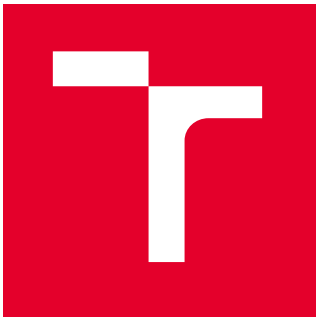
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2024 Martina Kaňoková



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

WEBOVÝ NÁSTROJ PRO RYCHLOU GENOTYPIZACI BAKTERIÁLNÍCH KMENŮ

WEB-BASED TOOL FOR FAST BACTERIAL GENOTYPING

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — Martina Kaňoková

AUTHOR

VEDOUCÍ PRÁCE — Ing. Helena Vítková, Ph.D.

SUPERVISOR

BRNO 2024

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Martina Kaňoková ***ID:*** 240518

Ročník: 3 ***Akademický rok:*** 2023/24

NÁZEV TÉMATU:

Webový nástroj pro rychlou genotypizaci bakteriálních kmenů

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši zaměřenou na problematiku bakteriální genotypizace na základě výskytu specifických segmentů DNA, zejména genů rezistence a virulence. 2) Navrhněte a realizujte v jazyce Python algoritmus detekce zvolených genů rezistence a virulence v bakteriálních genomech. 3) Seznamte se s dostupnými frameworky pro vytváření webových aplikací v jazyce Python a navrhněte uživatelské rozhraní webového nástroje. 4) Vytvořte nástroj pro vyhodnocení profilů rezistence a virulence bakteriálního genomu a proveďte klasifikaci neznámého bakteriálního kmene v rámci lokální databáze bakteriálních profilů. 5) Vytvořený nástroj doplňte o webové rozhraní umožňující vytvořit bakteriální report obsahující genomické profily, zjištěný genotyp v rámci lokální databáze a klasifikaci formou fylogenetického stromu. 6) Použijte vytvořený nástroj pro genotypizaci sekvenčních záznamů poskytnutých z Centra molekulární biologie a genetiky FN Brno. Výsledky diskutujte.

DOPORUČENÁ LITERATURA:

[1] RAMADAN, Asmaa A., 2022. Bacterial typing methods from past to present: A comprehensive overview. Gene Reports. Roč. 29, č. 101675, s. 1-13.

[2] GLOBAL, Emenwa, 2022. Python Flask and Django - Full Stack Python for Web Development: Build Web Applications in Python Using Flask and Django Frameworks. ISBN 979-8355105983.

Termín zadání: 5.2.2024 ***Termín odevzdání:*** 29.5.2024

Vedoucí práce: Ing. Helena Vítková, Ph.D.

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

ABSTRAKT

Tato práce se zabývá tvorbou algoritmu pro detekci zvolených genů rezistence a virulence. První část představuje genotypizační metody, zmiňuje jejich využití spolu s výhodami a nevýhodami. V další části jsou představeny některé z bioinformatických nástrojů, které lze pro genotypizaci použít. Praktická část pak tyto nástroje využívá. Dále je v praktické části představena vlastní webová aplikace umožňující určení sekvenčního typu spolu s detekcí genů virulence a rezistence v bakteriálním genomu.

KLÍČOVÁ SLOVA

typizace bakterií, MLST, genotypizace, webový nástroj pro genotypizaci

ABSTRACT

This thesis is focused on creating algorithm for detection of chosen resistance and virulence genes. First part introduces genotyping methods and mentions their use, pros and cons. In the next part, some of bioinformatics tools that can be used for genotyping are presented. The practical part then uses these tools. Practical part is also focused on creating new web application capable of detecting virulence and resistance genes in bacteria as well as determining the sequence type.

KEYWORDS

bacterial typing, MLST, genotyping, web-based genotyping tool

KAŇOKOVÁ, Martina. *Webový nástroj pro rychlou genotypizaci bakteriálních kmenů*. Bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2024. Vedoucí práce: Ing. Helena Vítková, Ph.D.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Martina Kaňoková
VUT ID autora:	240518
Typ práce:	Bakalářská práce
Akademický rok:	2023/24
Téma závěrečné práce:	Webový nástroj pro rychlou genotypizaci bakteriálních kmenů

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

Rád bych poděkovala vedoucí bakalářské práce paní Ing. Heleně Vítkové, Ph.D. za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Obsah

Seznam obrázků

Seznam tabulek

Seznam výpisů

Úvod

Tato práce se zabývá vytvořením nástroje pro genotypizaci klinicky významných bakterií. V dnešní době již není dostačující určit, které druhy bakterií se ve vzorku nacházejí, ale je vhodné zjistit o jaký bakteriální kmen a sekvenční typ se jedná, případně porovnat podobnost mezi blízce příbuznými bakteriemi z různých vzorků. Rychlá a správná identifikace patogenu je důležitá pro poskytnutí kvalitní péče pacientům a zajištění bezpečnosti nemocničního prostředí v podobě monitorování a prevence nozokomiálních nákaz.

Typizovat bakterie lze pomocí fenotypových metod, u kterých dochází k pozorování jejich morfologie pod mikroskopem a ke kultivaci na různých živých médiích, například pro určení rezistence vůči antibiotikům. Tyto metody však bývají časově náročné a často nedostatečně citlivé. Tyto nevýhody překonává genotypizace.

U genotypizace dochází k identifikaci bakterií na základě jejich genetického materiálu. Historicky se prováděla pouze na základě krátkých úseků DNA, většinou pomocí polymerázové řetězové reakce (PCR). V dnešní době je pro typizaci bakterií možno využít celogenomového sekvenování (WGS). Díky rozšíření sekvenátorů druhé a třetí generace je WGS dostupné i pro menší laboratoře. Nastává však problém s analýzou, jelikož WGS generuje velké množství dat, jenž je potřeba zpracovat, a malé laboratoře k tomu nemusejí mít dostatečné bioinformatické zázemí.

Došlo tak k vytvoření několika webových, volně dostupných nástrojů, které analýzu sekvenčních dat zajišťují. Existují nástroje, které identifikují geny rezistence vůči antibiotikům, další, které identifikují geny virulence a další, například pro určování sekvenčního typu. Nevýhodou je, že každý z nástrojů má pouze jednu funkci a funguje pouze pro vybrané bakterie.

Motivací této práce je vytvořit nástroj, který umožní genotypizaci klinicky významných bakterií a k jeho používání nebude potřeba mít programátorské znalosti. Zároveň umožní identifikaci genů rezistence, virulence a porovnání analyzovaného genomu s lokální databází. Při používání by mohl zefektivnit práci laborantům, a to odbouráním nutnosti obsluhovat několik různých nástrojů a následného manuálního kompletování výsledků.

Cílem teoretické části práce je představit metody, které se pro typizaci bakterií používají a nástroje, které jsou v současné době dostupné.

Praktická část se věnuje návrhu algoritmu detekce genů rezistence a virulence, spolu s určením sekvenčního typu metodou multilokusové sekvenční typizace (MLST), a to za využití nástrojů zmíněných v teoretické části. Tyto algoritmy jsou implementovány ve vytvořené webové aplikaci a doplněny tak, ať lze výsledky porovnat s předchozími provedenými analýzami.

1 Bakteriální genotypizace

Pro pochopení principů a využití bakteriální typizace je vhodné se nejdříve seznámit s tím, jak bakteriální genom vypadá. Genom je soubor veškeré DNA v organismu [1]. Genetická informace bakterií může být uložena v chromozomu a plazmidu, a ve většině případů se jedná pouze o jeden chromozom [2]. Ten je tvořen kruhovou dvoušroubovicí DNA, kde jsou geny uspořádány blízko sebe s minimálním množstvím nekódující DNA. Plazmid není nezbytnou součástí buňky, ale často obsahuje geny, které jsou pro ni výhodné. Jedná se například o geny rezistence či virulence. Velmi podobný pojem jako genom je genotyp, soubor všech genů a jejich variant – alel v organismu. Chybí zde však nekódující DNA. Ne u každého genu musí dojít k jeho expresi, to jaké znaky organismus bude vykazovat závisí i na vnějších podmínkách. Soubor všech znaků organismu se nazývá fenotyp.[1]

I bakterie stejného druhu mohou mít značně odlišný fenotyp, a tím i různý klinický význam. Typizace je tak obzvlášť důležitá pro efektivní léčbu a diagnózu nemocných, kde je podstatná hlavně virulence a rezistence vůči antibiotikům. Virulence je schopnost organismu infikovat hostitele a způsobit onemocnění [3]. Nalezení vztahů mezi jednotlivými bakteriálními liniemi dále pomáhá s monitorováním epidemiologické situace, určením ohniska infekce a s detekcí šíření nozokomiálních infekcí [4].

