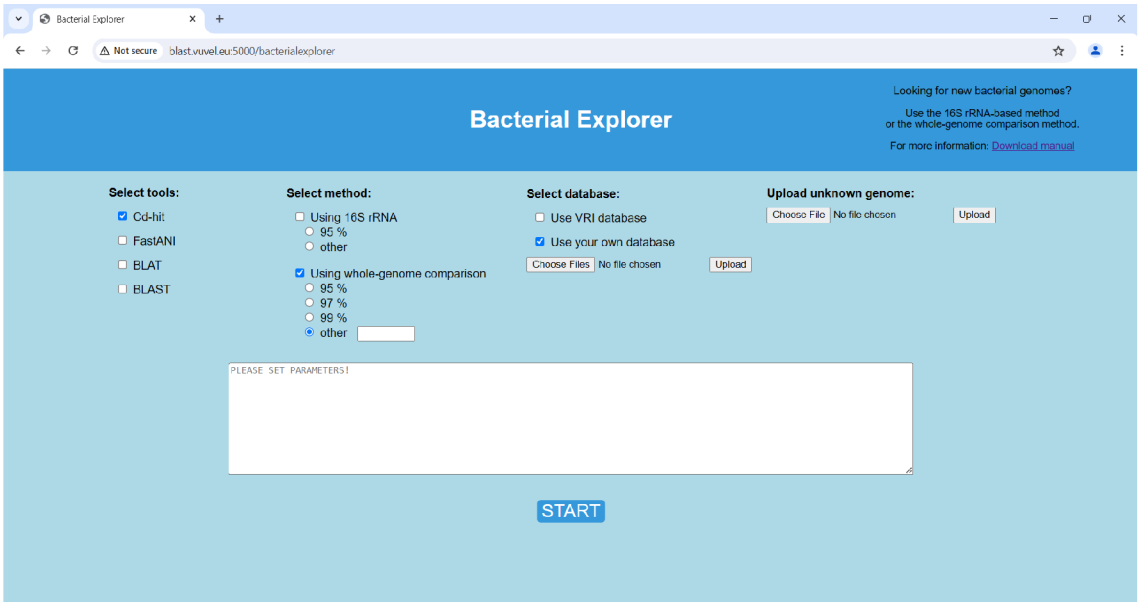
Stránka “bacterialexplorer” (viz obrázek 5.12) již představuje samotný nástroj Bacterial Explorer. Tato stránka byla implementována pomocí knihoven jQuery verze 3.6.4 [128] a SweetAlert2 verze 11 [129]. Bacterial Explorer zahrnuje čtyři hlavní sekce. V první sekci si uživatel volí nástroje, které chce použít pro bakteriální typizaci. Volba nástroje je vyřešena pomocí checkboxu, který umožňuje uživateli zvolit jeden až všechny nástroje. Pokud není uživatelem zvolen ani jeden nástroj, tak se objeví errorová hláška, která uživatele informuje o tom, že musí zvolit alespoň jeden nástroj.



Obr. 5.12: Hlavní stránka nástroje Bacterial Explorer.

Ve druhé sekci uživatel nastavuje metodu. Uživatel volí mezi metodou založenou na 16S rRNA nebo metodou založenou na celých genomech. Volba metody je rovněž vyřešena pomocí checkboxu. Pro každou metodu zvlášť pak uživatel volí práh. Pro metodu 16S rRNA může uživatel volit mezi prahem 95% nebo “other”. Pro celogenomovou metodu může uživatel volit mezi hodnotami 95%, 97%, 99% a “other”.

V tomto případě byl pro volbu prahu zvolen radiobutton, který umožňuje uživateli zvolit právě jednu hodnotu. Pokud se uživatel rozhodne pro možnost ”other”, zobrazí se uživateli pole, kam může ručně zadat hodnotu prahu (viz obrázek 5.13). Jestliže uživatel nezvolí ani jednu možnost, vypíše se errorová hláška, která uživatele informuje o tom, že musí zvolit nějaký prah.



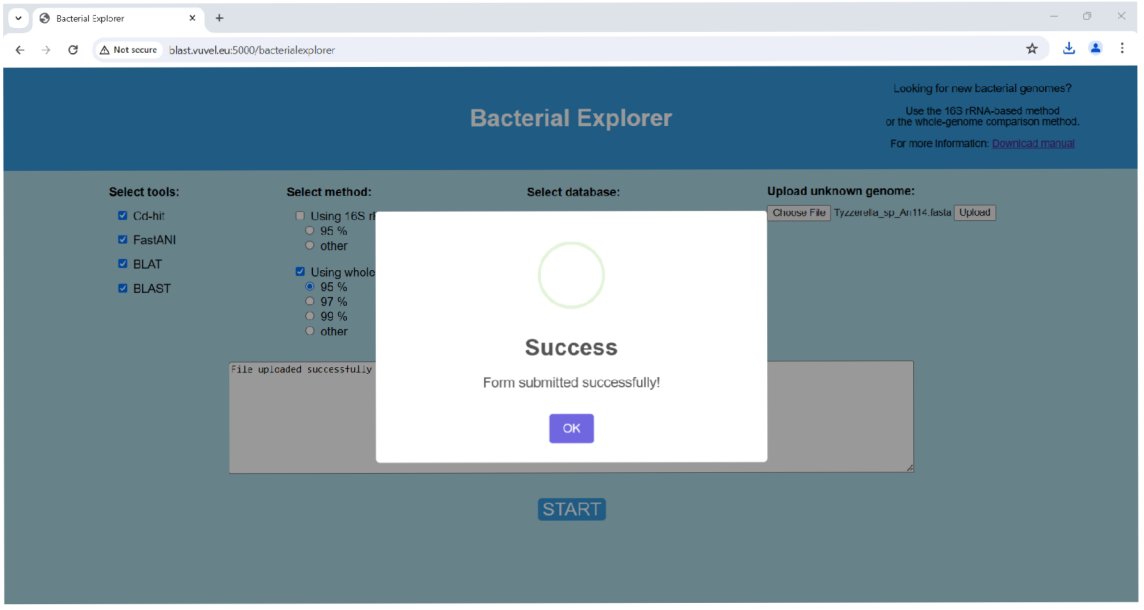
Obr. 5.13: Hlavní stránka nástroje Bacterial Explorer s ukázkou zakliknuté možnosti “other”.

Třetí sekce slouží ke zvolení databáze. Uživatel volí mezi referenční databází VÚVeLu, která je již v online prostředí propojená s nástrojem Bacterial Explorer nebo si může zvolit možnost “Use your own database”, která umožňuje použití vlastní databáze. Velikost vlastní databáze je omezena na 30 MB, což bylo konzultovány s odborníky z VÚVeLu. Velikost 30 MB je dostatečně veliká na to, aby bylo možné analyzovat bakteriální genomy, ale zároveň nedošlo k zahlcení serveru. Soubory v referenční databázi musí být ve *fasta* formátu.

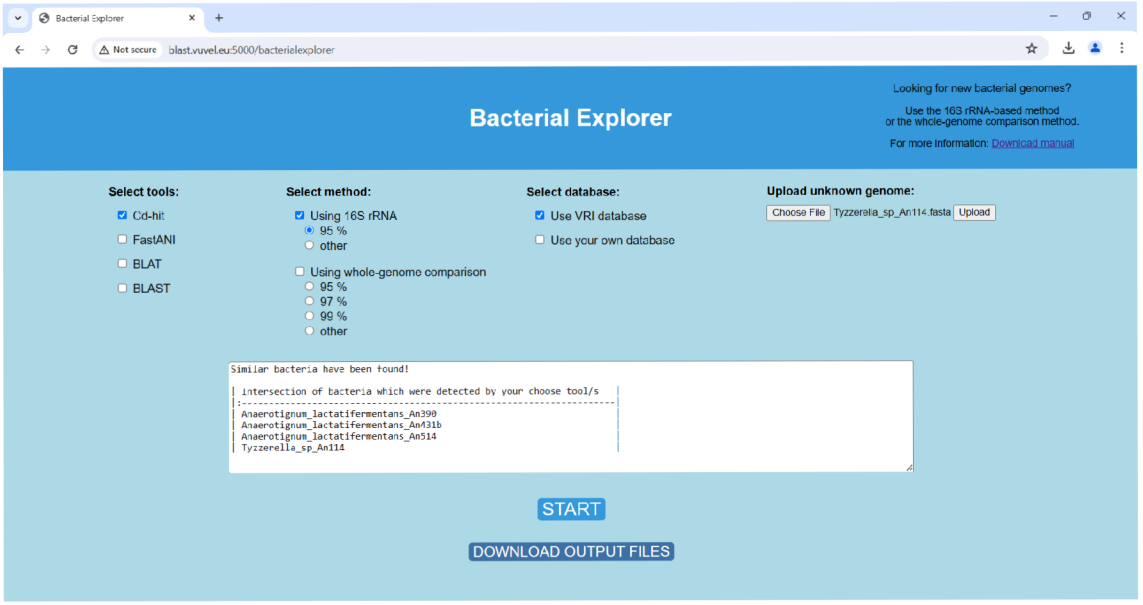
Poslední sekce aplikace Bacterial Explorer je určena k nahrání *fasta* souboru s neznámou bakterií. Pokud uživatel navolí všechny parametry a nahraje potřebné soubory, může stisknout tlačítko “START”. Pokud je vše v pořádku, zobrazí se načítací kolečko a proces Bacterial Exploreru je spuštěn.

Pokud proběhne bakteriální typizace v pořádku, zobrazí se po doběhnutí hláška, která uživatele informuje o tom, že vše proběhlo v pořádku (viz obrázek 5.14). Po doběhnutí nástroje se také zobrazí v informačním okně krátká hláška, která informuje uživatele o tom, zda byla nalezena nějaká bakterie podobná neznámé bakterii či nikoliv a pokud byly nalezeny nějaké významně podobné bakterie, vypíšou

se tyto bakterie v informačním okně. Dále se zobrazí tlačítko, které umožňuje stažení všech výstupních souborů z jednotlivých nástrojů. 5.15



Obr. 5.14: Ukázka hlášky, která se zobrazí po úspěšném doběhnutí nástroje Bacterial Explorer.



Obr. 5.15: Ukázka výstupu, který se zobrazí po doběhnutí nástroje Bacterial Explorer.

5.3 Testování nástroje

Nedílnou součástí vývoje nástroje Bacterial Explorer bylo jeho průběžné testování. V rámci této kapitoli je popsaný vývoj nástroje, ale také testování funkčnosti Bacterial Exploreru.

5.3.1 Testování nástroje během vývoje

Během vývoje nástroje vzniklo několik verzí nástroje Bacterial Explorer, které byly postupně testovány a následně upravovány podle požadavků uživatelů. Hlavními požadavky byla uživatelská přívětivost, rychlost a správná funkčnost nástroje.

Předchůdcem pro otestování hypotézy v offline modulu byl Bacterial Identifier, který byl prezentován na studentské konferenci EEICT 2024 [130] . Bacterial Identifier původně obsahoval pouze nástroje Cd-hit a BLAST a byl vytvořen pomocí knihovny Tkinter [131] v Pythonu3. Referenční databázi bylo potřeba vždy ručně nahrát a po doběhnutí Bacterial Identifieru byl uživatel informován pouze krátkou výstupní hláškou o tom, zda byly nalezeny nějaké významně podobné bakterie neznámé bakterii či nikoliv. Pokud byly nalezeny nějaké bakterie významně podobné neznámé vstupní bakterii, pak se do výstupního souboru uložily názvy podobných bakterií.

Při prvním testování Bacterial Identifieru byl vznesen požadavek na výpis výsledků přímo do výstupního okna nástroje, aby uživatel hned viděl, které bakterie jsou významně podobné neznámé bakterii. Automatický softwarový nástroj Bacterial Identifier byl také rozšířen o nástroje FastANI a BLAT a byl implementován do online databáze VÚVeLu. Jelikož hlavní cílovou skupinou uživatelů budou pracovníci VÚVeLu, byla jako výchozí referenční databáze nastavena databáze VÚVeLu.

Uživatelé také vznesli požadavek na možnost výběru vlastní databáze, takže byla do Bacterial Exploreru přidána možnost volby vlastní referenční databáze. Uživatel si může vybrat mezi databází VÚVeL, se kterou je nástroj propojený, nebo si může zvolit vlastní databázi, kterou je ale potřeba ručně nahrát.

Po těchto úpravách byl Bacterial Explorer otestován a zjištěné nedostatky byly opraveny. Jedním ze zjištěných nedostatků bylo špatné propojení databáze uživatelů s Docker kontejnerem Bacterial Exploreru. Tento problém byl vyřešen použitím knihovny MySQLdb místo flask_mysqldb. Dalším nedostatkem bylo špatné nastavení cest k výstupním souborům, což způsobilo, že se některé výstupní soubory neukládaly do výstupní složky Bacterial Exploreru. Cesty k souborům byly opraveny a všechny soubory se nyní ukládají do výstupní složky správně.

V současné době probíhá další testování nástroje Bacterial Explorer a již nyní jsou připravovány nové inovace. Jednou z inovací bude vytvoření jednotné indexo-

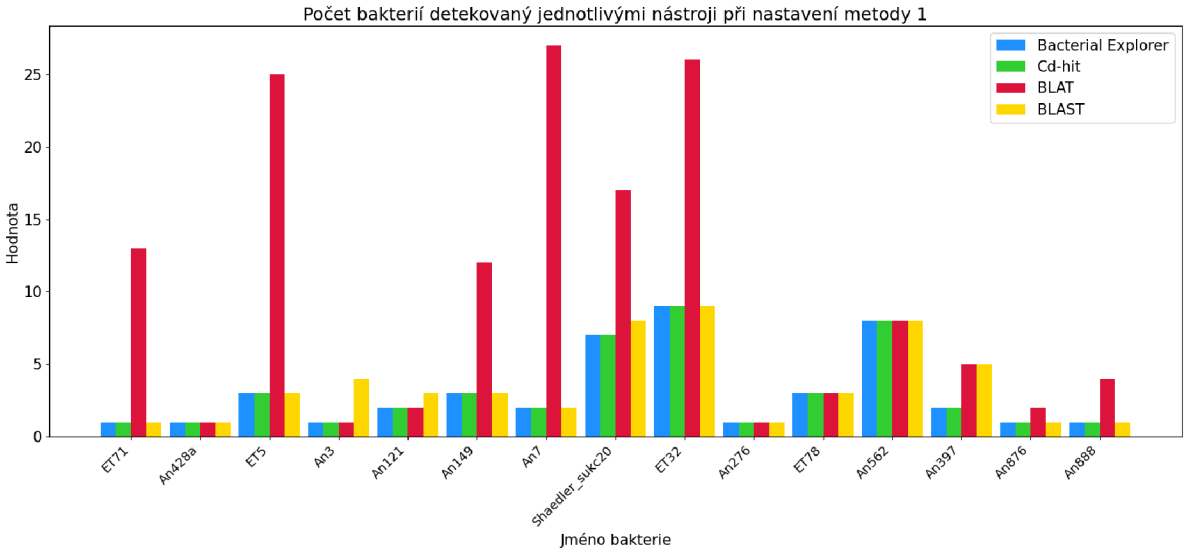
vané databáze pro všechny nástroje, čímž se urychlí proces Bacterial Exploreru. Do budoucna se pak počítá s dalším rozšířením nástroje.

5.3.2 Výsledky - Testování I

První testování se zaměřuje na kmeny, které mají v databázi VÚVeLu několik zástupců. Cílem tohoto experimentu bylo zjistit, jaké a kolik bakterií z databáze VÚVeLu jsou detekované nástrojem Bacterial Explorer a jednotlivými nástroji Cd-hit, FastANI, BLAT a BLAST. Výsledky testování jsou uvedeny v tabulkách (viz příloha A). Pro lepší přehlednost byly vytvořeny grafy 5.16, 5.17.

Z každého kmene, který obsahoval v databázi VÚVeLu více než jednoho zástupce, byly vybrány 3 bakterie, které byly testovány v nástroji Bacterial Explorer. Celkem bylo tedy testováno 15 bakterií při dvou různých nastaveních nástroje Bacterial Explorer.

Nejprve byl nástroj otestován v režimu detekce založené na 16S rRNA při nastavení prahu na 99% (viz graf 5.16) a výběru všech možných nástrojů (Cd-hit, BLAST a BLAT). Jak je patrné z grafu 5.16, Bacterial Explorer detekoval nejméně bakterií podobných referenčním bakteriím z databáze VÚVeLu, jelikož výsledkem Bacterial Exploreru (ve výstupním okně) je průnik výsledků ze všech použitých nástrojů.

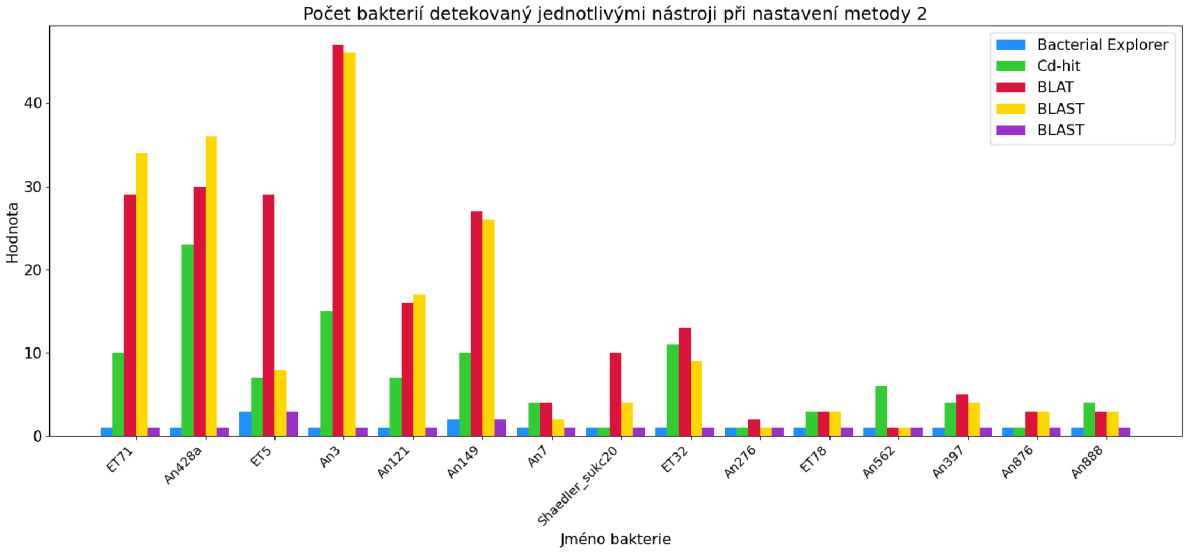


Obr. 5.16: Graf znázorňující počet bakterií predikovaný nástrojem Bacterial Explorer a nástroji Cd-hit, BLAT a BLAST. Bacterial Explorer byl nastaven na metodu 1 a práh byl nastaven na 99%.

V grafu si dále můžeme všimnout, že se u různých nástrojů liší počty detekovaných bakterií. Obecně lze říct, že BLAT a BLAST ve většině případů detekují mnohem více bakterií podobných referenční bakterii než Cd-hit. To je pravděpodobně

způsobené tím, že tyto nástroje "rozsekají" vstupní a referenční sekvenci na kratší úseky a následně jsou tyto úseky porovnávány. Je tudíž větší pravděpodobnost, že tyto nástroje najdou ke vstupní bakterii nějakou podobnou bakterii z referenční databáze s požadovanou shodou. Oproti tomu nástroj Cd-hit nalezne mezi vstupní sekvencí a referenční sekvencí nejdelší možné zarovnání a výsledná shoda mezi sekvencemi je pak určena na základě tohoto zarovnání. Z toho vyplývá, že celkový výsledek nástroje Bacterial Explorer, při volbě všech nástrojů, prahu 99% a metody 1, je nejvíce ovlivněn výsledkem Cd-hitu, jelikož detekuje nejméně podobných bakterií.

V grafu 5.17 jsou zaznamenány výsledky testování nástroje v režimu detekce založené na celých genomech. Z tohoto grafu vyplývá, že Bacterial Explorer, stejně jako v předchozí metodě, detekoval podle očekávání nejméně podobných bakterií vstupní bakterii. Z grafu můžeme pozorovat, že výsledek Bacterial Exploreru v tomto případě ovlivňuje nejvíce nástroj FastANI, který detekuje nejméně bakterií. To je způsobeno tím, že k výpočtu podobnosti používá celé sekvence, čímž se liší od ostatních nástrojů, které k výpočtu podobnosti nepoužívají celé sekvence, ale pouze kratší úseky. Ostatní nástroje mají tedy větší pravděpodobnost nalezení požadované shody. Výstupem FastANI je pak průměrná nukleotidová identita dvou sekvencí. Naopak, nejvíce podobných bakterií, opět detekují nástroje BLAT a BLAST.



Obr. 5.17: Graf znázorňující počet bakterií predikovaný nástrojem Bacterial Explorer a nástroji Cd-hit, BLAT, BLAST a FastANI. Bacterial Explorer byl nastavem na metodu 2 a práh byl nastaven na 99%.

U obou metod byly v grafech (viz příloha B) barevně znázorněny kmeny, do kterých spadají detekované bakterie. V drtivé většině případů spadají detekované bakterie do stejného kmene jako vstupní bakterie. Pouze ve třech případech spadala

detekovaná bakterie do jiného kmene než vstupní bakterie. Konkrétně u bakterie *Eubacterium* sp An3 z kmene *Firmicutes* byla nástrojem BLAST detekována bakterie *Gordonibacter* sp An232A z kmene *Actinobacteria* pro metodu 16S rRNA.

U metody založené na celogenomovém porovnávání byla pro bakterii *Collinsella tanakaei* ET32 z rodu *Actinobacteria* detekovaná nástrojem Cd-hit jako podobná bakterie *Elusimicrobium* sp An237 z rodu *Elusimicrobiota*. U bakterie *Fusobacterium perfoetens* An888 z kmene *Fusobacteria* byla identifikovaná jako podobná bakterie *Enterocloster clostridioformis* ET46 z kmene *Firmicutes*. Jelikož byly bakterie z jiného kmene detekovány vždy pouze jedním nástrojem, je pravděpodobné, že se jedná o odchylky nástrojů.

Navzdory očekávání, že metoda založená na 16S rRNA detekuje více podobných bakteriích, se toto očekávání zcela nenaplnilo, jelikož u nástrojů BLAT a BLAST bylo zaznamenáno větší množství podobných bakterií u metody 2, založené na porovnávání celých genomů. To je pravděpodobně způsobeno tím, že tyto nástroje sice využívají celé genomy, ale výsledná podobnost mezi neznámou bakterií a referenčními bakteriemi není vztažena na celé genomy, jelikož jsou genomy rozsekány a následně jsou porovnávány jednotlivé úseky. Aby bylo dosaženo větší přesnosti, bylo by možné nastavit přísnější požadavky na minimální délku porovnávaných úseků. Zde ale potom vyvstává otázka, jak minimální délku nastavit, aby nedošlo k nadměrné filtraci výstupních bakterií. A to je důvod, proč Bacterial Explorer kombinuje více nástrojů, jelikož umožňuje uživateli vzájemné srovnání výsledků jednotlivých nástrojů a ve výstupním souboru "output.txt" je souhrný přehled bakterií, ketré byly detekovány jednotlivými nástroji. Uživatel pak s touto informací může dále naložit podle svých potřeb a může například zkusit globální zarovnání pro jednotlivé bakterie, které umožňuje stanovení celkové podobnosti mezi bakteriemi.

V rámci tohoto experimentu byla také porovnána doba běhu Bacterial Exploreru pro metodu 1 a metodu 2. Hlavní výhodou metody 1 je její rychlost. Průměrná doba běhu Bacterial Exploreru pro metodu 1 byla 10 sekund pro referenční rRNA databázi 452 bakteriálních genomů z databáze VÚVeL. Oproti tomu metoda 2 běžela v průměru 15 minut pro 452 genomů z databáze VÚVeL.

5.3.3 Výsledky - Testování II

Druhé testování se zaměřuje na kmeny, které mají v databázi VÚVeL pouze jednoho zástupce. Konkrétně se jedná o kmeny *Verrumicrobiota*, *Elusimicrobiota* a *Synergistes*. Bacterial Explorer byl tedy testován na bakteriích - *Akkermansia muncinphila* An78, *Elusimicrobium* sp An273 a *Cloacibacillus* sp An23.

Jak ukazuje tabulka 5.7, Bacterial Explorer stanovil jako podobné bakterie pouze tytéž bakterie, které byly nahrány jako vstupní neznámé bakterie. Tento výsledek byl získán jak při volbě metody 1 (založené na 16S rRNA), tak i při volbě metody 2 Bacterial Exploreru. Zároveň i jednotlivé nástroje - Cd-hit, BLAT, BLAST a FastANI stanovily jako podobné bakterie pouze ty samé bakterie. Jelikož v databázi VÚVelu nebyly žádné jiné bakterie v těchto kmenech, naplnilo se očekávání, že žádné jiné bakterie nebudou Bacterial Explorerem ani jiným nástrojem detekovány jako podobné.

Tab. 5.7: Výsledky testování II pro nástroje Bacterial Explorer a Cd-hit. Metody 1 a 2 představují zvolené metody v nástroji Bacterial Explorer. Práh byl nastaven na 99%.

Název vstupní bakterie	**Metoda**	**Bacterial Explorer**	**Cd-hit**
Akkermansia muciniphila An78	1	*Akkermansia muciniphila* An78	*Akkermansia muciniphila* An78
Akkermansia muciniphila An78	2	*Akkermansia muciniphila* An78	*Akkermansia muciniphila* An78
Elusimicrobium sp An273	1	*Elusimicrobium* sp An273	*Elusimicrobium* sp An273
Elusimicrobium sp An273	2	*Elusimicrobium* sp An273	*Elusimicrobium* sp An273
Cloacibacillus sp An23	1	*Cloacibacillus* sp An23	*Cloacibacillus* sp An23
Cloacibacillus sp An23	2	*Cloacibacillus* sp An23	*Cloacibacillus* sp An23

Tab. 5.8: Výsledky testování II pro nástroje BLAT, BLAST a FastANI. Metody 1 a 2 představují zvolené metody v nástroji Bacterial Explorer. Práh byl nastaven na 99%.

Název vstupní bakterie	**Metoda**	**BLAT**	**BLAST**	**FastANI**
Akkermansia muciniphila An78	1	*Akkermansia muciniphila* An78	*Akkermansia muciniphila* An78	-
Akkermansia muciniphila An78	2	*Akkermansia muciniphila* An78	*Akkermansia muciniphila* An78	*Akkermansia muciniphila* An78
Elusimicrobium sp An273	1	*Elusimicrobium* sp An273	*Elusimicrobium* sp An273	-
Elusimicrobium sp An273	2	*Elusimicrobium* sp An273	*Elusimicrobium* sp An273	*Elusimicrobium* sp An273
Cloacibacillus sp An23	1	*Cloacibacillus* sp An23	*Cloacibacillus* sp An23	-
Cloacibacillus sp An23	2	*Cloacibacillus* sp An23	*Cloacibacillus* sp An23	*Cloacibacillus* sp An23

5.3.4 Výsledky - Testování III

Třetí testování se zaměřuje na bakterie, které nejsou součástí databáze VúVeLu, ale jsou volně dostupné na NCBI. Oproti testovací databázi VÚVeLu bylo zarovnání těchto genomů provedeno jinými zarovnávacími technikami. Pro testování byly vybrány jednak bakterie, které spadají do kmenů, které se nachází v referenční databázi VÚVeLu, ale také bakterie, které nespadají ani do jednoho z kmenů v databázi VÚVeLu. Testované bakterie a výsledky testování jsou uvedeny v tabulkách 5.9, 5.10.

Tab. 5.9: Výsledky testování III pro nástroje Bacterial Explorer a Cd-hit. Metody 1 a 2 představují zvolené metody v nástroji Bacterial Explorer. Ve všech experimentech byl nastaven práh 99%.

Název vstupní bakterie	**Metoda**	**Bacterial Explorer**	**Cd-hit**
Treponema pallidum	1	Žádná bakterie	Žádná bakterie
Treponema pallidum	2	Žádná bakterie	Žádná bakterie
Helicobacter pylori	1	Žádná bakterie	Žádná bakterie
Helicobacter pylori	2	Žádná bakterie	Žádná bakterie
Escherichia coli Nissle 1917	1	*Escherichia coli* An190 *Escherichia coli* An786 *Escherichia fergusonii* ET78	*Escherichia coli* An190 *Escherichia coli* An786 *Escherichia fergusonii* ET78
Escherichia coli Nissle 1917	2	Žádná bakterie	Žádná bakterie

Tab. 5.10: Výsledky testování III pro nástroje BLAT, BLAST a FastANI. Metody 1 a 2 představují zvolené metody v aplikaci Bacterial Explorer. Ve všech experimentech byl nastaven práh 99%.

Název vstupní bakterie	**Metoda**	**BLAT**	**BLAST**	**FastANI**
Treponema pallidum	1	Žádná bakterie	Žádná bakterie	-
Treponema pallidum	2	Žádná bakterie	Žádná bakterie	Žádná bakterie
Helicobacter pylori	1	Žádná bakterie	Bacillus aerophilus ET127	-
Helicobacter pylori	2	Žádná bakterie	Žádná bakterie	Žádná bakterie
Escherichia coli Nissle 1917	1	*Escherichia coli* An190 *Escherichia coli* An786 *Escherichia fergusonii* ET78	*Escherichia coli* An190 *Escherichia coli* An786 *Escherichia fergusonii* ET78	-
Escherichia coli Nissle 1917	2	*Escherichia coli* An190 Escherichia coli An786 *Escherichia fergusonii* ET78	*Escherichia coli* An190	Žádná bakterie

První testovací bakterií byla zvolena *Treponema pallidum* (PRJNA378185 [132]) z kmene *Spirochaete*. Tato bakterie je známá, ale v testovací databázi VÚVeLu chybí, jelikož není předmětem zájmu VÚVeLu. Proto byl u této bakterie předpoklad, že ji Bacterial Explorer s referenční databází VÚVeLu detekuje jako neznámou. Toto očekávání se naplnilo a Bacterial Explorer, ani žádný z nástrojů Cd-hit, FastANI, BLAT a BLAST, neidentifikoval žádnou bakterii z databáze VÚVeLu jako podobnou.

Druhou bakterií byla zvolena bakterie *Helicobacter pylori* (PRJNA715181 [133]) z kmene *Proteobacteria*. Tato bakterie také patří mezi známé bakterie, ale v databázi VÚVeLu chybí. Bacterial Explorer tuto bakterii identifikoval rovněž jako neznámou, ale při metodě 1 byla detekována nástrojem BLAST bakterie *Bacillus aerophilus* ET127 z kmene *Firmicutes* jako podobná.

Poslední testovanou bakterií byla *Escherichia coli Nissle* 1917 (PRJNA248167 [134]). Jedná se o nepatogení bakterii, která se vyskytuje v mikrobiomu a často bývá sou-

částí probiotik. Tato bakterie rovněž není v databázi VÚVeL, ale protože se v databázi nachází jiné druhy *Escherichia coli*, bylo pravděpodobné, že Bacterial Explorer detekuje nějaké bakterie z databáze jako podobné. Toto očekávání se naplnilo, jelikož při metodě 1 byly detekována bakterie *Escherichia coli* An190, *Escherichia coli* An786 a *Escherichia coli* ET78 jako podobné. V případě metody 2 ale nebyla žádná Bakterie detekována jako podobná, jelikož nástroj Cd-hit nedetekoval žádnou bakterii jako podobnou.

Závěr

Diplomová práce se zabývá problematikou bakteriální typizace. Bakteriální typizace hraje významnou roli při hledání nových bakterií, které jsou důležité například pro vytvoření nových probiotik.

První kapitola diplomové práce je věnována bakteriální typizaci. Jsou zde popsány jednak fenotypové metody, ale zejména genotypové metody. Konkrétně se kapitola zaměřuje na pulzní gelovou elektroforézu, polymorfismus délky restrikčních fragmentů, multilokusovou sekvenační typizaci a bakteriální typizaci analýzou jednonukleotidových polymorfismů.

Druhá kapitola se zaměřuje na metody analýzy genomických dat. Je zde popsán hladový algoritmus, heuristické vyhledávání podobných sekvencí, skrytý Markovův model, algoritmus mapování bez zarovnání a algoritmus minMLST.

Třetí kapitola popisuje dataset bakteriálních genomů poskytnutý VÚVeL. Čtvrtá kapitola je věnována dostupným nástrojům pro bakteriální typizaci. Konkrétně se kapitola zaměřuje na nástroje BLAST, BLAT Cd-hit, Barrnap, FastANI, BLAT a PROKKA.