1.1 Nesekvenační metody typizace

Bakterie můžeme typizovat buď na základě jejich fenotypu, nebo genotypu. Fenotypizace se věnuje hodnocení morfologie a dalších rysů kolonií, analýze antibiogramu nebo sérotypizaci. Výsledky často závisí na použitém médiu a životních podmínkách bakterií, například teplotě, pH a velikosti kolonie. Toto může vést k problémům s reproduktibilitou a zároveň jsou tyto metody časově náročné, jelikož je nutné čekat na kultivaci bakterií. U sérotypizace je dalším problémem nutnost přítomnosti specifických markerů. [4], [5]

Tyto nedokonalosti vedly k vývoji genotypizačních metod, které určují bakteriální typ pomocí analýzy genomu na základě několika různých přístupů. Nejstarší z nich využívá restrikčního štěpení DNA a nejpoužívanější metodou tohoto typu je pulzní gelová elektroforéza (PFGE). Další skupina metod využívá polymerázovou řetězovou reakci (PCR), a jako zástupce bude uvedena repetitivní PCR (rep-PCR). Nejmodernější přístup typizuje bakterie na základě identifikací polymorfismů v jejich genomu. [6]

1.1.1 Pulzní gelová elektroforéza (PFGE)

U metody PFGE je DNA kultivovaného izolátu rozštěpena pomocí restrikčních enzymů na velké fragmenty, které jsou následně separovány v gelové matrici pomocí pulzní elektroforézy. Ta se od konvenční elektroforézy liší použitím dvou elektrických polí odlišného směru, která se střídavě aktivují. Toto umožňuje separaci velkých fragmentů, které by jinak zůstaly příliš blízko u sebe.[5] Tato metoda je stále populární a jedná se o zlatý standard pro typizaci některých klinicky významných bakterií a to hlavně z důvodu její vysoké diskriminační síly. Zároveň se jedná o relativně levnou metodu, mezi jejíž nevýhody však patří časová náročnost, nutnost kultivace a použití vysoce kvalitní DNA, což může být obtížné při používání lidských vzorků. [7]

1.1.2 Repetitivní PCR (rep-PCR)

Metoda rep-PCR využívá primery, které hybridizují s opakujícími se úseky DNA v genomu. Za pomoci PCR jsou jednotlivé úseky DNA mezi primery zmnoženy a vznikají tak fragmenty různých velikostí podle délky vybraných úseků. Ty se rozdělí dle velikosti pomocí elektroforézy a následně dojde k analýze vzniklého elektroforeogramu, který je specifický pro každý druh bakterie. Při použití kapilární elektroforézy jde proces zautomatizovat, čehož využívají komerční kity určené pro konkrétní druhy bakterií. Mezi výhody této metody patří vysoká rychlost, jednoduchost získání výsledků, a také to, že stačí menší množství DNA. Mezi nevýhody patří nízká reproduktibilita způsobená variabilitou použitých reagencí a nastavením gelové elektroforézy. [5], [6]

1.2 Metody založené na sekvenaci

Problémy s reproduktibilitou a následným porovnáváním výsledků mezi jednotlivými laboratořemi překonává typizace bakterií na základě sekvenace jejich DNA [7] [6].

Jelikož sekvenování genomu poskytuje velké množství dat, umožňuje metodám na něm založených mít mnohem větší diskriminační sílu než mají metody zmíněné výše. Toho se dá využít při sledování šíření nozokomiálních infekcí, detekci ohnisek nákazy při epidemii nebo při určování zdroje kontaminace v potravinářství. Zároveň odpadá nutnost bakterie kultivovat, čímž jsou tyto metody vhodné i pro bakterie pomalu rostoucí nebo jinak obtížně kultivovatelné. Dále jsou tyto metody vhodné pro predikci fenotypových znaků, jako je například rezistence vůči antibiotikům, virulence nebo sérotyp. [13]

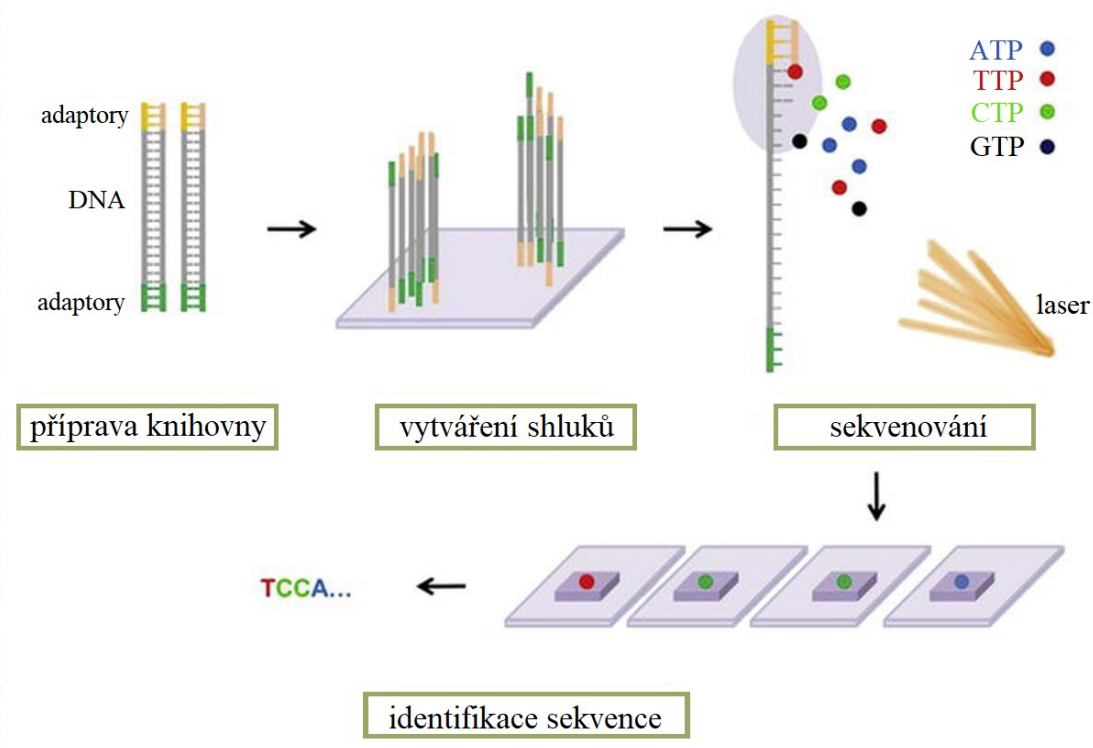
Jednou z nevýhod je nízká standardizace postupů, která limituje vzájemné porovnávání výsledků. Pro některé bakterie však již existují spravované databáze, což umožňuje i mezinárodní spolupráci. Další z nevýhod je stále ještě násobně vyšší cena v porovnání s PFGE nebo rep-PCR. Cena závisí i na tom, jaká platforma je k sekvenování využita. Další charakteristiky platforem rovněž ovlivňují, pro kterou typizační metodu je vhodné výsledky z nich použít. Z tohoto důvodu následující podkapitola často používané sekvenační platformy představí. [13]

1.2.1 Sekvenační platformy

Pro potřeby sekvenace nukleových kyselin bylo vyvinuto několik platforem, z nichž však budou zmíněny pouze sekvenátory Illumina a Oxford Nanopore, a to z důvodu použití dat získaných právě z nich.