Následná praktická část je věnována vlastní implementaci a testování nástroje Bacterial Explorer. V rámci této kapitoli je popsán backend nástroje, frontend nástroje a testování nástroje. Testování nástroje rovněž zahrnuje diskusi výsledků.

Nástroj Bacterial Explorer byl vytvořen pomocí frameworku Flask pro Python a byl implementován do online databáze VÚVeLu pomocí Docker kontejneru. Bacterial Explorer slouží k detekci nových bakterií, které nejsou zahrnuty v referenční databázi a nabízí dvě metody, jak detekovat podobné bakterie z databáze. První metoda (metoda 1) je založená na 16S rRNA, druhá metoda (metoda 2) pro detekci podobných bakterií používá celé genomy. K detekci podobných bakterií používá nástroj Bacterial Explorer nástroje Cd-hit, BLAT, BLAST a FastANI. V případě metody 1 je navíc ještě použit nástroj Barrnap k detekci 16S rRNA. Uživatelské prostředí nástroje bylo rovněž vytvořeno pomocí frameworku Flask pro Python.

V rámci testování byla jednak testována uživatelské přívětivost, ale také funkčnost nástroje. V průběhu vývoje nástroje byl Bacterial Explorer upravován dle požadavků VÚVeLu. V rámci testování funkčnosti byla provedena tři testování. Testování I se zabývalo testováním nástroje pro bakterie z databáze VŮVeLu, které spadají do kmenů, které jsou v databázi hojně zastoupeny. Testování odhalilo, že všechny testované bakterie byly správně detekovány jako nejvíce podobné bakterie z referenční databáze. Dále bylo zjištěno, že jako podobné bakterie se identifikují bakterie ze stejného kmene jako je vstupní bakterie. Pouze ve výjimečných případech jsou bakterie z jiných kmenů detekovány jako podobné. To je pravděpodobně způsobeno odchylkou použitých nástrojů.

Testování II testovalo Bacterial Explorer na bakteriích, které pochází z kmenů, které mají v databázi VÚVeLu pouze jednoho zástupce. Podle očekávání testování ukázalo, že se jako podobná bakterie z databáze detekovala pouze ona testovaná bakterie. Testování III bylo zaměřeno na testování nástroje na bakteriích, které nebyly v databázi. K testování byly zvoleny známé a dobře prozkoumané bakterie. Cílem experimentu bylo dokázat, zda se tyto bakterie identifikují jako neznámé. V případě *Treponemy pallidum* a *Helicobacter pylori* se tyto baterie identifikovaly jako neznámé, ale v případě bakterie *Escherichia coli Nissle* 1917 Bacterial Explorer detekoval podobné bakterie, jelikož se v databázi VÚVeLu vyskytovaly jiné druhy *Escherichia coli*.

Celkově z výsledků testování vyplývá, že Bacterial Explorer poskytuje poměrně rychlou možnost, jak detekovat neznámé bakterie. Zároveň Bacterial Explorer uživateli poskytuje výsledky ze všech nástrojů, které nástroj používá, což uživateli nabízí další možnosti.

Literatura

[1] FONTANA, Luis; BERMUDEZ-BRITO, Miriam; PLAZA-DIAZ, Julio; MUÑOZ-QUEZADA, Sergio a GIL, Angel, 2013. Sources, isolation, characterisation and evaluation of probiotics. Online. British Journal of Nutrition. Roč. 109, č. S2, s. S35-S50. ISSN 0007-1145. Dostupné z: https://doi.org/10.1017/S0007114512004011. [cit. 2024-05-17].

[2] SHARMA, Anshul; LEE, Sulhee a PARK, Young-Seo, 2020. Molecular typing tools for identifying and characterizing lactic acid bacteria: a review. Online. Food Science and Biotechnology. Roč. 29, č. 10, s. 1301-1318. ISSN 1226-7708. Dostupné z: https://doi.org/10.1007/s10068-020-00802-x.[cit. 2024-05-17].

[3] LI, Wenjun; RAOULT, Didier a FOURNIER, Pierre-Edouard. Bacterial strain typing in the genomic era. Online. FEMS Microbiology Reviews. 2009, roč. 33, č. 5, s. 892-916. ISSN 1574-6976. Dostupné z: https://doi.org/10.1111/j.1574-6976.2009.00182.x. [cit. 2023-11-17].

[4] FRASER-LIGGETT, Claire M. Insights on biology and evolution from microbial genome sequencing. Online. Genome Research. 2005, roč. 15, č. 12, s. 1603-1610. ISSN 1088-9051. Dostupné z: https://doi.org/10.1101/gr.3724205. [cit. 2023-11-17].

[5] SABAT, A. J., et al. Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Eurosurveillance, 2013, 18.4: 20380. Dostupné z: https://www.eurosurveillance.org/content/10.2807/ese.18.04.20380-en?TRACK=RSS

[6] RAMADAN, Asmaa A. Bacterial typing methods from past to present: A comprehensive overview. Online. Gene Reports. 2022, roč. 29. ISSN 24520144. Dostupné z: https://doi.org/10.1016/j.genrep.2022.101675. [cit. 2023-11-17].

[7] SCHWARZEROVA, Jana; LABANAVA, Anastasiya; RYCHLIK, Ivan; VARGA, Margaret a CEJKOVA, Darina. A minireview on the bioinformatics analysis of mobile gene elements in microbiome research. Online. Frontiers in Bacteriology. 2023, roč. 2. ISSN 2813-6144. Dostupné z: https://doi.org/10.3389/fbrio.2023.1275910. [cit. 2023-12-19].

[8] BONOFIGLIO, Laura; GARDELLA, Noella Mariel a MOLLERACH, Marta Eugenia. Application of Molecular Typing Methods to the Study of Medically

Relevant Gram-Positive Cocci. In: MAGDELIN, Sameh. Gel electrophoresis - Advanced Techniques. Rijeka, Croatia: InTech, 2012. ISBN 978-953-51-0457-5.

[9] TENOVER, F C; ARBEIT, R D; GOERING, R V; MICKELSEN, P A; MURRAY, B E et al. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. Online. Journal of Clinical Microbiology. 1995, roč. 33, č. 9, s. 2233-2239. ISSN 0095-1137. Dostupné z: https://doi.org/10.1128/jcm.33.9.2233-2239.1995. [cit. 2023-11-17].

[10] MACCANNELL, Duncan. Bacterial Strain Typing. Online. Clinics in Laboratory Medicine. 2013, roč. 33, č. 3, s. 629-650. ISSN 02722712. Dostupné z: https://doi.org/10.1016/j.cll.2013.03.005. [cit. 2023-11-17].

[11] VAN DER MERWE, R. G.; VAN HELDEN, P. D.; WARREN, R. M.; SAMPSON, S. L. a GEY VAN PITTIUS, N. C. Phage-based detection of bacterial pathogens. Online. The Analyst. 2014, roč. 139, č. 11, s. 2617-2626. ISSN 0003-2654. Dostupné z: https://doi.org/10.1039/C4AN00208C. [cit. 2023-11-17].

[12] JUNG, Ryan; KIM, Minzae; BHATT, Bhoomi; CHOI, Jong a ROH, Jung. Identification of Pathogenic Bacteria from Public Libraries via Proteomics Analysis. Online. International Journal of Environmental Research and Public Health. 2019, roč. 16, č. 6. ISSN 1660-4601. Dostupné z: https://doi.org/10.3390/ijerph16060912. [cit. 2023-11-17].

[13] SAYERS, Eric W; BOLTON, Evan E; BRISTER, J Rodney; CANESE, Kathi; CHAN, Jessica et al. Database resources of the national center for biotechnology information. Online. Nucleic Acids Research. 2022, roč. 50, č. D1, s. D20-D26. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkab1112. [cit. 2023-11-18].

[14] LEINONEN, R.; AKHTAR, R.; BIRNEY, E.; BOWER, L.; CERDENO-TARRAGA, A. et al. The European Nucleotide Archive. Online. Nucleic Acids Research. 2010, roč. 39, č. Database, s. D28-D31. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkq967. [cit. 2023-11-18].

[15] SMIGIELSKI, E. M. DbSNP: a database of single nucleotide polymorphisms. Online. Nucleic Acids Research. Roč. 28, č. 1, s. 352-355. ISSN 13624962. Dostupné z: https://doi.org/10.1093/nar/28.1.352. [cit. 2023-11-18].

[16] The International HapMap Project. Online. Nature. 2003, roč. 426, č. 6968, s. 789-796. ISSN 0028-0836. Dostupné z: https://doi.org/10.1038/nature02168. [cit. 2023-11-18].

[17] MARTAK, D.; MEUNIER, A.; SAUGET, M.; CHOLLEY, P.; THOUVEREZ, M. et al. Comparison of pulsed-field gel electrophoresis and whole-genome-sequencing-based typing confirms the accuracy of pulsed-field gel electrophoresis for the investigation of local Pseudomonas aeruginosa outbreaks. Online. Journal of Hospital Infection. 2020, roč. 105, č. 4, s. 643-647. ISSN 01956701. Dostupné z: https://doi.org/10.1016/j.jhin.2020.06.013. [cit. 2023-11-17].

[18] NEOH, Hui-min; TAN, Xin-Ee; SAPRI, Hassriana Fazilla a TAN, Toh Leong. Pulsed-field gel electrophoresis (PFGE): A review of the "gold standard" for bacteria typing and current alternatives. Online. Infection, Genetics and Evolution. 2019, roč. 74. ISSN 15671348. Dostupné z: https://doi.org/10.1016/j.meegid.2019.103935. [cit. 2023-11-17].

[19] YAN, Zhongqiang; ZHOU, Yu; DU, Mingmei; BAI, Yanling; LIU, Bowei et al. Prospective investigation of carbapenem-resistant Klebsiella pneumonia transmission among the staff, environment and patients in five major intensive care units, Beijing. Online. Journal of Hospital Infection. 2019, roč. 101, č. 2, s. 150-157. ISSN 01956701. Dostupné z: https://doi.org/10.1016/j.jhin.2018.11.019. [cit. 2023-11-17].

[20] PIRŠ, M.; CERAR KIŠEK, T.; KRIŽAN HERGOUTH, V.; SEME, K.; MUELLER PREMRU, M. et al. Successful control of the first OXA-48 and/or NDM carbapenemase-producing Klebsiella pneumoniae outbreak in Slovenia 2014–2016. Online. Journal of Hospital Infection. 2019, roč. 101, č. 2, s. 142-149. ISSN 01956701. Dostupné z: https://doi.org/10.1016/j.jhin.2018.10.022. [cit. 2023-11-17].

[21] DE LA ROSA-ZAMBONI, Daniela, et al. Everybody hands-on to avoid ESKAPE: effect of sustained hand hygiene compliance on healthcare-associated infections and multidrug resistance in a paediatric hospital. Journal of Medical Microbiology, 2018, 67.12: 1761-1771. Dostupné z: https://www.microbiologyresearch.org/content/journal/jmm/10.1099/jmm.0.000863

[22] LOPEZ-CANOVAS, Lilia; MARTINEZ BENITEZ, Maximo B.; HERRERA ISIDRON, Jose A. a FLORES SOTO, Eduardo. Pulsed Field Gel Electrophoresis: Past, present, and future. Online. Analytical Biochemistry. 2019, roč. 573, s. 17-29. ISSN 00032697. Dostupné z: https://doi.org/10.1016/j.ab.2019.02.020. [cit. 2023-11-17].

[23] ABUZENADAH, Adel. Restriction Fragment Length Polymorphism (RFLP). Online. In: . Dostupné z:https://www.kau.edu.sa/Files/0002923/Files/

18591_Restriction%20Fragment%20Length%20Polymorphism.pdf. [cit. 2023-11-18].

[24] RFLP Method - Restriction Fragment Length Polymorphism. Online. In: Www.bio.davidson.edu. 2001. Dostupné z: https://www.bio.davidson.edu/courses/genomics/method/rflp.html. [cit. 2023-11-18].

[25] Restriction Fragment Length Polymorphism (RFLP). Online. In: Www.ncbi.nlm.nih.gov. Dostupné z: https://www.ncbi.nlm.nih.gov/probe/docs/techrflp/. [cit. 2023-11-18].

[26] RFLP – restrikční reakce. Online. In: . Dostupné z: https://cit.vfu.cz/opvk2011/?title=popis_metod-rflp&lang=cz. [cit. 2023-11-18].

[27] SANGER, F.; NICKLEN, S. a COULSON, A. R. DNA sequencing with chain-terminating inhibitors. Online. Proceedings of the National Academy of Sciences. 1977, roč. 74, č. 12, s. 5463-5467. ISSN 0027-8424. Dostupné z: https://doi.org/10.1073/pnas.74.12.5463. [cit. 2023-11-30].

[28] LEWIS, T.; LOMAN, N.J.; BINGLE, L.; JUMAA, P.; WEINSTOCK, G.M. et al. High-throughput whole-genome sequencing to dissect the epidemiology of Acinetobacter baumannii isolates from a hospital outbreak. Online. Journal of Hospital Infection. 2010, roč. 75, č. 1, s. 37-41. ISSN 01956701. Dostupné z: https://doi.org/10.1016/j.jhin.2010.01.012. [cit. 2023-11-30].

[29] METZKER, Michael L. Sequencing technologies — the next generation. Online. Nature Reviews Genetics. 2010, roč. 11, č. 1, s. 31-46. ISSN 1471-0056. Dostupné z: https://doi.org/10.1038/nrg2626. [cit. 2023-11-30].

[30] PALLEN, Mark J; LOMAN, Nicholas J a PENN, Charles W. High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. Online. Current Opinion in Microbiology. 2010, roč. 13, č. 5, s. 625-631. ISSN 13695274. Dostupné z: https://doi.org/10.1016/j.mib.2010.08.003. [cit. 2023-11-30].

[31] ROSSEN, J.W.A.; FRIEDRICH, A.W. a MORAN-GILAD, J. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. Online. Clinical Microbiology and Infection. 2018, roč. 24, č. 4, s. 355-360. ISSN 1198743X. Dostupné z: https://doi.org/10.1016/j.cmi.2017.11.001. [cit. 2023-11-30].

[32] DIDELOT, Xavier, et al. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome biology, 2012, 13: 1-13. Dostupné z: https://link.springer.com/article/10.1186/gb-2012-13-12-r118

[33] DIDELOT, Xavier; EYRE, David W; CULE, Madeleine; IP, Camilla LC; ANSARI, M et al. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Online. Genome Biology. 2012, roč. 13, č. 12. ISSN 1465-6906. Dostupné z: https://doi.org/10.1186/gb-2012-13-12-r118. [cit. 2023-11-30].

[34] HASAN, Nur A.; CHOI, Seon Young; EPPINGER, Mark; CLARK, Philip W.; CHEN, Arlene et al. Genomic diversity of 2010 Haitian cholera outbreak strains. Online. Proceedings of the National Academy of Sciences. 2012, roč. 109, č. 29. ISSN 0027-8424. Dostupné z: https://doi.org/10.1073/pnas.1207359109. [cit. 2023-11-30].

[35] BOONSILP, Siriphan; THAIPADUNGPANIT, Janjira; AMORNCHAI, Premjit; WUTHIEKANUN, Vanaporn; BAILEY, Mark S. et al. A Single Multilocus Sequence Typing (MLST) Scheme for Seven Pathogenic Leptospira Species. Online. PLoS Neglected Tropical Diseases. 2013, roč. 7, č. 1. ISSN 1935-2735. Dostupné z: https://doi.org/10.1371/journal.pntd.0001954. [cit. 2023-11-19].

[36] GODORNES, Charmie; GIACANI, Lorenzo; BARRY, Alyssa E.; MITJA, Oriol; LUKEHART, Sheila A. et al. Development of a Multilocus Sequence Typing (MLST) scheme for Treponema pallidum subsp. pertenue: Application to yaws in Lihir Island, Papua New Guinea. Online. PLOS Neglected Tropical Diseases. 2017, roč. 11, č. 12. ISSN 1935-2735. Dostupné z: https://doi.org/10.1371/journal.pntd.0006113. [cit. 2023-11-19].

[37] MAIDEN, Martin C.J. Multilocus Sequence Typing of Bacteria. Online. Annual Review of Microbiology. 2006, roč. 60, č. 1, s. 561-588. ISSN 0066-4227. Dostupné z: https://doi.org/10.1146/annurev.micro.59.030804.121325. [cit. 2023-11-30].