Sekvenátory značky Illumina se řadí mezi sekvenování nové generace (NGS), které je charakteristické masivní paralelizací, jenž vede k větší rychlosti a nižší ceně sekvenování. Na obrázku 1.1 je uveden stručný postup sekvenování, jehož prvním krokem je příprava knihovny. Dochází k fragmentaci DNA, kdy jednotlivé fragmenty mají velikost 150–600 bp. Následuje ligace adaptorů a denaturace dsDNA na ssDNA. Jednovláknová DNA pak nasedne na destičku a pomocí můstkové PCR dojde k vytvoření shluků. Na toto navazuje denaturace dsDNA na ssDNA a přidání primerů. [8], [9]

Samotné sekvenování probíhá tak, že destička je promývána fluorescenčně značenými nukleotidy s terminátory, které zabraňují navázání několika nukleotidů najednou. Pomocí laseru jsou nukleotidy excitovány a fluorescence je detekována kamerou. Podle barvy pak lze určit jaký nukleotid se na dané místo navázal. Poté dojde k odstranění terminátorů a cyklus se opakuje. Dalším krokem je odstranění nasyntetizovaného vlákna a opakování čtení v druhém směru. [8], [9]

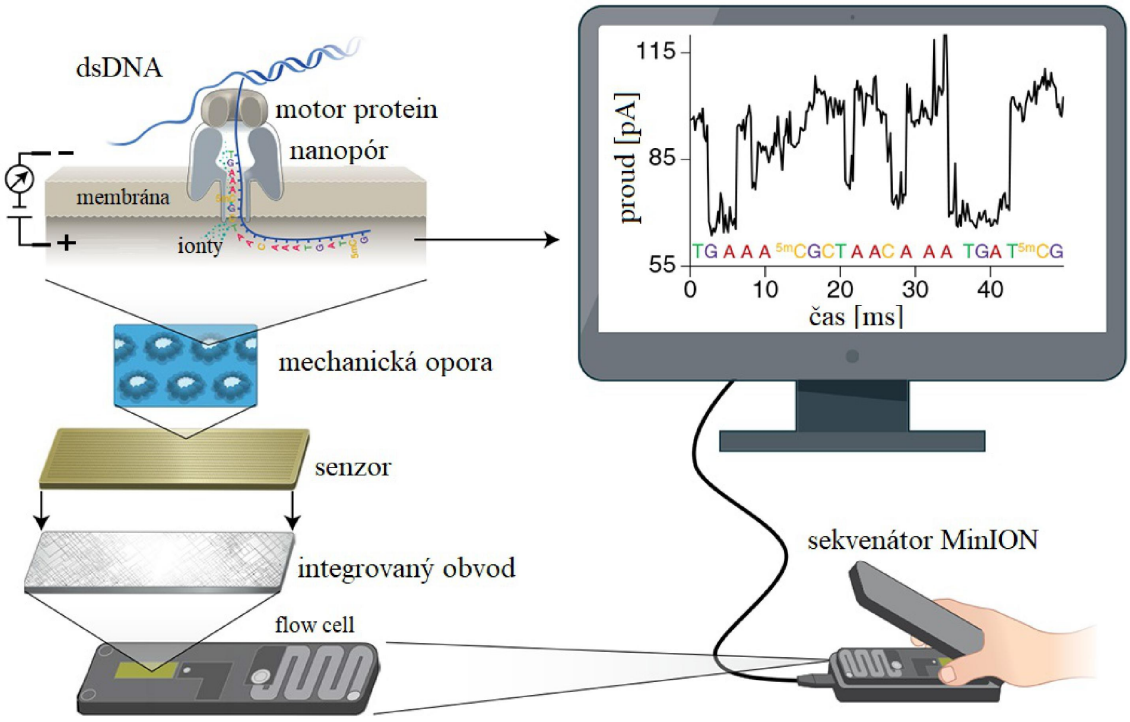


Obr. 1.1: Schéma Illumina sekvenování, převzato z [10]

Sekvenování třetí generace, mezi které patří i platforma Oxford Nanopore, se od NGS liší hlavně tím, že není třeba DNA amplifikovat a jsou možné dlouhé délky čtení, což usnadňuje následné zpracování dat. [11]

Nanopórové sekvenování je založeno na průchodu nukleotidů skrz nanopór v membráně a měření změn elektrického proudu, schéma tohoto postupu je na obrázku 1.2. Na polymerovou membránu obsahující nanopóry z proteinů je přivedeno napětí, díky kterému dochází k prostupu DNA skrz póry. Rychlost tohoto prostupu kontrolují motorové proteiny navázané na sekvenovanou dsDNA. Ty ji zároveň denaturují na ssDNA. Nukleotidy právě procházející nanopórem vedou k charakteristickým změnám v iontovém proudu a tyto změny jsou pomocí počítačových algoritmů dekódovány. [11]

Nevýhoda sekvenátorů třetí generace je ve větší chybovosti a vyšší ceně. Díky svým dlouhým čtením, však umožňují analýzu repetitivních sekvencí nebo určení, zda se geny rezistence nacházejí na plasmidu nebo chromozomu. Přesnost výsledného složeného genomu lze zvýšit namapováním krátkých čtení ze sekvenátoru Illumina. [12]



Obr. 1.2: Schéma Oxford Nanopore sekvenování, převzato z [11]

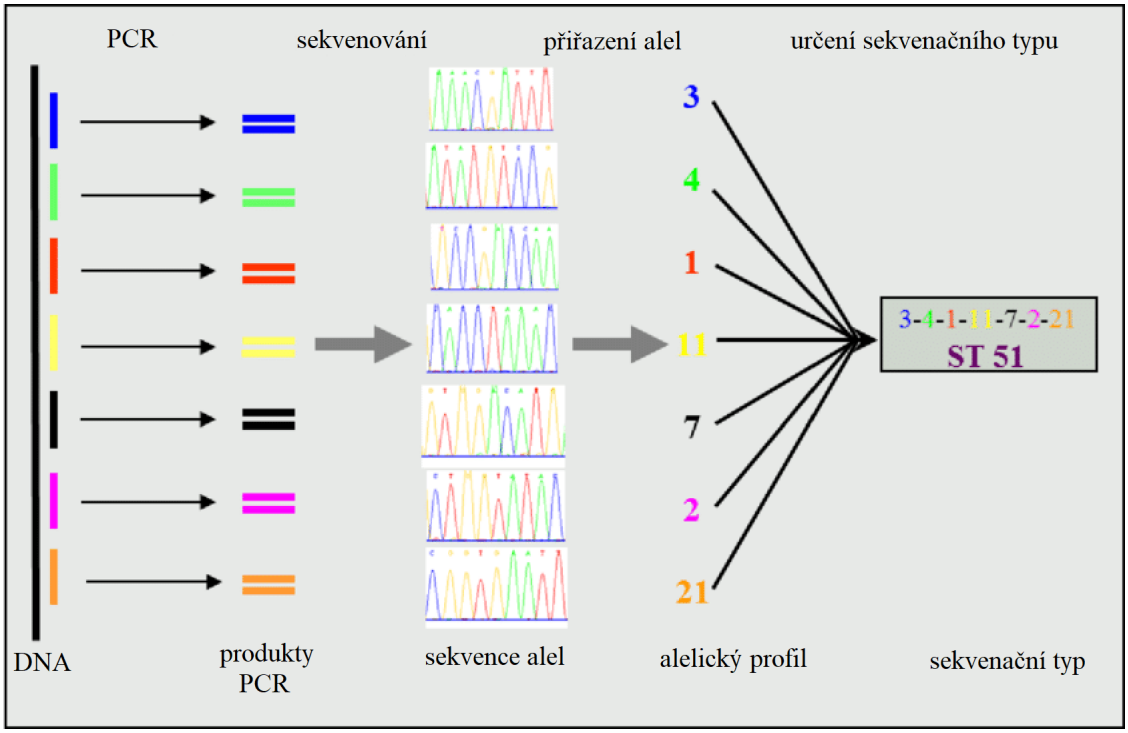
1.2.2 Jednolokusová sekvenční typizace

Jednolokusová sekvenční typizace (SLST), jak už z názvu vyplývá, analyzuje pouze jeden lokus, proto je vhodné vybrat k analýze geny, které se vyznačují dostatečnou variabilitou, zároveň jsou však přítomny u všech identifikovaných vzorků. Jeden z často používaných je gen pro 16S ribozomální RNA (16S rRNA), který je přítomný ve všech bakteriích. Tento gen se tak využívá k pochopení jejich fylogeneze a pro taxonomické účely. [5]

1.2.3 Multilokusová sekvenční typizace

Multilokusová sekvenční typizace (MLST) využívá k sekvenaci několik z udržovacích genů (z anglického housekeeping genes), obvykle sedm [7]. Tyto geny jsou nezbytné pro udržování základních funkcí a pro přežití bakterie. Zajišťují například replikaci nebo opravu DNA. Jejich mutace bývají většinou neutrální, jelikož nepodléhají selekčnímu tlaku tolik jako ostatní geny [6]. Prvním krokem je odebrání vzorku obsahující bakterii, která se bude identifikovat. Jednotlivé kroky MLST shrnuje obrázek 1.3. Pro MLST není potřeba získat vysoce kvalitní DNA ani pracovat s živými bakteriemi, což zvyšuje bezpečnost laboratorních pracovníků a usnadňuje získání použitelného vzorku, například z mozkomíšního moku [14]. Vybrané geny, obvykle podle již zavedeného MLST schématu, jsou pomocí PCR amplifikovány a následuje

jejich sekvenace. Získaná data se porovnají s databází, ve které mají jednotlivé alely přidělená čísla. Toto se provede pro všechny zvolené geny a podle výsledné číselné kombinace se určí sekvenční typ (ST) bakterie. [15], [14],[5]



Obr. 1.3: Schéma metody MLST, převzato z [15]

1.2.4 Core genome a whole genome MLST

Metoda cgMLST (z anglického core genome multilocus sequence typing) je principiálně podobná MLST, rozšiřuje však množství genů k sekvenaci. Využívá geny, které se vyskytují u většiny bakteriálních kmenů – core genome. Výhodou tohoto přístupu je větší diskriminační síla oproti MLST a možnost fylogenetické analýzy. Nevýhodou je nepřítomnost jedné centrální databáze s MLST schématy, což vede ke snížení reproduktibility výsledků. [13],[5]

Metoda, která k analýze využívá celogenomového sekvenování, se nazývá wg-MLST (z anglického whole genome MLST). Dá se očekávat, že tato metoda předčí svou diskriminační silou cgMLST, což platí, pokud je genom osekvenován a jednotlivá čtení složena s dostatečnou kvalitou. Metoda tak vyžaduje kvalitní referenční genom, a proto je z hlediska bioinformatického zpracování a znalostí náročnější.[13]

1.2.5 Detekce bodových mutací

Celogenomového sekvenování (WGS) využívají i další metody, tentokrát založené na detekci bodových mutací (SNP). Ty nacházejí využití hlavně při rozlišování blízce příbuzných bakteriálních kmenů. Jednou z možností je porovnávání sekvenovaných vzorků s referenčním blízce příbuzným genomem. Tato technika se nazývá mapování a umožňuje detekovat SNP, podle jejich počtu a typu lze určit příbuznost jednotlivých vzorků. Tento přístup je vhodný, pokud je k dispozici malé množství vzorků. Nevýhodou je omezená reprodukovatelnost, obzvlášť pro bakterie, u kterých nejsou jednoznačně určeny referenční genomy. Nutnost referenčního genomu obchází metoda, která identifikuje SNP na základě vzájemného porovnávání k-merů (sekvencí nukleotidů o délce k). Porovnávají se k-mery dvou různých sekvencí mezi sebou, a to pro všechny osekvenované vzorky. Tento přístup je vhodný, pokud se očekává velká podobnost vzorků.[13], [5]

1.2.6 Mini-MLST

Poslední zmíněnou metodou je mini-MLST, která poskytuje rychlejší a levnější alternativu MLST a je méně výpočetně náročná než ostatní výše zmíněné metody. Kompromisem je její nižší diskriminační síla.

Stejně jako u MLST je zde vybráno několik genů, které jsou pomocí PCR amplifikovány, dalším krokem ale není jejich sekvenace, nýbrž vysokorozlišovací analýza křivek tání (HRM). Při HRM je obarvená DNA zahřívána, což vede k jejímu rozpadu a snížení měřené fluorescence. Výsledná křivka tání koreluje s obsahem cytosinu, guaninu a s délkou daného úseku DNA. Toto umožňuje detekci SNP a následné určení melt typu. Tato metoda je vhodná pro preventivní monitorování epidemiologické situace v nemocnicích nebo pro rychlou genotypizaci v případě podezření na šíření nákazy. [16]

2 Výpočetní nástroje pro genotypizaci

Genotypizační metody založené na sekvenaci vyžadují po uživatelích značné bioinformatické znalosti a alespoň základy programování. To je jedna z věcí, která limituje jejich rozsáhlé využívání. Zároveň i znalí uživatelé profitují z jednoduchých a hlavně optimalizovaných a udržovaných nástrojů pro genotypizaci. Tato kapitola přiblíží několik z nich, představí princip jejich fungování, poukáže na to, zda mají webovou aplikaci a zda je možné je integrovat do vlastního kódu.

2.1 Basic local alignment search tool (BLAST)

BLAST je nástroj provozovaný Národním centrem pro biotechnologické informace (NCBI – National Center for Biotechnology Information), dostupný z `https://blast.ncbi.nlm.nih.gov/Blast.cgi`. Jedná se o nástroj k hledání homologních sekvencí v databázi. BLAST se dělí na několik typů podle toho, jaká sekvence a v jaké databázi se hledá. Jejich přehled se nachází v tabulce 2.1. [17]

Tab. 2.1: Typy BLAST

typ BLAST	**hledaná sekvence**	**databáze**
blastn	nukleotidy	nukleotidy
blastp	protein	proteiny
blastx	nukleotidy	proteiny
tblastn	protein	nukleotidy
tblastx	translatované nukleotidy	translatované nukleotidy

Vzhledem k tomu, že se tato práce věnuje hledání genů (nukleotidová sekvence) v sekvenčních datech (nukleotidové sekvence), bude dále popisován a ukázán blastn, jehož princip je však podobný dalším typům BLAST. Zároveň je nutné zdůraznit, že se nejedná o nástroj přímo určený pro genotypizaci, lze ho však pro hledání genů, alel, v genomu úspěšně využít.

2.1.1 Princip algoritmu

K popisu fungování BLAST je využito několika pojmů a parametrů, které jsou uvedeny spolu s vysvětlením v následujícím seznamu:

- maximal segment pair (MSP) - úsek dvou sekvencí s nejvyšším možným skóre zarovnání pro dané sekvence;
- high scoring pair (HSP) - část dvou sekvencí s vysokým skóre zarovnání;
- words - části sekvence o délce w;

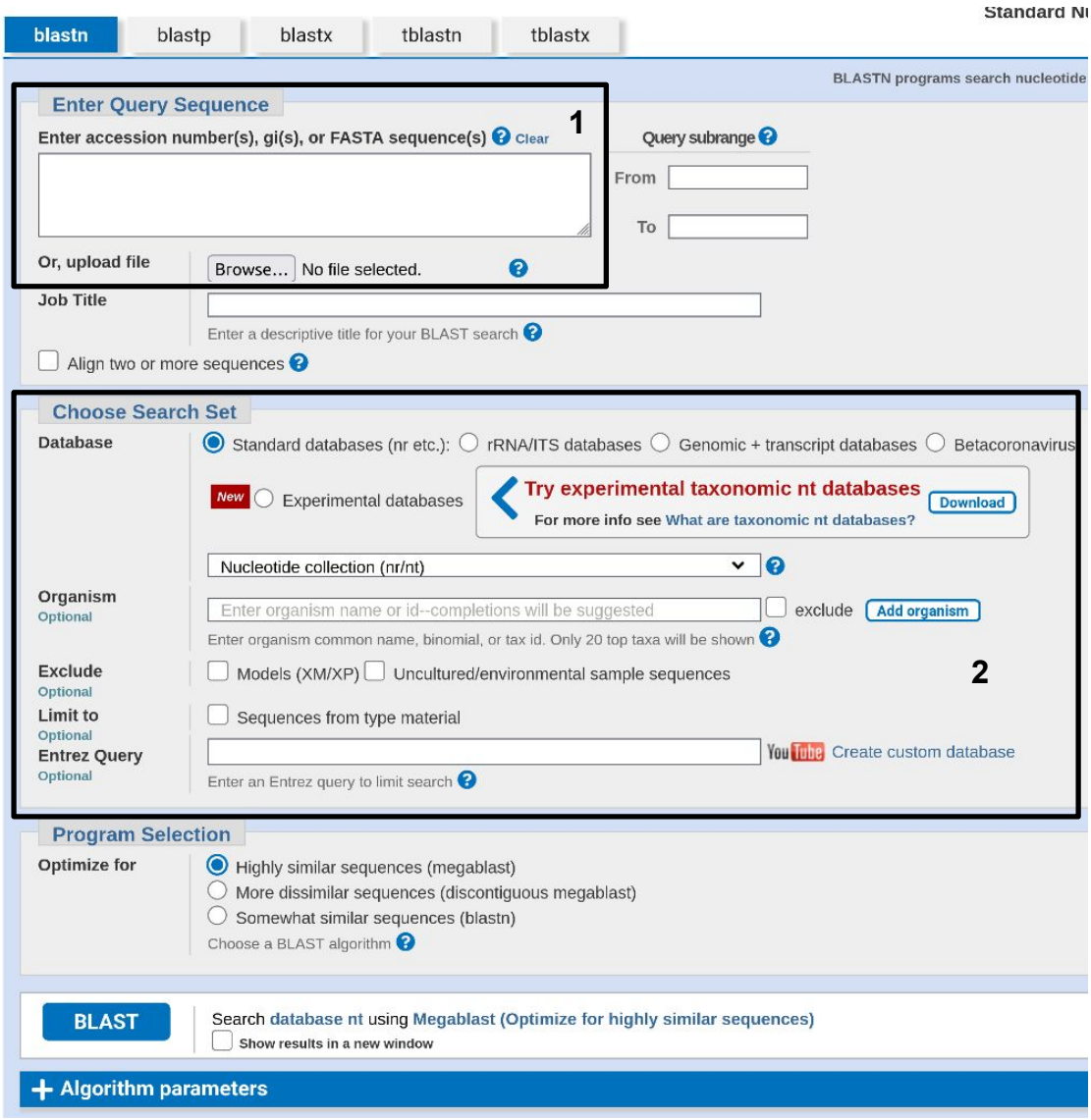
- T, S - minimální skóre zarovnání;
- X - hodnota zastavující rozšiřování zarovnání.
- bit score - normalizované skóre zarovnání, tak ať je porovnatelné mezi různými vyhledáváními;
- e value - představuje počet různých zarovnání o stejné nebo lepší shodě náhodně se vyskytujících v databázi;
- percent identity - procentuální shoda daného zarovnání;
- query cover - procentuální část hledané sekvence použitá v zarovnání.[18]