[38] AANENSEN, D. M. a SPRATT, B. G. The multilocus sequence typing network: mlst.net. Online. Nucleic Acids Research. 2005, roč. 33, č. Web Server, s. W728-W733. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gki415. [cit. 2023-11-30].

[39] MAIDEN, Martin C. J.; BYGRAVES, Jane A.; FEIL, Edward; MORELLI, Giovanna; RUSSELL, Joanne E. et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenicmicroorganisms. Online. Proceedings of the National Academy of Sciences. 1998, roč. 95, č. 6, s. 3140-3145. ISSN 0027-8424. Dostupné z: https://doi.org/10.1073/pnas.95.6.3140. [cit. 2023-11-30].

[40] OAKLEY, Brian B.; GONZALEZ-ESCALONA, Narjol a MOLINA, Marirosa. 12. Molecular Typing and Differentiation. Online. In: SALFINGER, Yvonne a TORTORELLO, Mary Lou (ed.). Compendium of Methods for the Microbiological Examination of Foods. American Public Health Association, 2015. ISBN 978-0-87553-022-2. Dostupné z: https://doi.org/10.2105/MBEF.0222.017. [cit. 2023-11-30].

[41] FOLEY, Steven L.; LYNNE, Aaron M. a NAYAK, Rajesh. Molecular typing methodologies for microbial source tracking and epidemiological investigations of Gram-negative bacterial foodborne pathogens. Online. Infection, Genetics and Evolution. 2009, roč. 9, č. 4, s. 430-440. ISSN 15671348. Dostupné z: https://doi.org/10.1016/j.meegid.2009.03.004. [cit. 2023-11-30].

[42] BESSER, J.; CARLETON, H.A.; GERNER-SMIDT, P.; LINDSEY, R.L. a TREES, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Online. Clinical Microbiology and Infection. 2018, roč. 24, č. 4, s. 335-341. ISSN 1198743X. Dostupné z: https://doi.org/10.1016/j.cmi.2017.10.013. [cit. 2023-11-30].

[43] PARCELL, B.J.; ORAVCOVA, K.; PINHEIRO, M.; HOLDEN, M.T.G.; PHILLIPS, G. et al. Pseudomonas aeruginosa intensive care unit outbreak: winnowing of transmissions with molecular and genomic typing. Online. Journal of Hospital Infection. 2018, roč. 98, č. 3, s. 282-288. ISSN 01956701. Dostupné z: https://doi.org/10.1016/j.jhin.2017.12.005. [cit. 2023-11-30].

[44] DAVIS, Steve; PETTENGILL, James B.; LUO, Yan; PAYNE, Justin; SHPUNTOFF, Al et al. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. Online. PeerJ Computer Science. 2015, roč. 1. ISSN 2376-5992. Dostupné z: https://doi.org/10.7717/peerj-cs.20. [cit. 2023-11-30].

[45] KATZ, Lee S.; GRISWOLD, Taylor; WILLIAMS-NEWKIRK, Amanda J.; WAGNER, Darlene; PETKAU, Aaron et al. A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. Online. Frontiers in Microbiology. 2017, roč. 8. ISSN 1664-302X. Dostupné z: https://doi.org/10.3389/fmicb.2017.00375. [cit. 2023-11-30].

[46] ZOU, Quan; LIN, Gang; JIANG, Xingpeng; LIU, Xiangrong a ZENG, Xiangxiang. Sequence clustering in bioinformatics: an empirical study. Online. Briefings in Bioinformatics. 2018. ISSN 1467-5463. Dostupné z: https://doi.org/10.1093/bib/bby090. [cit. 2023-11-07].

[47] KIM, Tùng T. a POOR, H. Vincent. Strategic Protection Against Data Injection Attacks on Power Grids. Online. IEEE Transactions on Smart Grid. 2011, roč. 2, č. 2, s. 326-333. ISSN 1949-3053. Dostupné z: https://doi.org/10.1109/TSG.2011.2119336. [cit. 2023-11-07].

[48] JU, Zhen; ZHANG, Huiling; MENG, Jintao; ZHANG, Jingjing; FAN, Jianping et al. NGIA: A novel Greedy Incremental Alignment based algorithm for gene sequence clustering. Online. Future Generation Computer Systems. 2022, roč. 136, s. 221-230. ISSN 0167739X. Dostupné z: https://doi.org/10.1016/j.future.2022.05.024. [cit. 2023-11-07].

[49] *JONES, Neil C. a PEVZNER, Pavel. An introduction to bioinformatics algorithms. Cambridge, MA: MIT Press, c2004. ISBN 0262101068.*

[50] SONG, Mingjun a RAJASEKARAN, Sanguthevar. A greedy algorithm for gene selection based on SVM and correlation. Online. International Journal of Bioinformatics Research and Applications. 2010, roč. 6, č. 3. ISSN 1744-5485. Dostupné z: https://doi.org/10.1504/IJBRA.2010.034077. [cit. 2023-11-07].

[51] TUFFERY, Pierre; GUYON, Frédéric a DERREUMAUX, Philippe. Improved greedy algorithm for protein structure reconstruction. Online. Journal of Computational Chemistry. 2005, roč. 26, č. 5, s. 506-513. ISSN 0192-8651. Dostupné z: https://doi.org/10.1002/jcc.20181. [cit. 2023-11-07].

[52] FINK, Andreas a VOSS, Stefan. Applications of modern heuristic search methods to pattern sequencing problems. Online. Computers & Operations Research. 1999, roč. 26, č. 1, s. 17-34. ISSN 03050548. Dostupné z: https://doi.org/10.1016/S0305-0548(98)80001-4. [cit. 2023-11-07].

[53] RASHEDI, Esmat; NEZAMABADI-POUR, Hossien a SARYAZDI, Saeid. Filter modeling using gravitational search algorithm. Online. Engineering Applications of Artificial Intelligence. 2011, roč. 24, č. 1, s. 117-122. ISSN 09521976. Dostupné z: https://doi.org/10.1016/j.engappai.2010.05.007. [cit. 2023-11-07].

[54] SUN, Yanni a BUHLER, Jeremy. Online. BMC Bioinformatics. Roč. 7, č. 1. ISSN 14712105. Dostupné z: https://doi.org/10.1186/1471-2105-7-133. [cit. 2023-11-07].

[55] DE KRETSER, Owen a MOFFAT, Alistair. Effective document presentation with a locality-based similarity heuristic. Online. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 1999, s. 113-120.

ISBN 1581130961. Dostupné z: https://doi.org/10.1145/312624.312664. [cit. 2023-11-07].

[56] EDDY, Sean R. What is a hidden Markov model? Online. Nature Biotechnology. 2004, roč. 22, č. 10, s. 1315-1316. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/nbt1004-1315. [cit. 2023-11-30].

[57] EDDY, Sean R. Hidden Markov models. Online. Current Opinion in Structural Biology. 1996, roč. 6, č. 3, s. 361-365. ISSN 0959440X. Dostupné z: https://doi.org/10.1016/S0959-440X(96)80056-X. [cit. 2023-11-30].

[58] YOON, Byung-Jun. Hidden Markov Models and their Applications in Biological Sequence Analysis. Online. Current Genomics. 2009, roč. 10, č. 6, s. 402-415. ISSN 13892029. Dostupné z: https://doi.org/10.2174/138920209789177575. [cit. 2023-11-07].

[59] HENDERSON, JOHN; SALZBERG, STEVEN a FASMAN, KENNETH H. Finding Genes in DNA with a Hidden Markov Model. Online. Journal of Computational Biology. 1997, roč. 4, č. 2, s. 127-141. ISSN 1066-5277. Dostupné z: https://doi.org/10.1089/cmb.1997.4.127. [cit. 2023-11-07].

[60] KAMAL, Md. Sarwar; CHOWDHURY, Linkon; KHAN, Mohammad Ibrahim; ASHOUR, Amira S.; TAVARES, João Manuel R.S. et al. Hidden Markov model and Chapman Kolmogrov for protein structures prediction from images. Online. Computational Biology and Chemistry. 2017, roč. 68, s. 231-244. ISSN 14769271. Dostupné z: https://doi.org/10.1016/j.compbiolchem.2017.04.003. [cit. 2023-11-07].

[61] VINGA, S. Editorial: Alignment-free methods in computational biology. Online. Briefings in Bioinformatics. 2014, roč. 15, č. 3, s. 341-342. ISSN 1467-5463. Dostupné z: https://doi.org/10.1093/bib/bbu005. [cit. 2023-11-07].

[62] VINGA, Susana a ALMEIDA, Jonas. Alignment-free sequence comparison—a review. Online. Bioinformatics. 2003, roč. 19, č. 4, s. 513-523. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/btg005. [cit. 2023-11-07].

[63] ZHANG, Qian; JUN, Se-Ran; LEUZE, Michael; USSERY, David a NOOKAEW, Intawat. Viral Phylogenomics Using an Alignment-Free Method: A Three-Step Approach to Determine Optimal Length of k-mer. Online. Scientific Reports. 2017, roč. 7, č. 1. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/srep40712. [cit. 2023-11-07].

[64] COHEN, Shani; ROKACH, Lior; MOTRO, Yair; MORAN-GILAD, Jacob; VEKSLER-LUBLINSKY, Isana et al. MinMLST: machine learning for optimization of bacterial strain typing. Online. Bioinformatics. 2021, roč. 37, č. 3, s. 303-311. ISSN 1367-4803. Dostupné z: https://doi.org/10.1093/bioinformatics/btaa724. [cit. 2023-12-21].

[65] DEKKER, John P.; FRANK, Karen M. a BOURBEAU, P. Commentary: Next-Generation Epidemiology. Online. Journal of Clinical Microbiology. 2016, roč. 54, č. 12, s. 2850-2853. ISSN 0095-1137. Dostupné z: https://doi.org/10.1128/JCM.01714-16. [cit. 2023-12-21].

[66] MAIDEN, Martin C. J.; VAN RENSBURG, Melissa J. Jansen; BRAY, James E.; EARLE, Sarah G.; FORD, Suzanne A. et al. MLST revisited: the gene-by-gene approach to bacterial genomics. Online. Nature Reviews Microbiology. 2013, roč. 11, č. 10, s. 728-736. ISSN 1740-1526. Dostupné z: https://doi.org/10.1038/nrmicro3093. [cit. 2023-12-21].

[67] MAIMON, Oded a ROKACH, Lior (ed.). Data Mining and Knowledge Discovery Handbook. Online. Boston, MA: Springer US, 2010. ISBN 978-0-387-09822-7. Dostupné z: https://doi.org/10.1007/978-0-387-09823-4. [cit. 2023-12-21].

[68] LUNDBERG, Scott M.; ERION, Gabriel; CHEN, Hugh; DEGRAVE, Alex; PRUTKIN, Jordan M. et al. From local explanations to global understanding with explainable AI for trees. Online. Nature Machine Intelligence. 2020, roč. 2, č. 1, s. 56-67. ISSN 2522-5839. Dostupné z: https://doi.org/10.1038/s42256-019-0138-9. [cit. 2023-12-21].

[69] HUBERT, Lawrence a ARABIE, Phipps. Comparing partitions. Online. Journal of Classification. 1985, roč. 2, č. 1, s. 193-218. ISSN 0176-4268. Dostupné z: https://doi.org/10.1007/BF01908075. [cit. 2023-12-21].

[70] SCHWARZEROVA, Jana; ZEMAN, Michal; BABAK, Vladimir; JURECKOVA, Katerina; NYKRYNOVA, Marketa et al. Detecting horizontal gene transfer among microbiota: an innovative pipeline for identifying co-shared genes within the mobilome through advanced comparative analysis. Online. Microbiology Spectrum. S. e01964-23. ISSN 2165-0497. Dostupné z: https://doi.org/10.1128/spectrum.01964-23. [cit. 2023-12-19].

[71] MEDVECKY, Matej; CEJKOVA, Darina; POLANSKY, Ondrej; KARASOVA, Daniela; KUBASOVA, Tereza et al. Whole genome sequencing and function prediction of 133 gut anaerobes isolated from chicken caecum in pure cultures.

Online. BMC Genomics. 2018, roč. 19, č. 1. ISSN 1471-2164. Dostupné z: https://doi.org/10.1186/s12864-018-4959-4. [cit. 2023-11-30].

[72] JURICOVA, Helena; MATIASOVICOVA, Jitka; KUBASOVA, Tereza; CEJKOVA, Darina a RYCHLIK, Ivan, 2021. The distribution of antibiotic resistance genes in chicken gut microbiota commensals. Online. Scientific Reports. Roč. 11, č. 1. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-021-82640-3. [cit. 2024-05-12].

[73] BOLGER, Anthony M.; LOHSE, Marc a USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. Online. Bioinformatics. 2014, roč. 30, č. 15, s. 2114-2120. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/btu170. [cit. 2023-04-01].

[74] PENG, Yu; LEUNG, Henry C. M.; YIU, S. M. a CHIN, Francis Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Online. Bioinformatics. 2012, roč. 28, č. 11, s. 1420-1428. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/bts174. [cit. 2023-11-30].

[75] JOLLEY, Keith A.; BLISS, Carly M.; BENNETT, Julia S.; BRATCHER, Holly B.; BREHONY, Carina et al. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Online. Microbiology. 2012, roč. 158, č. 4, s. 1005-1015. ISSN 1350-0872. Dostupné z: https://doi.org/10.1099/mic.0.055459-0. [cit. 2023-11-30].

[76] OVERBEEK, Ross; OLSON, Robert; PUSCH, Gordon D.; OLSEN, Gary J.; DAVIS, James J. et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Online. Nucleic Acids Research. 2013, roč. 42, č. D1, s. D206-D214. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkt1226. [cit. 2023-11-30].

[77] NA, Seong-In; KIM, Yeong Ouk; YOON, Seok-Hwan; HA, Sung-min; BAEK, Inwoo et al. UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. Online. Journal of Microbiology. 2018, roč. 56, č. 4, s. 280-285. ISSN 1225-8873. Dostupné z: https://doi.org/10.1007/s12275-018-8014-6. [cit. 2023-12-21].

[78] LETUNIC, Ivica a BORK, Peer. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Online. Nucleic Acids Research. 2021, roč. 49, č. W1, s. W293-W296. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkab301. [cit. 2023-12-21].

[79] TAMANA, Sukhpreet K.; TUN, Hein M.; KONYA, Theodore; CHARI, Radha S.; FIELD, Catherine J. et al. Bacteroides-dominant gut microbiome of late infancy is associated with enhanced neurodevelopment. Online. Gut Microbes. 2021, roč. 13, č. 1. ISSN 1949-0976. Dostupné z: https://doi.org/10.1080/19490976.2021.1930875. [cit. 2023-11-22].

[80] GRIGOR'EVA, Irina N. Gallstone Disease, Obesity and the Firmicutes/Bacteroidetes Ratio as a Possible Biomarker of Gut Dysbiosis. Online. Journal of Personalized Medicine. 2021, roč. 11, č. 1. ISSN 2075-4426. Dostupné z: https://doi.org/10.3390/jpm11010013. [cit. 2023-11-22].

[81] CHANDRANGSU, Pete; LOI, Vu Van; ANTELMANN, Haike a HELMANN, John D. The Role of Bacillithiol in Gram-Positive Firmicutes. Online. Antioxidants & Redox Signaling. 2018, roč. 28, č. 6, s. 445-462. ISSN 1523-0864. Dostupné z: https://doi.org/10.1089/ars.2017.7057. [cit. 2023-11-22].

[82] VESTH, Tammi; OZEN, Aslı; ANDERSEN, Sandra C.; KAAS, Rolf Sommer; LUKJANCENKO, Oksana et al. Veillonella, Firmicutes: Microbes disguised as Gram negatives. Online. Standards in Genomic Sciences. 2013, roč. 9, č. 3, s. 431-448. ISSN 1944-3277. Dostupné z: https://doi.org/10.4056/sigs.2981345. [cit. 2023-11-22].

[83] HAZARIKA, Shabiha Nudrat a THAKUR, Debajit. Actinobacteria. Online. In: Beneficial Microbes in Agro-Ecology. Elsevier, 2020, s. 443-476. ISBN 9780128234143. Dostupné z: https://doi.org/10.1016/B978-0-12-823414-3.00021-6. [cit. 2023-11-22].

[84] UL-HASSAN, A. a WELLINGTON, E.M. Actinobacteria. Online. In: Encyclopedia of Microbiology. Elsevier, 2009, s. 25-44. ISBN 9780123739445. Dostupné z: https://doi.org/10.1016/B978-012373944-5.00044-4. [cit. 2023-11-22].

[85] VAZ-MOREIRA, Ivone; NUNES, Olga C. a MANAIA, Célia M. Ubiquitous and persistent Proteobacteria and other Gram-negative bacteria in drinking water. Online. Science of The Total Environment. 2017, roč. 586, s. 1141-1149. ISSN 00489697. Dostupné z: https://doi.org/10.1016/j.scitotenv.2017.02.104. [cit. 2023-11-22].

[86] GENG, Jianing; LUO, Sainan; SHIEH, Hui-Ru; WANG, Hsing-Yi; HU, Songnian et al. Identification of a Putative CodY Regulon in the Gram-Negative Phylum Synergistetes. Online. International Journal of Molecular Sciences. 2022, roč. 23, č. 14. ISSN 1422-0067. Dostupné z: https://doi.org/10.3390/ijms23147911. [cit. 2023-11-22].

[87] RIZZATTI, G.; LOPETUSO, L. R.; GIBIINO, G.; BINDA, C. a GASBARRINI, A. Proteobacteria: A Common Factor in Human Diseases. Online. BioMed Research International. 2017, roč. 2017, s. 1-7. ISSN 2314-6133. Dostupné z: https://doi.org/10.1155/2017/9351507. [cit. 2023-11-22].

[88] BIAN, Xiaoyuan; WU, Wenrui; YANG, Liya; LV, Longxian; WANG, Qing et al. Administration of Akkermansia muciniphila Ameliorates Dextran Sulfate Sodium-Induced Ulcerative Colitis in Mice. Online. Frontiers in Microbiology. 2019, roč. 10. ISSN 1664-302X. Dostupné z: https://doi.org/10.3389/fmicb.2019.02259. [cit. 2023-11-22].

[89] BRUNE, A.; MIES, U. S.; TRUJILLO, M. E.; DEDYSH, S.; DEVOS, P. et al. Elusimicrobiota. Online. In: Bergey's Manual of Systematics of Archaea and Bacteria. 2023. Dostupné z: https://hdl.handle.net/21.11116/0000-000D-5B51-2. [cit. 2023-11-22].

[90] MCCRACKEN, Barbara Anne a NATHALIA GARCIA, M. Phylum Synergistetes in the oral cavity: A possible contributor to periodontal disease. Online. Anaerobe. 2021, roč. 68. ISSN 10759964. Dostupné z: https://doi.org/10.1016/j.anaerobe.2020.102250. [cit. 2023-11-22].

[91] DE WITTE, Chloë; DEMEYERE, Kristel; DE BRUYCKERE, Sofie; TAMINIAU, Bernard; DAUBE, Georges et al. Characterization of the non-glandular gastric region microbiota in Helicobacter suis-infected versus non-infected pigs identifies a potential role for Fusobacterium gastrosuis in gastric ulceration. Online. Veterinary Research. 2019, roč. 50, č. 1. ISSN 1297-9716. Dostupné z: https://doi.org/10.1186/s13567-019-0656-9. [cit. 2023-12-03].

[92] ALTSCHUL, Stephen F.; GISH, Warren; MILLER, Webb; MYERS, Eugene W. a LIPMAN, David J. Basic local alignment search tool. Online. Journal of Molecular Biology. 1990, roč. 215, č. 3, s. 403-410. ISSN 00222836. Dostupné z: https://doi.org/10.1016/S0022-2836(05)80360-2. [cit. 2023-11-30].

[93] KERFELD, Cheryl A.; SCOTT, Kathleen M. a KERFELD, Cheryl A. Using BLAST to Teach "E-value-tionary" Concepts. Online. PLoS Biology. 2011, roč. 9, č. 2. ISSN 1545-7885. Dostupné z: https://doi.org/10.1371/journal.pbio.1001014. [cit. 2023-11-30].

[94] NCBI C++ Toolkit Cross Reference. Online. In: . Dostupné z: https://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/scripts/projects/blast/LICENSE. [cit. 2024-05-19].

[95] MCGINNIS, S. a MADDEN, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Online. Nucleic Acids Research. 2004, roč. 32, č. Web Server, s. W20-W25. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkh435. [cit. 2023-11-30].

[96] WHEELER, David a BHAGWAT, Medha. BLAST QuickStart. Online. In: BERGMAN, Nicholas H. (ed.). Comparative Genomics. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2008, s. 149-175. ISBN 978-1-58829-693-1. Dostupné z: https://doi.org/10.1007/978-1-59745-514-5_9. [cit. 2023-11-30].

[97] DISEGHA, G. C.; JEAPUDOARI, T. F. Sequence Alignment as a Method of Bacterial Identification. Current Studies in Comparative Education, Science and Technology, 2017, 4.1: 221-238. Dostupné z: http://www.journal.iscest.org/wp-content/uploads/2016/07/DISEGHA-G.C.-JEAPUDOARI-T.-F.pdf. [cit. 2023-11-30].

[98] ZOLFO, Moreno; TETT, Adrian; JOUSSON, Olivier; DONATI, Claudio a SEGATA, Nicola. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples. Online. Nucleic Acids Research. 2017, roč. 45, č. 2, s. e7-e7. ISSN 0305-1048. Dostupné z: https://doi.org/10.1093/nar/gkw837. [cit. 2023-11-30].

[99] LICENSE, GNU General Public, 1989. Gnu general public license.

[100] LI, Weizhong; JAROSZEWSKI, Lukasz a GODZIK, Adam. Tolerating some redundancy significantly speeds up clustering of large protein databases. Online. Bioinformatics. 2002, roč. 18, č. 1, s. 77-82. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/18.1.77. [cit. 2023-11-30].

[101] LI, Weizhong; JAROSZEWSKI, Lukasz a GODZIK, Adam. Tolerating some redundancy significantly speeds up clustering of large protein databases. Online. Bioinformatics. 2002, roč. 18, č. 1, s. 77-82. ISSN 1367-4811. Dostupné z: https://doi.org/10.1093/bioinformatics/18.1.77. [cit. 2023-11-30].

[102] ARREDONDO-ALONSO, Sergio; GLADSTONE, Rebecca A; PÖNTINEN, Anna K; GAMA, João A; SCHÜRCH, Anita C et al. Mge-cluster: a reference-free approach for typing bacterial plasmids. Online. NAR Genomics and Bioinformatics. 2023, roč. 5, č. 3. ISSN 2631-9268. Dostupné z: https://doi.org/10.1093/nargab/lqad066. [cit. 2023-11-30].

[103] GNU General Public License, version 3, 2007. Online. Dostupné z: http://www.gnu.org/licenses/gpl.html. [cit. 2024-05-19].

[104] SEEMANN, Torsten. Barrnap. Online. 2018. Dostupné z: https://github.com/tseemann/barrnap. [cit. 2023-11-30].

[105] PAGE, Andrew J.; AINSWORTH, Emma V. a LANGRIDGE, Gemma C. Socru: typing of genome-level order and orientation around ribosomal operons in bacteria. Online. Microbial Genomics. 2020, roč. 6, č. 7. ISSN 2057-5858. Dostupné z: https://doi.org/10.1099/mgen.0.000396. [cit. 2023-11-30].

[106] PAGE, Andrew J. a LANGRIDGE, Gemma C. Socru: Typing of genome level order and orientation in bacteria. Online. Dostupné z: https://doi.org/doi.org/10.1101/543702. [cit. 2023-11-30].

[107] CHAU, Stephanie; ROJAS, Carlos; JETCHEVA, Jorjeta G.; VIJAYAKUMAR, Sudha; YUAN, Sophia et al. On the synergies between ribosomal assembly and machine learning tools for microbial identification. Online. Dostupné z: https://doi.org/10.1101/2022.09.30.510284. [cit. 2023-11-30].

[108] ALBANESE, Davide a DONATI, Claudio. Large-scale quality assessment of prokaryotic genomes with metashot/prok-quality. Online. F1000Research. 2021, roč. 10. ISSN 2046-1402. Dostupné z: https://doi.org/10.12688/f1000research.54418.1. [cit. 2023-11-30].

[109] FERRERA, Isabel; GINER, Caterina R.; REÑÉ, Albert; CAMP, Jordi; MASSANA, Ramon et al. Evaluation of Alternative High-Throughput Sequencing Methodologies for the Monitoring of Marine Picoplanktonic Biodiversity Based on rRNA Gene Amplicons. Online. Frontiers in Marine Science. 2016, roč. 3. ISSN 2296-7745. Dostupné z: https://doi.org/10.3389/fmars.2016.00147. [cit. 2023-11-26].

[110] SINCLAIR, Andrew, 2010. Licence Profile: Apache License, Version 2.0. Online. International Free and Open Source Software Law Review. 2010-12-31, roč. 2, č. 2, s. 107-114. ISSN 18776922. Dostupné z: https://doi.org/10.5033/ifosslr.v2i2.42. [cit. 2024-05-19].

[111] JAIN, Chirag; RODRIGUEZ-R, Luis M.; PHILLIPPY, Adam M.; KONSTANTINIDIS, Konstantinos T. a ALURU, Srinivas. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Online. Nature Communications. 2018, roč. 9, č. 1. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/s41467-018-07641-9. [cit. 2023-11-30].

[112] CIUFO, Stacy; KANNAN, Sivakumar; SHARMA, Shobha; BADRETDIN, Azat; CLARK, Karen et al. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Online. International

Journal of Systematic and Evolutionary Microbiology. 2018, roč. 68, č. 7, s. 2386-2392. ISSN 1466-5026. Dostupné z: https://doi.org/10.1099/ijsem.0.002809. [cit. 2023-11-26].

[113] GORIS, Johan; KONSTANTINIDIS, Konstantinos T.; KLAPPENBACH, Joel A.; COENYE, Tom; VANDAMME, Peter et al. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Online. International Journal of Systematic and Evolutionary Microbiology. 2007, roč. 57, č. 1, s. 81-91. ISSN 1466-5026. Dostupné z: https://doi.org/10.1099/ijs.0.64483-0. [cit. 2023-11-26].

[114] LIANG, Qian; LIU, Chengzhi; XU, Rong; SONG, Minghui; ZHOU, Zhihui et al. FIDBAC: A Platform for Fast Bacterial Genome Identification and Typing. Online. Frontiers in Microbiology. 2021, roč. 12. ISSN 1664-302X. Dostupné z: https://doi.org/10.3389/fmicb.2021.723577. [cit. 2023-11-30].

[115] ROBERTSON, James; BESSONOV, Kyrylo; SCHONFELD, Justin a NASH, John H. E. Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. Online. Microbial Genomics. 2020, roč. 6, č. 10. ISSN 2057-5858. Dostupné z: https://doi.org/10.1099/mgen.0.000435. [cit. 2023-11-30].

[116] BLAT: BLAST-Like Alignment Tool. Online. GitHub. Dostupné z: https://github.com/djhshih/blat/blob/master/LICENSE.txt. [cit. 2024-05-19].

[117] KENT, W. James, 2002. BLAT —The BLAST -Like Alignment Tool. Online. Genome Research. 2002-04-01, roč. 12, č. 4, s. 656-664. ISSN 1088-9051. Dostupné z: https://doi.org/10.1101/gr.229202. [cit. 2024-05-02].

[118] GRINBERG, Miguel, 2018. Flask web development: developing web applications with python. O'Reilly Media.

[119] The Web Server Gateway Interface (WSGI), 2009. Online. In: The Definitive Guide to Pylons. Berkeley, CA: Apress, s. 369-388. ISBN 978-1-59059-934-1. Dostupné z: https://doi.org/10.1007/978-1-4302-0534-0_16. [cit. 2024-05-02].

[120] COPPERWAITE, Matt a LEIFER, Charles, 2015. Learning flask framework. Packt Publishing Ltd. ISBN 978-1-78398-336-0.

[121] Lekce 1 - Úvod do frameworku Flask a webových aplikací v Pythonu. Online. In: . Dostupné z: https://www.itnetwork.cz/python/flask/

uvod-do-frameworku-flask-a-webovych-aplikaci-v-pythonu. [cit. 2024-05-02].

[122] MERKEL, Dirk, et al., 2014. Docker: lightweight linux containers for consistent development and deployment. Linux j.

[123] CHELLADHURAI, Jeeva S.; SINGH, Vinod a RAJ, Pethuru, 2017. Learning Docker: faster app development and deployment with Docker containers. 2nd ed. Birmingham: Packt Publishing. ISBN 978-1-78646-292-3.

[124] AXMARK, David a WIDENIUS, Michael. MySQL 8.3 reference manual. Online. In: . Dostupné z: https://dev.mysql.com/doc/refman/8.3/en/. [cit. 2024-05-12].

[125] FRENCH, Robert M., 2000. The Turing Test: the first 50 years. Online. Trends in Cognitive Sciences. Roč. 4, č. 3, s. 115-122. ISSN 13646613. Dostupné z: https://doi.org/10.1016/S1364-6613(00)01453-4. [cit. 2024-04-18].

[126] PINAR SAYGIN, Ayse; CICEKLI, Ilyas a AKMAN, Varol. Online. Minds and Machines. Roč. 10, č. 4, s. 463-518. ISSN 09246495. Dostupné z: https://doi.org/10.1023/A:1011288000451. [cit. 2024-04-18].

[127] Captcha. Online. Dostupné z: https://captcha.lepture.com/. [cit. 2024-04-18].

[128] Jquery. Online. Dostupné z: https://jquery.com/. [cit. 2024-04-18].

[129] Sweetalert2. Online. Dostupné z: https://sweetalert2.github.io/. [cit. 2024-04-18].

[130] NEJEZCHLEBOVÁ, Julie; RYCHLÍK, Ivan a JANA, Schwarzerová, 2024. Bacterial Identifier: Accelerating Bacterial Genome Detection. In: STUDENT EEICT 2024. Brno: Fakulta elektrotechniky a komunikacnich technologii VUT v Brne. [cit. 2024-05-21]. (v tisku)

[131] LUNDH, Fredrik. An introduction to tkinter. Online. 1999. Dostupné z: www.pythonware.com/library/tkinter/introduction/index.htm. [cit. 2023-11-05].

[132] STROUHAL, Michal; MIKALOVÁ, Lenka; HAVIERNIK, Jan; KNAUF, Sascha; BRUISTEN, Sylvia et al., 2018. Complete genome sequences of two strains of Treponema pallidum subsp. pertenue from Indonesia: Modular structure of several treponemal genes. Online. PLOS Neglected Tropical Diseases. 2018-10-10, roč. 12, č. 10. ISSN 1935-2735. Dostupné z: https://doi.org/10.1371/journal.pntd.0006867. [cit. 2024-05-22].

[133] MANNION, Anthony; DZINK-FOX, JoAnn; SHEN, Zeli; PIAZUELO, M. Blanca; WILSON, Keith T. et al., 2021. Helicobacter pylori Antimicrobial Resistance and Gene Variants in High- and Low-Gastric-Cancer-Risk Populations. Online. Journal of Clinical Microbiology. 2021-04-20, roč. 59, č. 5, s. e03203-20. ISSN 0095-1137. Dostupné z: https://doi.org/10.1128/JCM.03203-20. [cit. 2024-05-22].

[134] REISTER, Marten; HOFFMEIER, Klaus; KREZDORN, Nicolas; ROTTER, Bjoern; LIANG, Chunguang et al., 2014. Complete genome sequence of the Gram-negative probiotic Escherichia coli strain Nissle 1917. Online. Journal of Biotechnology. Roč. 187, s. 106-107. ISSN 01681656. Dostupné z: https://doi.org/10.1016/j.jbiotec.2014.07.442. [cit. 2024-05-22].