Nejjednodušší algoritmus vypadá tak, že se vytvoří words, která mají s hledanou sekvencí skóre zarovnání větší než T. Dojde k mapováno words vůči databázi za použití algoritmu rychlého vyhledávání, např. hashovací tabulky. Jednotlivá zarovnávání jsou prodlužována, dokud nedojde k poklesu skóre o X. Výsledkem jsou všechny segmenty, které mají skóre alespoň S. MSP je zarovnání s nejvyšším skóre. Jednotlivá skóre jsou počítána pomocí substituční matice.[19]

Tento základní algoritmus je dále modifikován, tak aby bylo dosaženo optimálních výsledků. Jednou z modifikací je například prodlužování pouze těch částí, ve kterých bylo nalezeno více words vedle sebe. HSP jsou tak získány rychleji. Další varianta umožňuje brát v potaz při rozšiřování zarovnání i inzerci nebo deleci. [18]

2.1.2 Webové rozhraní BLAST

Hledanou sekvenci lze zadat několika různými způsoby, zadáním identifikačního čísla sekvence (accession number, gi) z NCBI, napsáním sekvence ve formátu fasta nebo přímo nahráním souboru v tomto nebo genbank formátu. Toto zadávání probíhá ve fialově vyznačeném poli 1 na obrázku 2.1, což je screenshot úvodní stránky nástroje blastn. [17]



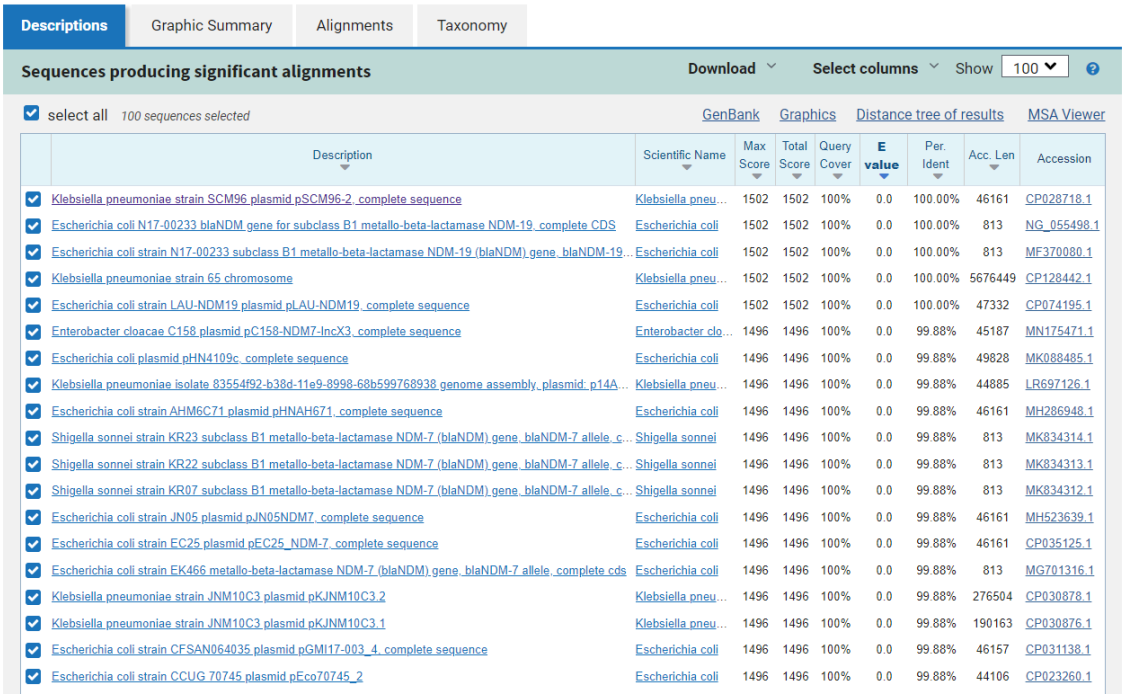
Obr. 2.1: Úvodní stránka blastn, dostupné z [20]

Vyznačené pole 2 v tomtéž obrázku značí část, ve které se dá specifikovat to, v jaké databázi se budou hledat podobné sekvence. V základním nastavení se jedná o neredundantní nukleotidové databáze NCBI, v rozbalovacím seznamu je možnost specifikovat, které konkrétní databáze mají být použity – např. databáze referenčních sekvencí. Při zadání organismu v další části lze prohledávat pouze sekvence k němu patřící. Stejně tak lze organismus z vyhledávání zcela vyřadit. [17]

Mezi další parametry, které jdou u blastn zvolit, patří například skórovací parametry, maximální počet zobrazovaných výsledků nebo minimální velikost e-value. [17]

Výsledkem blastn je seznam zarovnaných sekvencí rozšířený o několik parametrů. Jak je vidět na obrázku 2.2, patří mezi ně již výše zmíněná e-value, dále

percent identity a query cover. Podle všech zobrazených parametrů lze výsledky také filtrovat a všechna zarovnání jsou rovněž k dispozici ke stažení. [17]



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Select columns — Show 100

select all *100 sequences selected* — GenBank | Graphics | Distance tree of results | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Klebsiella pneumoniae strain SCM96 plasmid pSCM96-2, complete sequence	Klebsiella pneu...	1502	1502	100%	0.0	100.00%	46161	CP028718.1
Escherichia coli N17-00233 blaNDM gene for subclass B1 metallo-beta-lactamase NDM-19, complete CDS	Escherichia coli	1502	1502	100%	0.0	100.00%	813	NG_055498.1
Escherichia coli strain N17-00233 subclass B1 metallo-beta-lactamase NDM-19 (blaNDM) gene, blaNDM-19...	Escherichia coli	1502	1502	100%	0.0	100.00%	813	MF370080.1
Klebsiella pneumoniae strain 65 chromosome	Klebsiella pneu...	1502	1502	100%	0.0	100.00%	5676449	CP128442.1
Escherichia coli strain LAU-NDM19 plasmid pLAU-NDM19, complete sequence	Escherichia coli	1502	1502	100%	0.0	100.00%	47332	CP074195.1
Enterobacter cloacae C158 plasmid pC158-NDM7-IncX3, complete sequence	Enterobacter clo...	1496	1496	100%	0.0	99.88%	45187	MN175471.1
Escherichia coli plasmid pHN4109c, complete sequence	Escherichia coli	1496	1496	100%	0.0	99.88%	49828	MK088485.1
Klebsiella pneumoniae isolate 83554f92-b38d-11e9-8998-68b599768938 genome assembly, plasmid: p14A...	Klebsiella pneu...	1496	1496	100%	0.0	99.88%	44885	LR697126.1
Escherichia coli strain AHM6C71 plasmid pHNAH671, complete sequence	Escherichia coli	1496	1496	100%	0.0	99.88%	46161	MH286948.1
Shigella sonnei strain KR23 subclass B1 metallo-beta-lactamase NDM-7 (blaNDM) gene, blaNDM-7 allele, c...	Shigella sonnei	1496	1496	100%	0.0	99.88%	813	MK834314.1
Shigella sonnei strain KR22 subclass B1 metallo-beta-lactamase NDM-7 (blaNDM) gene, blaNDM-7 allele, c...	Shigella sonnei	1496	1496	100%	0.0	99.88%	813	MK834313.1
Shigella sonnei strain KR07 subclass B1 metallo-beta-lactamase NDM-7 (blaNDM) gene, blaNDM-7 allele, c...	Shigella sonnei	1496	1496	100%	0.0	99.88%	813	MK834312.1
Escherichia coli strain JN05 plasmid pJN05NDM7, complete sequence	Escherichia coli	1496	1496	100%	0.0	99.88%	46161	MH523639.1
Escherichia coli strain EC25 plasmid pEC25_NDM-7, complete sequence	Escherichia coli	1496	1496	100%	0.0	99.88%	46161	CP035125.1
Escherichia coli strain EK466 metallo-beta-lactamase NDM-7 (blaNDM) gene, blaNDM-7 allele, complete cds	Escherichia coli	1496	1496	100%	0.0	99.88%	813	MG701316.1
Klebsiella pneumoniae strain JNM10C3 plasmid pKJNM10C3.2	Klebsiella pneu...	1496	1496	100%	0.0	99.88%	276504	CP030878.1
Klebsiella pneumoniae strain JNM10C3 plasmid pKJNM10C3.1	Klebsiella pneu...	1496	1496	100%	0.0	99.88%	190163	CP030876.1
Escherichia coli strain CFSAN064035 plasmid pGMI17-003_4, complete sequence	Escherichia coli	1496	1496	100%	0.0	99.88%	46157	CP031138.1
Escherichia coli strain CCUG 70745 plasmid pEco70745_2	Escherichia coli	1496	1496	100%	0.0	99.88%	44106	CP023260.1