Seznam symbolů a zkratek

AFMA Algoritmus mapování bez zarovnání (Alignment-free mapping algoritmus)

ANI Průměrná nukleotidová identita (Average nucleotide identity)

ARI Upravený náhodný index (Adjusted Rand Index)

BBH Obousměrně nejlepší shody (Bidirectional best hits)

BLAST Basic Local Alignment Search Tool

BLAT Bast-Like Alignment Tool

Barrnap Bacterial ribosomal rRNA predictor

cgMLST Multilokusová sekvenční typizace (Core genome Multilocus sequence typing)

DNA Deoxyribonukleová kyselina

ERIC PCR Enterobakteriální repetitivní intergenní konsenzus polymerázové řetězové reakce

HMM Hidden Markov model

iTOL Interaktivní strom života (Interactive Tree of Life)

MLST Multilokusová sekvenční typizace (Multilocus sequence typing)

MLVA Analýza variabilního počtu repetic ve více lokusech (Multiple-locus variable number tandem repeat analysis)

NCBI Národní centrum pro bioinformatické informace (National Center for Biotechnology Information)

NGS Nová generace sekvenování (Next Generation Sequencing)

PCR Polymerázová řetězcová reakce (Polymerase chain reaction)

PFGE Pulsní gelová elektroforéza (Pulsed-field gel electrophoresis)

rep-PCR Repetitivní sekvenční polymerázová řetězcová reakce (Repetitive-element polymerase chain reaction)

RFLP Polymorfismus délky restrikčních enzymů (Restriction fragment length polymorphism)

SHAP	Shapleyho aditivní hodnoty (Shapley Additive explanations)
SLST	Jednolokusová sekvenční typizace (Single locus sequences typing)
SNP	Jednonukleotidový polymorfismus (Single nucleotide polymorphism)
WCHA	Wilkins-Chalgrenův anaerobní agar
wgMLST	Multilokusová sekvenční typizace (Whole genome Multi locus sequence typing)
WGS	Celogenomové sekvenování (Whole genome sequencing)
WSGI	Web Server Gateway Interface

A Tabulky

Tab. A.1: Výsledky testování I pro bakterie z kmene *Bacteroides* pro metodu 1 s prahem 99% při volbě nástrojů Cd-hit, BLAT a BLAST.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**
Bacteroides helcogenes ET71	ET71	ET71	ET71 68 SSukc20 109 WCHN ET225 ET238 ET44 ET42 An878 An109 An421 An502 125 WCHN An199	ET71
Bacteroides caecigallinarum An876	An876	An876	An876	An876
Mediterranea sp ET5	ET5 ET6 ET36	ET5 ET6 ET36	ET5 An768 ET2 68 SSukc20 An189 ET48 ET474 An793 An822 ET44 An878 ET42 An824 An502 ET36 ET6 ET5 An277 ET47 An923 An925 An893 ET15 ET37	ET5 ET36 ET36

Tab. A.2: Výsledky testování I pro bakterie z kmene *Firmicutes* pro metodu 1 s prahem 99% při volbě nástrojů Cd-hit, BLAT a BLAST.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**
Eubacterium sp An3	An3	An3	An3	An3 An179 An180 An232A
Faecalibacterium sp An121	An121 An58	An121 An58	An121 An58	An121 An58 An77
Clostridium spiroforme An149	An149 An158 An26	An149 An158 An26	An149 An158 An26 An779 An15 An173 An105 An134 An142 An80 An13 An731	An149 An158 An26

Tab. A.3: Výsledky testování I pro vstupní bakterie z kmene *Actinobacteria* pro metodu 1 s prahem 99% při volbě nástrojů Cd-hit, BLAT a BLAST.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**
Collinsella sp An7	An7 An840	An7 An840	An7 An840 An5 An2 An307 An340a An70 An788 154 Feed An718 An794 An188 An270 An285 An293 An732 An820 13 COKtk 10 COK 153 Feed 172 Sheadler sukc20 23 COK 26 COK 3 COKtk 51 SSukc10 7 COKtk	An7 An840
Thermophilibacter provencensis 172 Shaedler sukc20	172 Shaedler sukc20 10 COK 23 COK 3 COKtk 7 COKtk An270 13 COKtk	172 Shaedler sukc20 10 COK 23 COK 3 COKtk 7 COKtk An270 13 COKtk	172 Shaedler sukc20 10 COK 23 COK 3 COKtk 7 COKtk 153 Feed 26 COK 51 SSukc10 An188 An270 An285 An293 An732 An733 An820 An794 13 COKtk	172 Shaedler sukc20 10 COK 23 COK 3 COKtk 7 COKtk An270 13 COKtk
Collinsella tanakaei ET32	ET32 An271 15 COKtk 176 SSukc20 53 Shaedler sukc10 An712 An789 An792 An833	ET32 An271 15 COKtk 176 SSukc20 53 Shaedler sukc10 An712 An789 An792 An833	ET32 An271 15 COKtk 176 SSukc20 53 Shaedler sukc10 57 Shaedler sukc10 An712 An789 An792 An833 ET10 ET226 ET30 ET32 An718 13 COKtk	ET32 An271 15 COKtk 176 SSukc20 53 Shaedler sukc10 An712 An789 An792 An833 ET32

Tab. A.4: Výsledky testování I pro vstupní bakterie z kmenů *Proteobacteria* a *Fusobacteria* pro metodu 1 s prahem 99% při volbě nástrojů Cd-hit, BLAT a BLAST.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**
Desulfovibrio sp An276	An276	An276	An276	An276
Escherichia fergusonii ET78	ET78 An786 An190	ET78 An786 An190	ET78 An786 An190	ET78 An786 An190
Parasutterella secunda An562	An562 88 Shaedler sukc20 95 BHI 97 BHI ET72 ET73 ET74 ET75	An562 88 Shaedler sukc20 95 BHI 97 BHI ET72 ET73 ET74 secunda ET75	An562 88 Shaedler sukc20 95 BHI 97 BHI ET72 ET73 ET74 ET75	An562 88 Shaedler sukc20 95 BHI 97 BHI ET72 ET73 ET74 ET75
Fusobacterium mortiferum An397	An397 ET45	An397 ET45	An397 ET45 An425 An814 An874	An397 ET45 An425 An814 An874
Fusobacterium varium An876	An876	An876 An888	An876	
Fusobacterium perfoetens An888	An888	An888 An814 An874 An876	An888	An888

Tab. A.5: Výsledky testování I pro vstupní bakterii *Bacteroides helcogenes* ET71 pro metodu 2 s prahem 99% při volbě všech nástrojů.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**	**FastANI**
Bacteroides helcogenes ET71	ET71	ET71 ET21 An558 ET225 ET474 An793 An801 An279 An767 An426	ET71 ET2 68 SSukc20 An406 ET19 ET22 ET4 84 SSukc20 An825 E238 ET474 An801 ET489 An19 An269 An322 An20 An277 ET11 An818 ET3 143 Shaedl plusK 133 WCHN An905 ET47 ET15 ET37	ET71 68 SSukc20 An406 An496 ET19 ET22 ET4 84 SSukc20 An825 An801 An822 ET489 An19 An269 An322 An51A An62 118 WCHN An421 An502 An772 An20 An277 ET11 ET13 An767 An818 ET3 ET8 133 WCHN An905 ET47 ET15 ET37	ET71

Tab. A.6: Výsledky testování I pro bakterii *Bacteroides caecigallinarum* An428a pro metodu 2 s prahem 99% při volbě všech nástrojů.

Název bakterie	Bacterial Explorer	Cd-hit	BLAT	BLAST	FastANI
Bacteroides caecigallinarum An428a	An428a	An428a 139 WCHN 98 BHI 1 COKtk 68 SSukc20 An406 ET19 ET21 ET22 ET4 ET238 ET71 An822 An269 An62 118 WCHN ET490 46 SSukc10 64 SSukc20 An850 An818 An426 ET15	An428a 139 WCHN ET2 1 COKtk 45 SSukc10 68 SSukc20 An406 An428b An496 ET12 ET19 ET21 ET22 109 WCHN 84 SSukc20 An558 ET238 An279 An51A An819 118 WCHN An20 An39 An45 An42 37 SSukc10 63 SSukc20 An767 An850 An41	An428a 139 WCHN An768 ET2 1 COKtk 45 SSukc10 An406 An428b An496 ET12 ET19 ET21 ET22 84 SSukc20 An558 An825 ET238 ET474 An822 An279 An322 An51A An55 An62 118 WCHN An20 An39 An45 ET490 ET13 An767 An850 143 Shaedl plusK An426 ET80 ET15 ET37	An428a

Tab. A.7: Výsledky testování I pro bakterii *Mediterranea* sp ET5 pro metodu 2 s prahem 99% při volbě všech nástrojů.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**	**FastANI**
Mediterranea sp ET5	ET5 ET36 ET6	ET5 An428b ET21 An819 An824 ET36 ET6	ET5 1 COKtk 45 SSukc10 An428a ET12 121 Egel 84 SSukc20 An558 An825 An822 An279 ET7 An62 An819 An421 ET36 ET6 ET490 63 SSukc20 An767 An850 An818 An16 An426 An475 An905 ET15	ET5 1 COKtk ET12 ET21 An62 ET36 ET6 125 WCHN	ET5 ET36 ET6

Tab. A.8: Výsledky testování I pro bakterii *Eubacterium* sp An3 pro metodu 1 s prahem 99% při volbě všech nástrojů.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**	**FastANI**
Eubacterium sp An3	An3	An3 An915 An81 An179 An177 An210 An11 An135 An503 36 SSukc10 An817 39 SSukc10 70 Shaedler sukc20 An499 71 SSukc20	An3 An201 An172 An250 An75 An174 An175 An883 An915 An249 An46 An81 An179 An149 An158 An779 An12 An210 ET318 An15 An11 An3 An122 An192 77 SSukc20 An248 An10 An135 An4 An82 An9 An91 ET340 An503 An138 An14 An169 An196 An76 36 SSukc10 An427 An785 ET229 An499 An85 An869 ET50 71 SSukc20	An3 An172 An250 An75 An174 An175 An883 An915 An249 An46 An81 An179 An149 An158 An779 An12 An210 ET318 An15 An11 An3 An122 77 SSukc20 An10 An135 An4 An82 An9 An91 ET340 An503 An138 An14 An169 An196 An76 36 SSukc10 An427 An785 ET18 ET229 An499 An507 An85 An869 ET50 71 SSukc20	An3

Tab. A.9: Výsledky testování I pro bakterie *Faecalibacterium* sp An121 a *Clostridium spiroforme* An149 pro metodu 1 s prahem 99% při volbě všech nástrojů.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**	**FastANI**
Faecalibacterium sp An121	An121	An121 An915 An58 An77 An785 ET229 An569	An121 An867 An75 An249 An46 An122 An192 An58 An936 ET51 An194 An411 An860 An901 An85	An121 An867 An75 An174 An816 An249 An46 An192 An58 An936 ET51 An912 An194 ET229 An411 An901 An85	An121
Clostridium spiroforme An149	An149 An158	An149 An250 An46 An158 An26 An142 An80 An731 ET341 ET229	An149 An250 An174 An175 An251 An249 *Blautia* sp An46 An158 An26 An15 An3 An773 An248 An4 ET340 An138 An14 An168 An105 An134 An13 ET341 An499 An559 71 SSukc20	An149 An250 An514 An174 An251 An249 An46 An81 An158 An26 An12 An15 An3 An248 An306 An4 ET340 An138 An168 An105 An134 An142 An80 An13 ET341 71 SSukc20	An149 An158

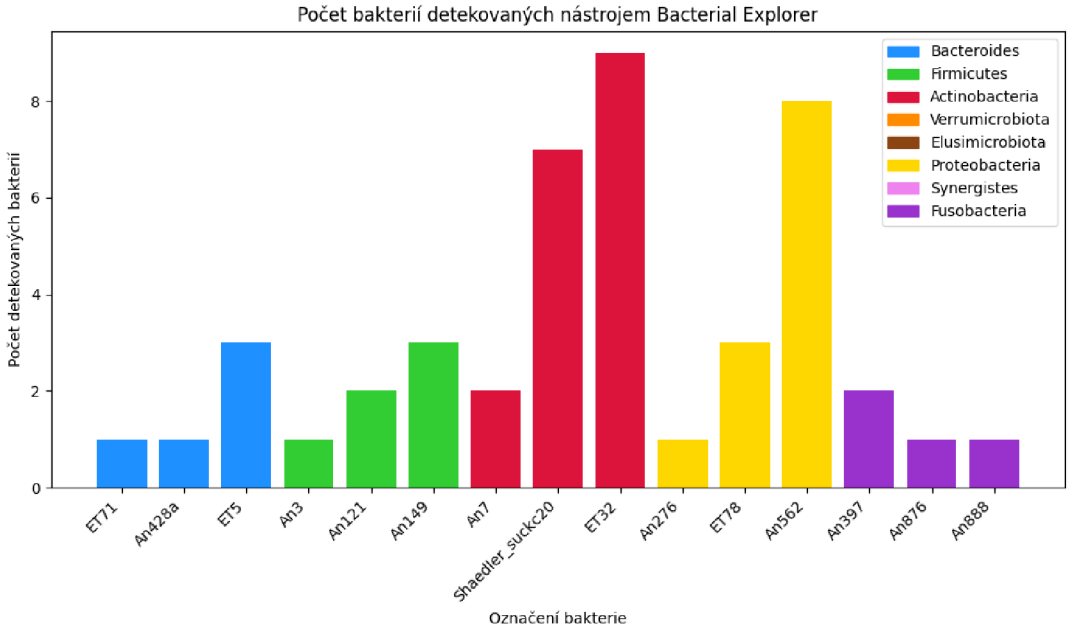
Tab. A.10: Výsledky testování I pro bakterie z kmene *Actinobacteria* pro metodu 2 s prahem 99% při volbě všech nástrojů.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**	**FastANI**
Collinsella sp An7	An7	An7 An719 An840 ET32 An718	An7 An719 An840 An268	An7 An840	An7
Thermophilibacter provencensis 172 Shaedler sukc20	172 Shaedler sukc20	172 Shaedler sukc20	172 Shaedler sukc20 An188 An270 An290 An732 An82 10 COK 153 Feed 51 SSukc10 7 COKtk	172 Shaedler sukc20 An270 10 COK 7 COKtk	172 Shaedler sukc20
Collinsella tanakaei ET32	ET32	ET32 An271 15 CPKtk 176 SSukc20 An368 An712 An789 An792 An833 An273 An718	ET32 An271 15 CPKtk 176 SSukc20 50 SSukc10 53 Shaedler sukc10 57 Shaedler sukc10 An712 An789 An791 An792 An833 ET226 ET32	ET32 An271 15 CPKtk 176 SSukc20 53 Shaedler sukc10 An712 An792 An833	ET32

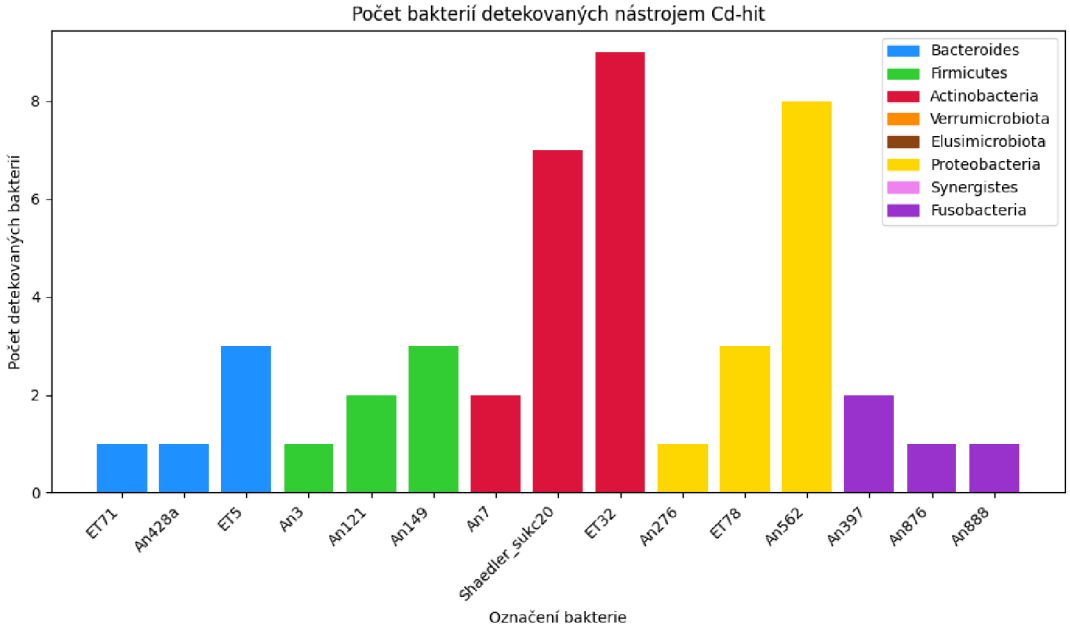
Tab. A.11: Výsledky testování I pro bakterie z kmenů *Proteobacteria* a *Fusobacteria* pro metodu 2 s prahem 99% při volbě všech nástrojů.