Obr. 2.2: Stránka s výsledky blastn

BLAST je k dispozici nejen jako webový nástroj, lze ho také stáhnout a používat lokálně. Stažený umožňuje rychlý a efektivní přístup pomocí příkazového řádku nebo integraci do svého programu. Hlavní výhodou je však možnost vytvoření vlastní databáze, ve které se podobné sekvence hledají. Tato funkce umožňuje BLAST využít i pro genotypizaci, kdy se ze sekvenačních dat vytvoří databáze a v ní jsou hledány známé geny, případně jejich alely. [17]

2.2 ResFinder

ResFinder je nástroj provozovaný Centrem pro genomickou epidemiologii (CGE – Center for Genomic Epidemiology), dostupný z `http://genepi.food.dtu.dk/resfinder`, a určený pro detekci genů antimikrobiální rezistence (AMR geny). Byl vyvinut za účelem zpřístupnění analýzy sekvenačních dat i méně zkušeným uživatelům a laboratorním pracovníkům. Uživatelské rozhraní webu je proto jednoduché a přehledné. Zároveň je k dispozici verze ke stažení pro používání pomocí příkazového řádku. Je určena pro operační systém Linux. [21]

2.2.1 Princip nástroje

Nástroj se skládá ze čtyř databází viz tabulka 2.2. Databází je myšlena strukturovaná sbírka vzájemně souvisejících dat, často v textovém formátu. Jedná se o databázi:

- ResFinder, která obsahuje získané AMR geny.
- PointFinder obsahující mutace chromozamálních genů způsobujících AMR.
- pro určení fenotypu na základě genotypu.
- panelů pro in silico druhově specifické antibiogramy.

Tyto databáze lze stáhnout samostatně a použít k vlastní analýze. [21]

Tab. 2.2: Přehled databází nástroje ResFinder

databáze	**bakteriální databáze**
ResFinder-AMR geny	nezávislé
PointFinder – chromozomální mutace	*Campylobacter* *Enterococcus faecalis, faecium* *Escherichia coli* *Helicobacter pylori* *Klebsiella* *Mycobacterium tuberculosis* *Neisseria gonorrhoeae* *Plasmodium falciparum* *Salmonella* *Staphylococcus aureus*
fenotyp z genotypu	nezávislé
panely pro in silico antibiogramy	*Campylobacter* *Campylobacter jejuni, coli* *Enterococcus faecalis, faecium* *Escherichia coli* *Mycobacterium tuberculosis* *Salmonella* *Staphylococcus aureus*

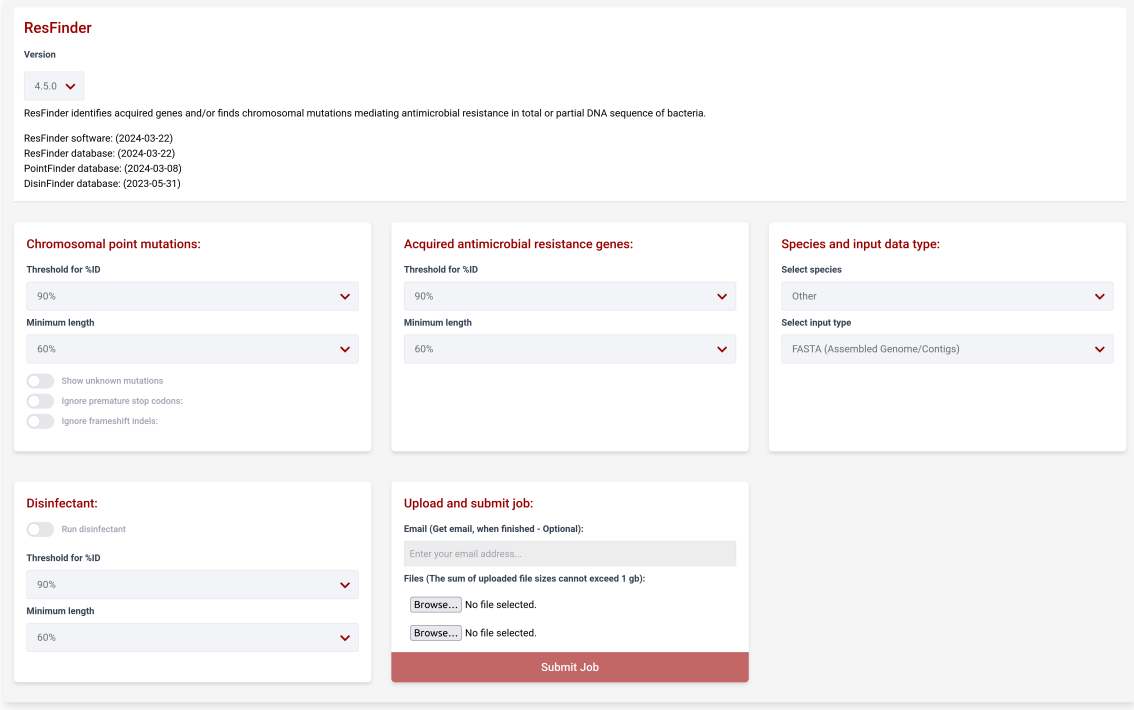
Nástroj ResFinder pracuje se sestaveným genomem ve formátu fasta nebo s jednotlivými čteními ve formátu fastq. Hledají se nejlepší shody v databázích s nahranými daty. Pro tato porovnávání je v případě sestaveného genomu využit BLAST a pro nesestavená data je využito metody zarovnání k-mer (KMA, k-mer alignment). KMA je metoda vyvinuta pro mapování čtení vůči redundantní databázi [22], jako je například databáze ResFinder. [21]

Následně je z těchto nalezených genů a mutací předpovězen fenotyp. Zde je nutné brát v potaz, že ResFinder a jemu podobné nástroje odhalí rezistenci pouze v případě, že je již znám gen nebo mutace, které ji způsobuje a jsou zavedeny v databázi. Nedojde tak k odhalení nově vzniklých mechanismů rezistence nebo rezistence závislé na samotné stavbě bakterie, jako je např. chybějící struktura, na kterou dané antibiotikum cílí nebo tlustá buněčná stěna. [21]

2.2.2 Webové rozhraní ResFinder

Na obrázku 2.3 ukazujícím hlavní stránku webové verze ResFinderu, se nachází několik volitelných parametrů, které určují, co bude považováno za shodu při hledání genů v databázi. Konkrétně se jedná o parametry Treshold for %ID a Minimum length, které odpovídají, ve stejném pořadí, parametrům percent identity a query cover, zmíněných v předchozí kapitole o nástroji BLAST.

Žádný z těchto parametrů nelze nastavit libovolně, je nutno vybrat přednastavenou hodnotu z rozbalovacího seznamu. Stejně tak je nutno specifikovat, zda je nahráván sestavený genom nebo jednotlivá čtení, která se ještě rozlišují na čtení nanopórová a nenanopórová. Posledním z povinných parametrů je výběr druhu bakterie.

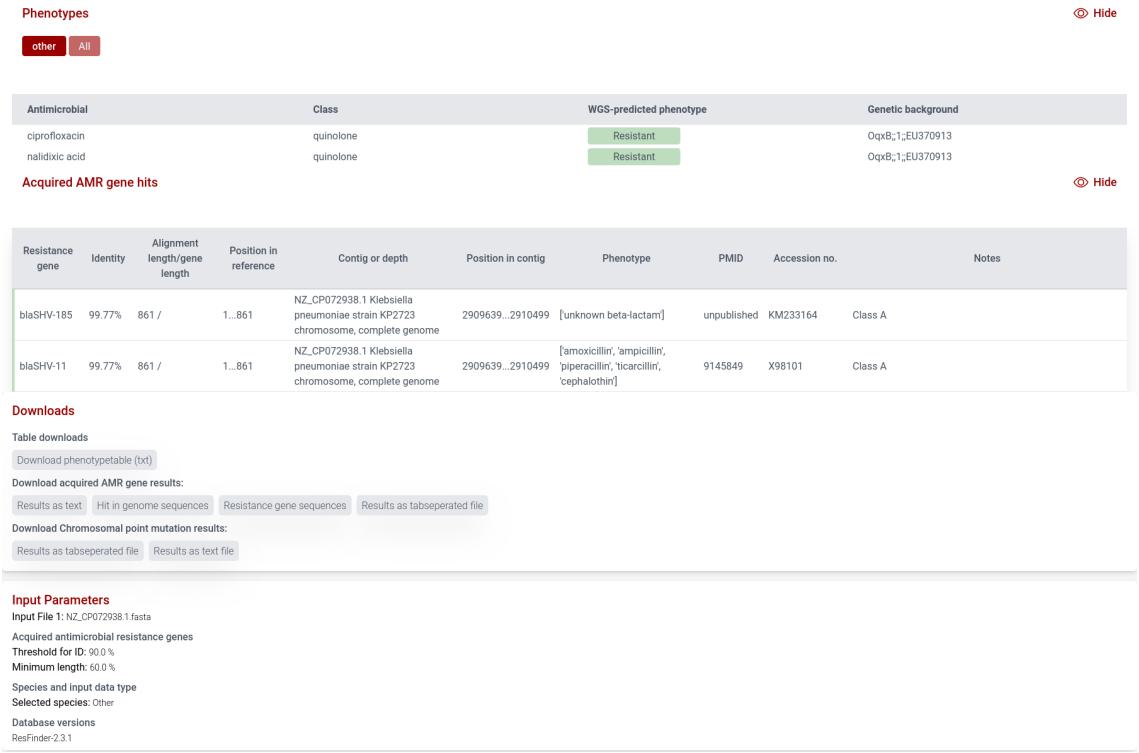


Obr. 2.3: Hlavní stránka nástroje ResFinder, dostupné z [23]

Výsledky jsou zobrazeny jako webová stránka, obrázek 2.4, ve kterém jsou jednotlivé části pro přehlednost zkráceny. Zároveň je však možnost výsledky stáhnout

ve formě textových souborů, které jsou výsledkem i v případě použití stažené verze ResFinderu.

V první části obrázku 2.4 lze vidět predikci fenotypu bakterie. Jedná se o seznam antibiotik spolu s jejich třídou a graficky vyznačenou odpovídající rezistencí. Je zmíněn i gen, na jehož základě byla rezistence predikována. Informace o něm jsou více rozepsány v další části, kde se lze dozvědět, s jakou procentuální shodou byl nalezen (identity) a jaká je délka zarovnání (alignment length). Mezi další informace patří, jakou fenotypovou rezistenci predikuje a pokud je k dispozici, tak číslo článku, který se tomu věnuje. Výsledky chromozomálních mutací jsou prezentovány obdobně.



Obr. 2.4: Stránka s výsledky nástroje ResFinder

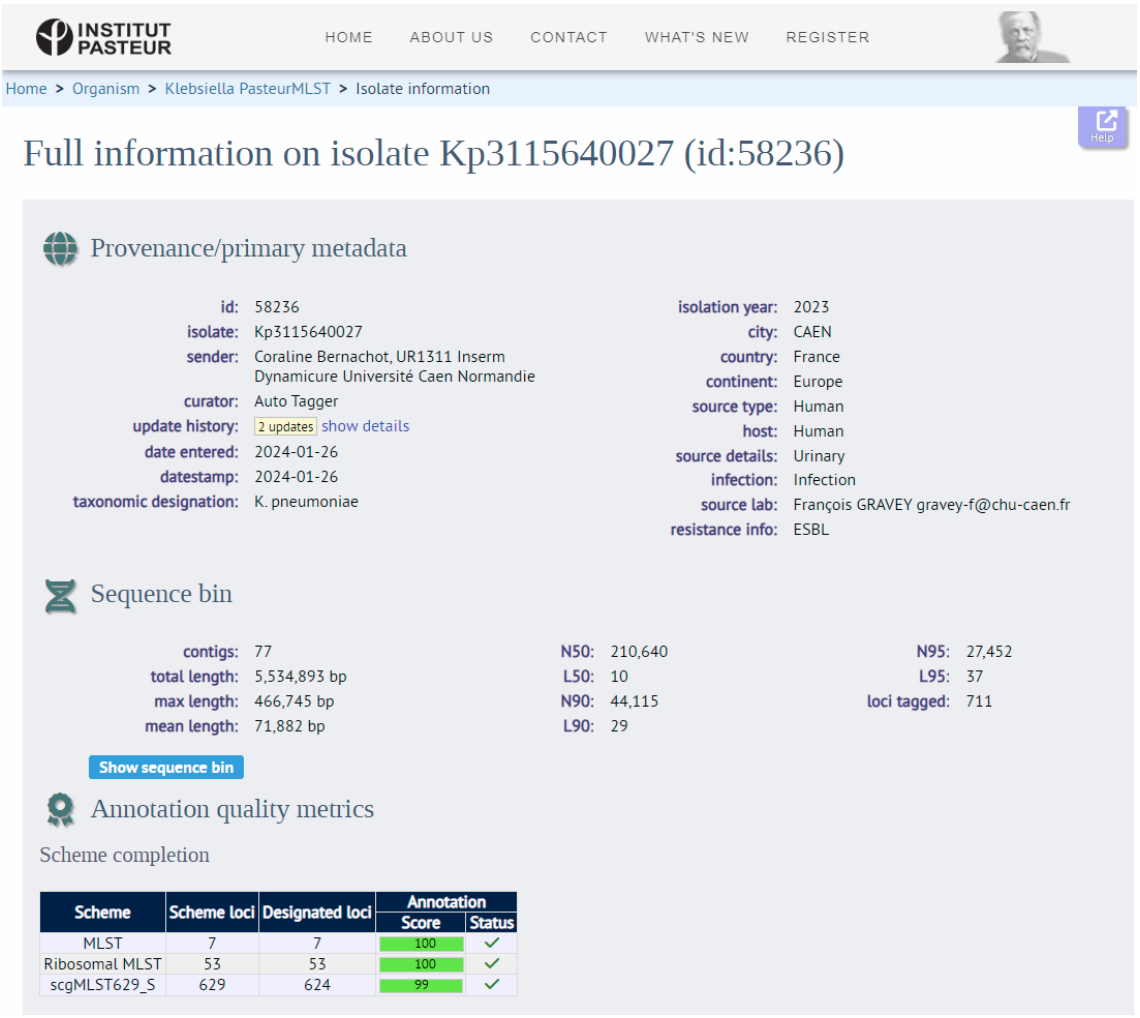
2.3 BIGSdb-Pasteur

BIGSdb (Bacterial Isolate Genome Sequence database) je software vytvořen Oxfordskou univerzitou pro analýzu a ukládání sekvenačních dat z bakteriálních izolátů. BIGSdbPasteur je webová platforma využívající tohoto softwaru, dostupná z https://bigsdb.pasteur.fr/ a provozuje ji Institut Pasteur.

Nachází se na ní velké množství soukromých i veřejných databází genomových sekvencí založených na MLST, cgMLST a dalších doplňkových schématech. Momentálně je zde veřejně dostupných 13 aktivních databází pro jednotlivé bakterie.

Vzhledem k jejímu použití dále v této práci bude blíže představena databáze pro bakterii *Klebsiella Pneumoniae*, která má dvě hlavní části. První obsahuje jednotlivé nahrané genomy spolu s informacemi o daném izolátu a druhá sekvence existujících alel s typizačními schématy.