Název bakterie	**Bacterial Explorer**	**Cd-hit**	**BLAT**	**BLAST**	**FastANI**
Desulfovibrio sp An276	An 276	An 276	An 276 An401	An276	An276
Escherichia fergusonii ET78	ET78	ET78 An190 An786	ET78 An190 An786	ET78 An190 An786	ET78
Parasutterella secunda An562	An562	An562 88 Shaedler sukc20 95 BHI ET72 ET74 ET75	An562	An562	An562
Fusobacterium mortiferum An397	An397	An397 An425 An814 An874	An397 An425 An814 An874 ET45	An397 An425 An814 ET45	An397
Fusobacterium varium An876	An876	An876	An876 An874 An888	An876 An874 An888	An876
Fusobacterium perfoetens An888	An888	An888 ET46 An425 An874	An888 An874 876	An888 An874 876	An888

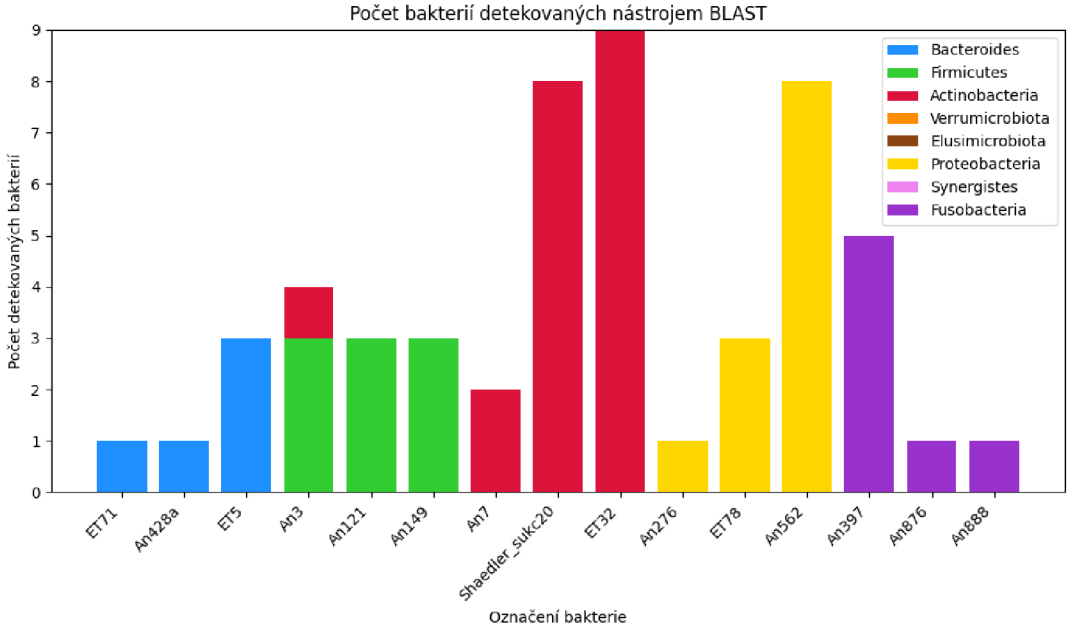
B Grafy



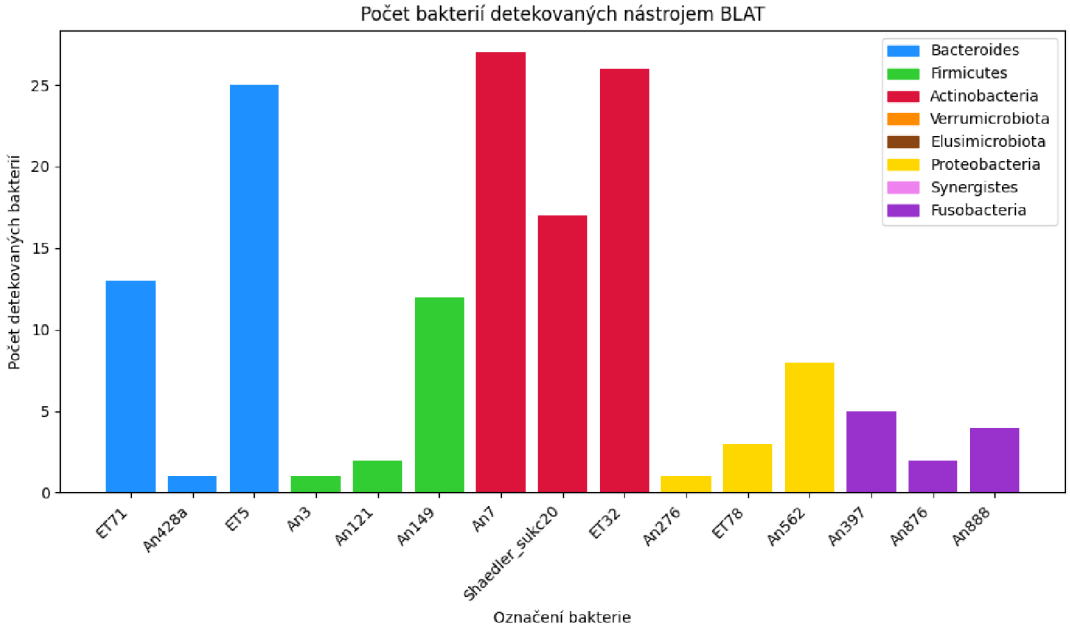
Obr. B.1: Graf s počtem detekovaných bakterií pomocí nástroje Bacterial Explorer. Parametry Bacterial Exploreru: metoda 1, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST.



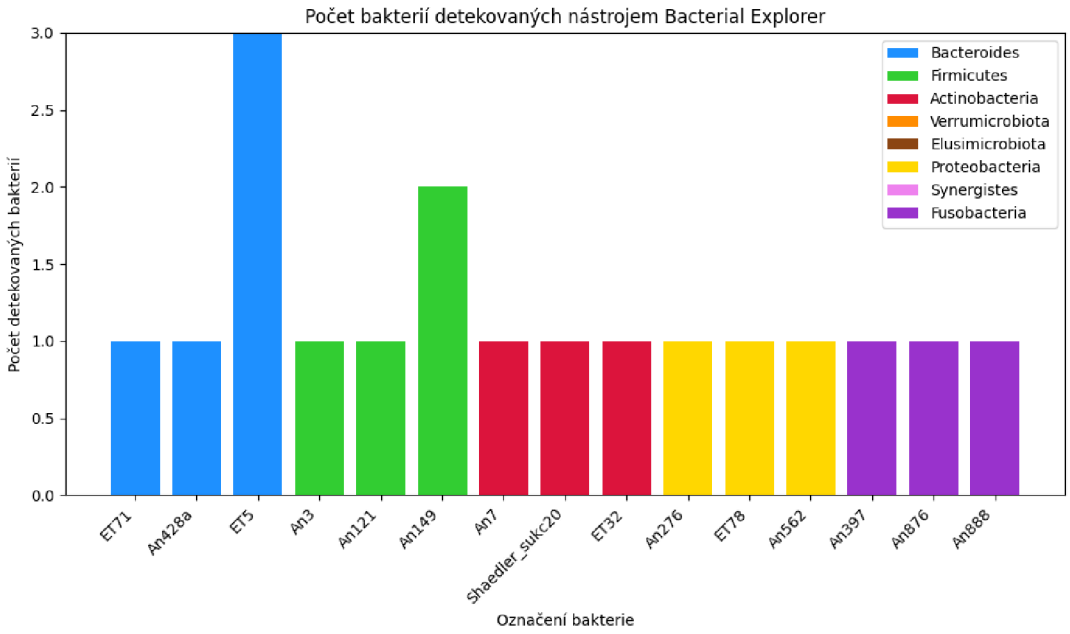
Obr. B.2: Graf s počtem detekovaných bakterií pomocí nástroje Cd-hit v Bacterial Exploreru. Parametry Bacterial Exploreru: metoda 1, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST.



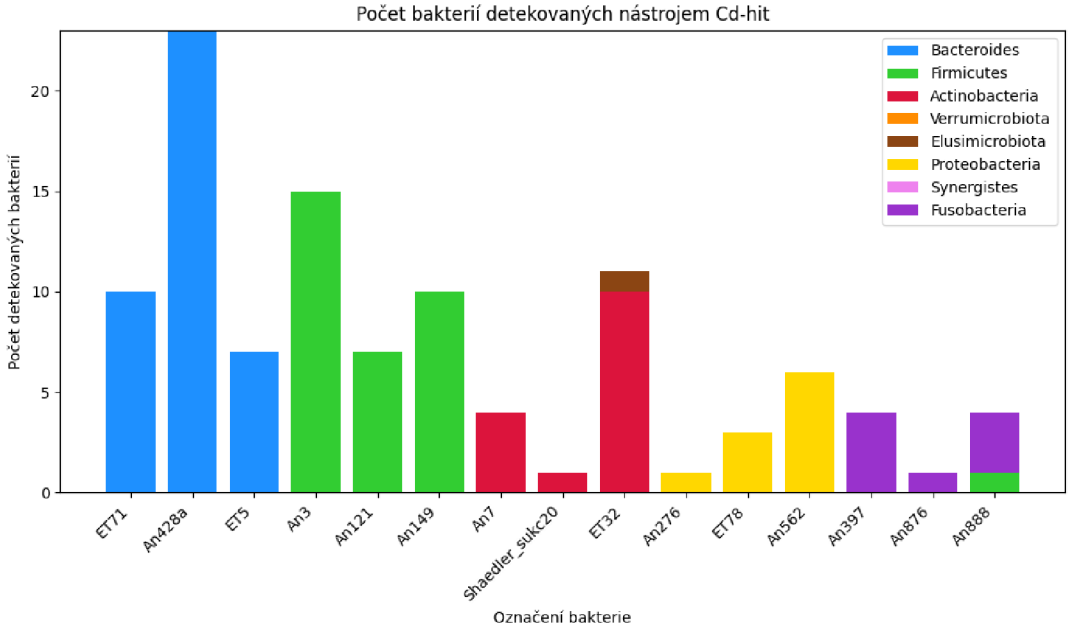
Obr. B.3: Graf s počtem detekovaných bakterií pomocí nástroje BLAST v Bacterial Exploreru. Parametry Bacterial Exploreru: metoda 1, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST.



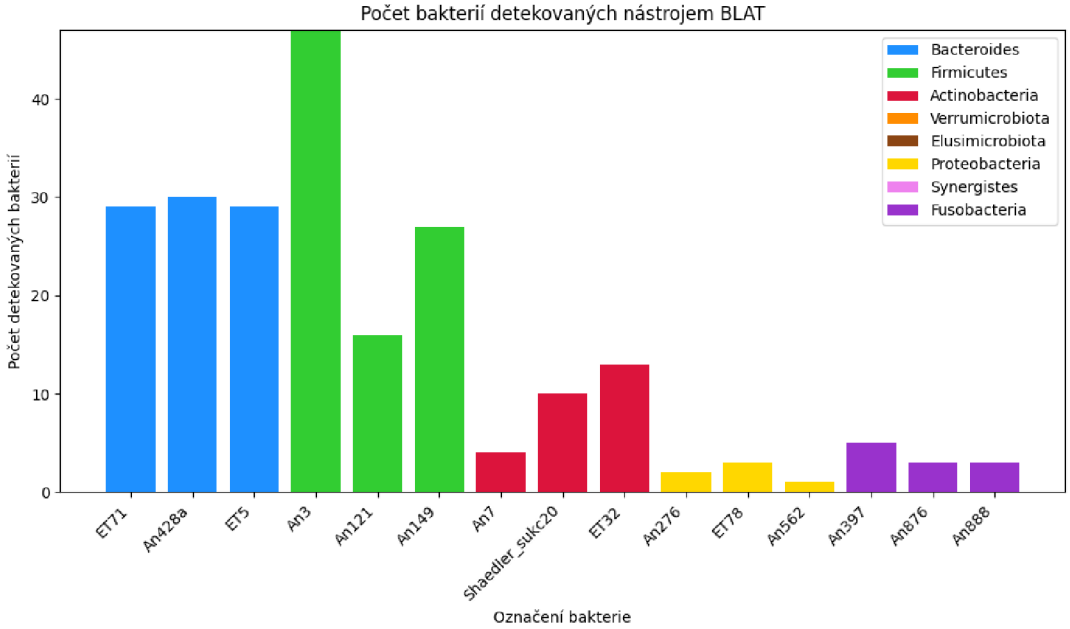
Obr. B.4: Graf s počtem detekovaných bakterií pomocí nástroje BLAT v Bacterial Exploreru. Parametry Bacterial Exploreru: metoda 1, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST.



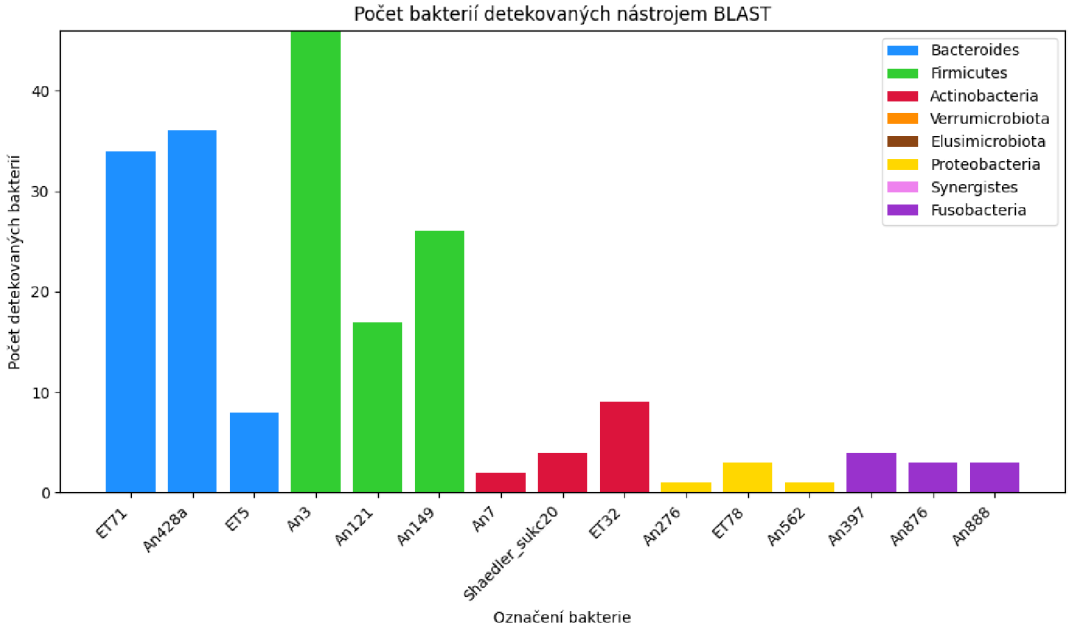
Obr. B.5: Graf s počtem detekovaných bakterií pomocí nástroje Bacterial Explorer. Parametry Bacterial Exploreru: metoda 2, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST, FastANI.



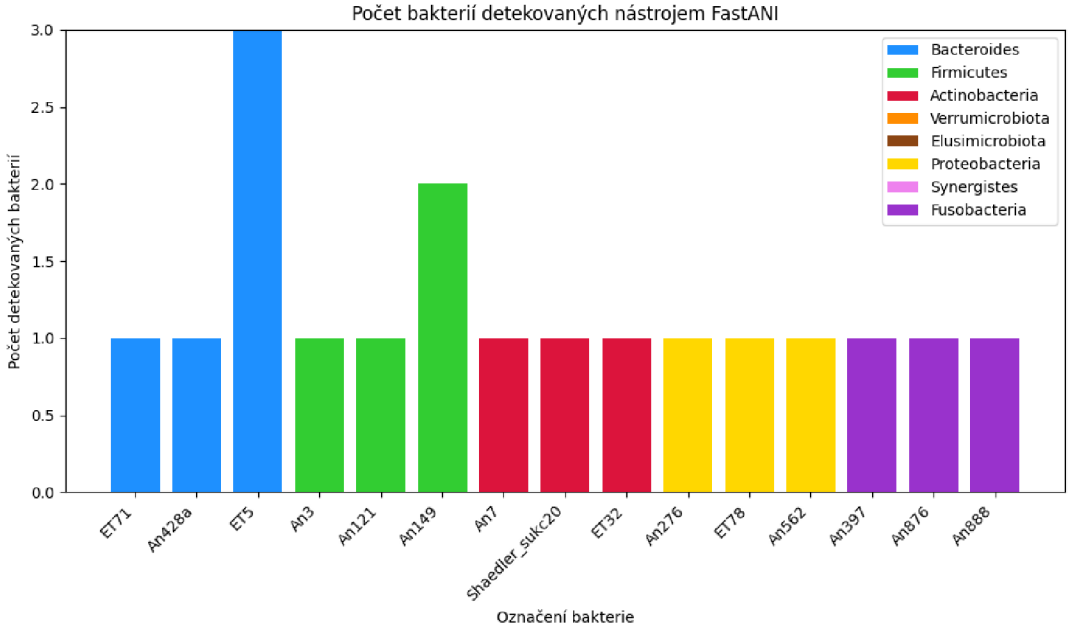
Obr. B.6: Graf s počtem detekovaných bakterií pomocí nástroje Cd-hit v Bacterial Exploreru. Parametry Bacterial Exploreru: metoda 2, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST, FastANI.



Obr. B.7: Graf s počtem detekovaných bakterií pomocí nástroje BLAT v Bacterial Exploreru. Parametry Bacterial Exploreru: metoda 2, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST, FastANI.



Obr. B.8: Graf s počtem detekovaných bakterií pomocí nástroje BLAST v Bacterial Exploreru. Parametry Bacterial Exploreru: metoda 2, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST, FastANI.



Obr. B.9: Graf s počtem detekovaných bakterií pomocí nástroje FastANI v Bacterial Exploreru. Parametry Bacterial Exploreru: metoda 2, práh 99%, nastavené nástroje pro detekci Cd-hit, BLAT, BLAST, FastANI.

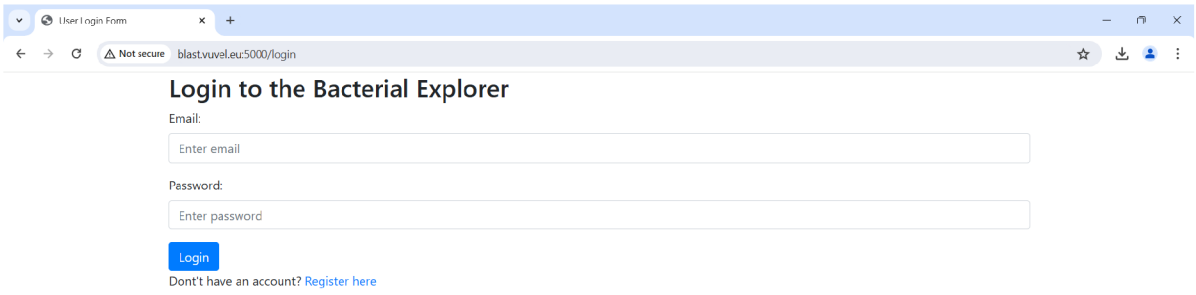
C Manuál nástroje Bacterial Explorer

Toto je manuál k nástroji Bacterial Explorer. Bacterial Explorer je nástroj, který slouží k odhalení nových bakterií. Níže jsou popsány bližší pokyny k obsluze nástroje Bacterial Explorer.

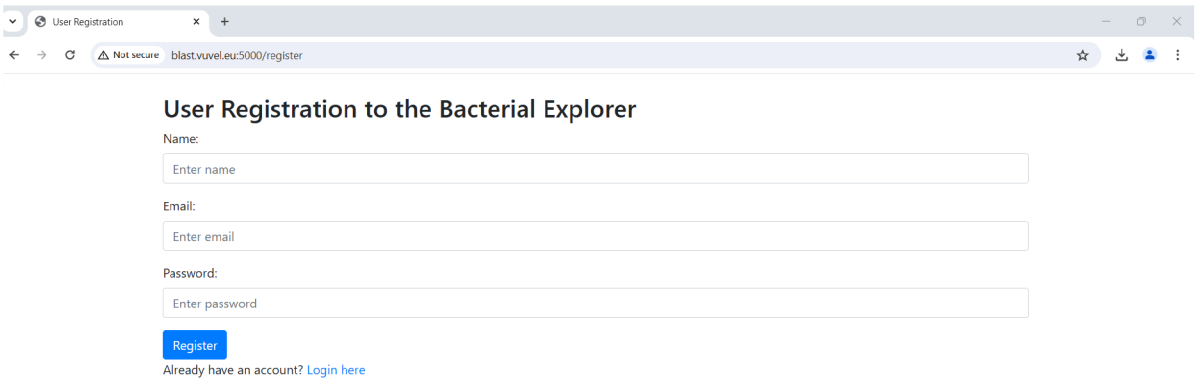
Vítejte v Bacterial Exploreru! Chcete odhalit nové bakteriální genomy? Pak je pro Vás tento nástroj ideální! Bacterial Explorer odhaluje nové bakteriální genomy pomocí dostupných bioinformatických nástrojů – Barrnap, Cd-hit, BLAT, BLAST a FastANI a poskytuje uživateli dva způsoby odhalení nových bakteriálních genomů. První metoda je založená na porovnávání 16S rRNA, druhá metoda je založena na porovnání celých genomů.

1) Přihlášení a registrace

Po kliknutí na odkaz nástroje Bacterial Explorer se zobrazí přihlašovací okno. Přihlašte se do Bacterial Exploreru pomocí mailu a hesla a stiskněte tlačítko “Login”.



Pokud nejste v Bacterial Exploreru registrováni, stiskněte Register here. Následně se zobrazí registrační stránka. Vyplňte všechny údaje a klikněte na tlačítko “Register”.



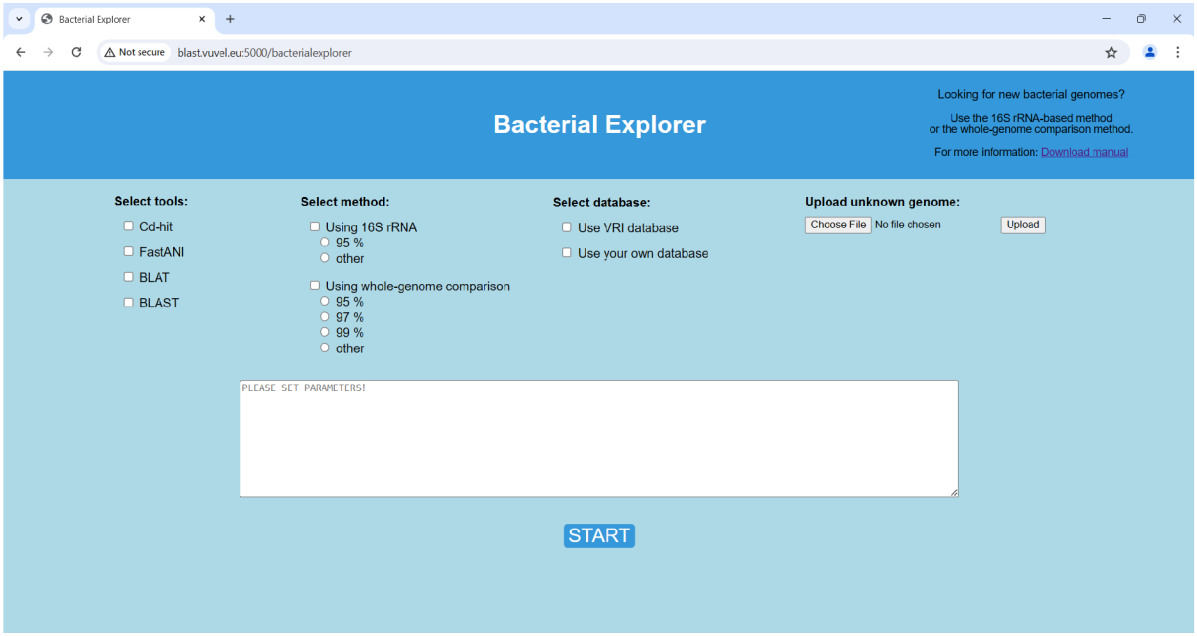
Nyní jste zaregistrováni do nástroje Bacterial Explorer. Klikněte na Login here, což Vás znovu přesměruje na přihlašovací stránku.

2) Captcha

Po úspěšném přihlášení vyplňte Captcha test a klikněte na tlačítko Submit.

2) Bacterial Explorer

Po správném vyplnění Captcha testu budete přesměrováni na hlavní stránku nástroje Bacterial Explorer.



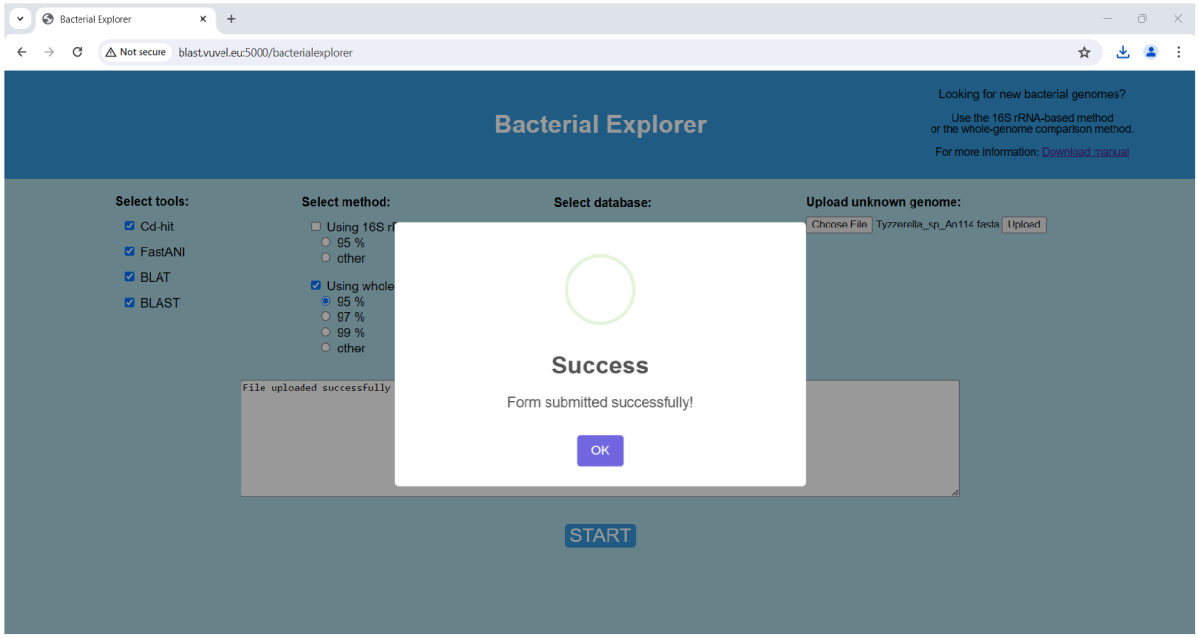
Pro správný běh nástroje je potřeba nastavit parametry. V části „Select tools“ vyberte nástroje, které chcete použít. Nástroj FastANI nevolte, pokud zvolíte v sekci „Select method“ metodu 16S rRNA. Jinak můžete zvolit jeden až všechny nástroje.

V sekci „Select method“ vyberte jednu z metod. Metoda „16S rRNA“ je založená na porovnávání 16S rRNA, metoda „whole-genome comparison“ je založená na porovnávání celých genomů. Taky zvolte požadovaný práh podobnosti. Nejedná se ale o celkovou podobnost porovnávaných bakterií, ale o minimální podobnost vybraných úseků bakterií!!! Pokud zvolíte práh „other“, napište do pole, které se zobrazí, číselnou hodnotu požadované procentuální shody.

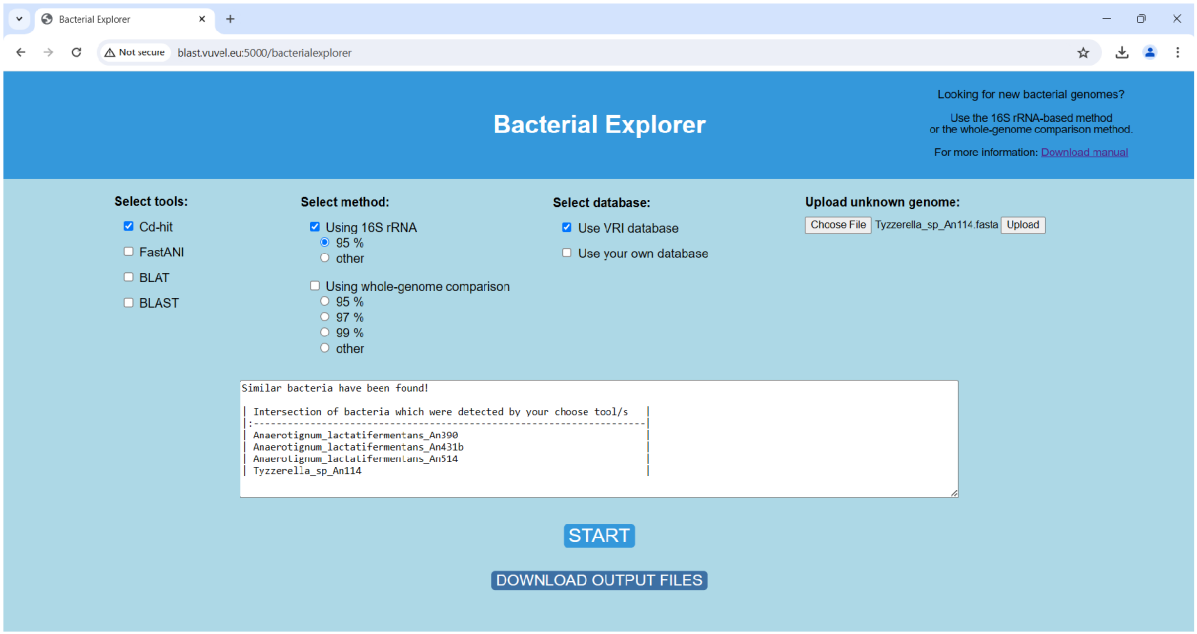
V sekci „Select database“ vyberte databázi, kterou chcete použít. Můžete použít databázi Výzkumného ústavu veterinárního lékařství nebo vlastní databázi. Pokud zvolíte vlastní databázi, nahrajte vaše bakterie ve fasta formátu. Jeden fasta soubor může obsahovat pouze jednu bakterii a celková velikost vlastní databáze je omezená na 30MB! Poté, co máte vybrané soubory z vaší databáze, klikněte na tlačítko Upload, aby se soubory nahrály na server.

V sekci „Upload unknown genome“ nahrajte fasta soubor s bakterií, kterou chcete analyzovat. Opět nezapomeňte zmáčknout tlačítko Upload, aby se soubor nahrál na server!!!

Pokud máte všechny parametry navoleny, stiskněte tlačítko START. Po doběhnutí aplikace se zobrazí informační hláška o úspěšném doběhnutí nástroje a zobrazí se výsledky ve výstupním okně nástroje Bacterial Explorer.



Bakterie zobrazené ve výstupním okně, které jsou detekované jako podobné, jsou průnikem výsledků jednotlivých nástrojů – Cd-hit, FastANI, BLAT a BLAST. Po stisknutí tlačítka DOWNLOAD OUTPUT FILES si můžete stáhnout složku BacterialExplorer.zip ve které jsou uloženy výstupní soubory jednotlivých nástrojů, použitých pro analýzu, ale také souhrnný soubor output.txt, který obsahuje bakterie detekované jednotlivými nástroji jako podobné. Součástí složky je také soubor README.txt, kde jsou popsány výstupy jednotlivých nástrojů.


Bacterial Explorer
Not secure
blast.vuvel.eu:5000/bacterialexplorer
Bacterial Explorer
Looking for new bacterial genomes?
Use the 16S rRNA-based method
or the whole-genome comparison method.
For more information: Download manual
Select tools:
Cd-hit
FastANI
BLAT
BLAST
Select method:
Using 16S rRNA
95 %
other
Using whole-genome comparison
95 %
97 %
99 %
other
Select database:
Use VRI database
Use your own database
Upload unknown genome:
Choose File
Tyzzerella_sp_An114.fasta
Upload
Similar bacteria have been found!
Intersection of bacteria which were detected by your choose tool/s
Anaerotignum_lactatifermentans_An390
Anaerotignum_lactatifermentans_An431b
Anaerotignum_lactatifermentans_An514
Tyzzerella_sp_An114
START
DOWNLOAD OUTPUT FILES

D Obsah elektronické přílohy

Elektronická příloha obsahuje složku BacterialExplorer.zip s výstupy nástroje Bacterial Explorer. Dále obsahuje zdrojový kód nástroje - “pipeline.sh”, manuál nástroje Bacterial Explorer a soubor “vysledky.xlsx”, kde jsou uloženy tabulky s výsledky testování I Bacterial Exploreru.