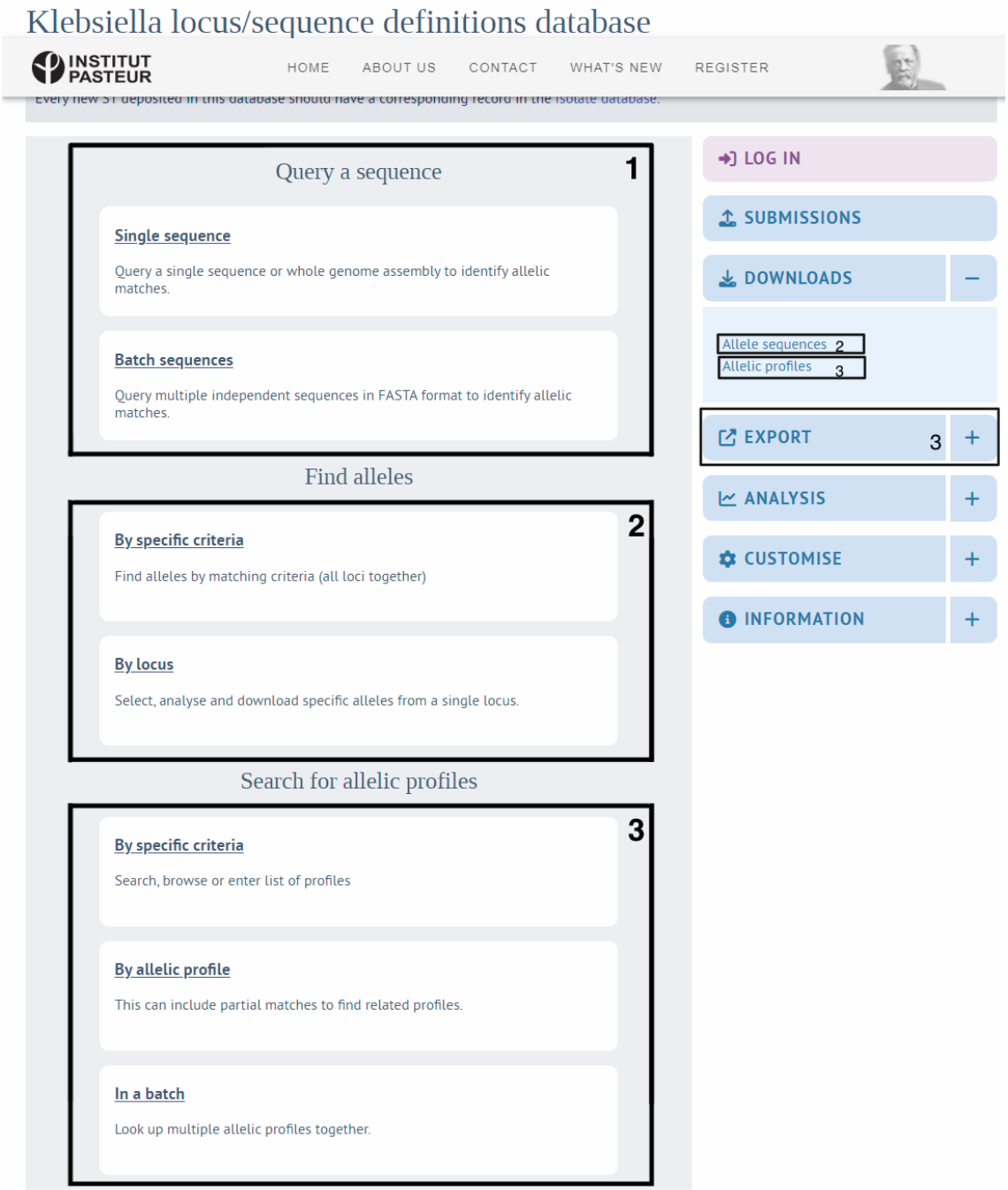
Databáze izolátů zahrnuje v současnosti více než 44 000 záznamů [24], obsahujících složené sekvence a data informující o kvalitě daného sekvenování a sestavení genomu. Krom toho záznamy obsahují také metadata týkající se původu izolátu, z jaké laboratoře pochází a podobně. Obrázek 2.5 ukazuje kompletní stránku jednoho záznamu s uváděnými informacemi. V celé databázi jde pomocí těchto polí vyhledávat a filtrovat.



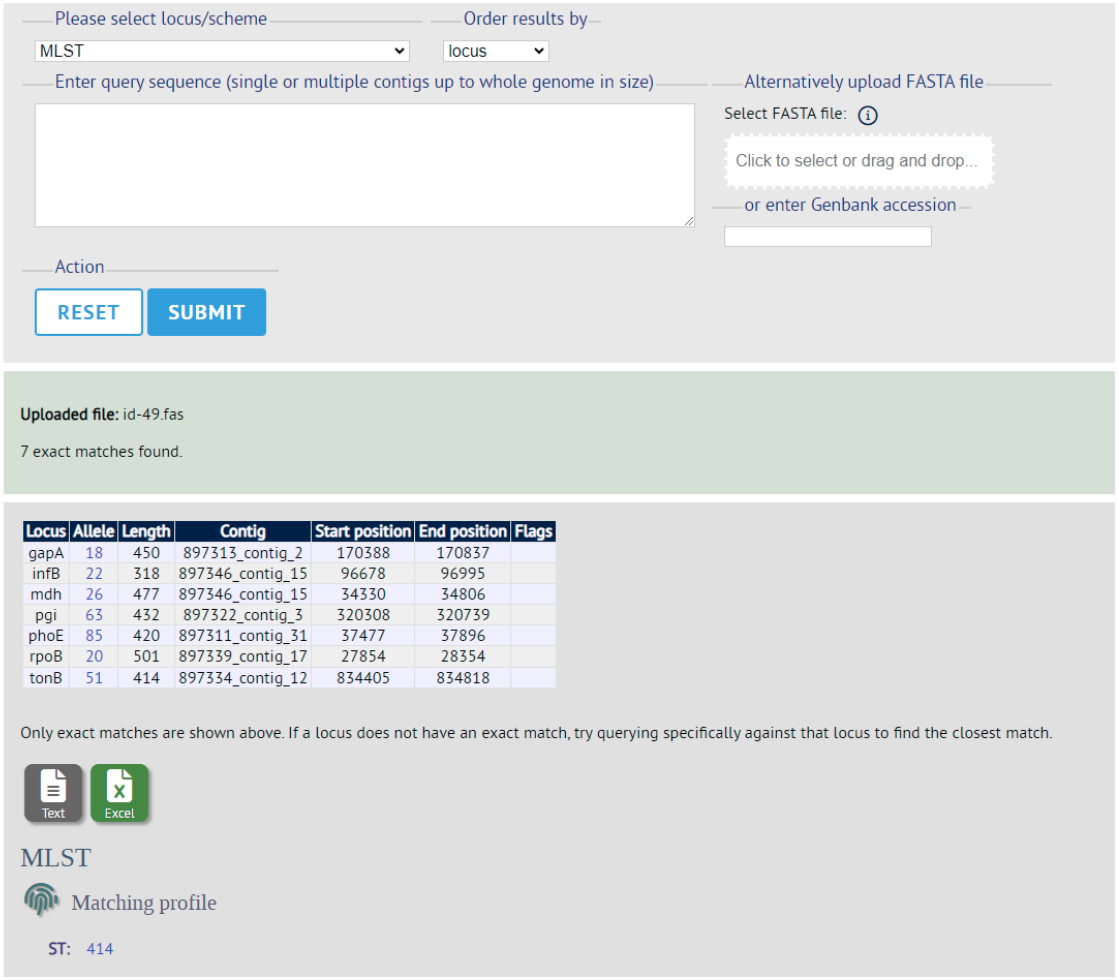
Obr. 2.5: Ukázka záznamu izolátu z BIGSdbPasteur, dostupné z [24]

Databázi alel, jejíž domovská strana je ukázána na obrázku 2.6, lze dále rozdělit na několik částí podle jejich funkce, kdy první důležitou funkcí, pole 1 obrázku 2.6, je možnost vyhledání alel přítomných ve zvolené sekvenci a případně určení sekvenčního typu. Je možnost zadat celý složený genom, jednotlivé kontigy, anebo

hromadně vyhledávat pro několik nezávislých sekvencí. Nezávislé sekvence lze zadat pro vyhledávání pouze ve formátu fasta, zatímco při jednom vzorku lze zadat i identifikační číslo sekvence z Genbank. Vyhledávání přítomných alel opět v pozadí využívá BLAST algoritmu. Způsob zadávání spolu s výsledky lze vidět na obrázku 2.7. Výstupy jsou zároveň k dispozici ke stažení.

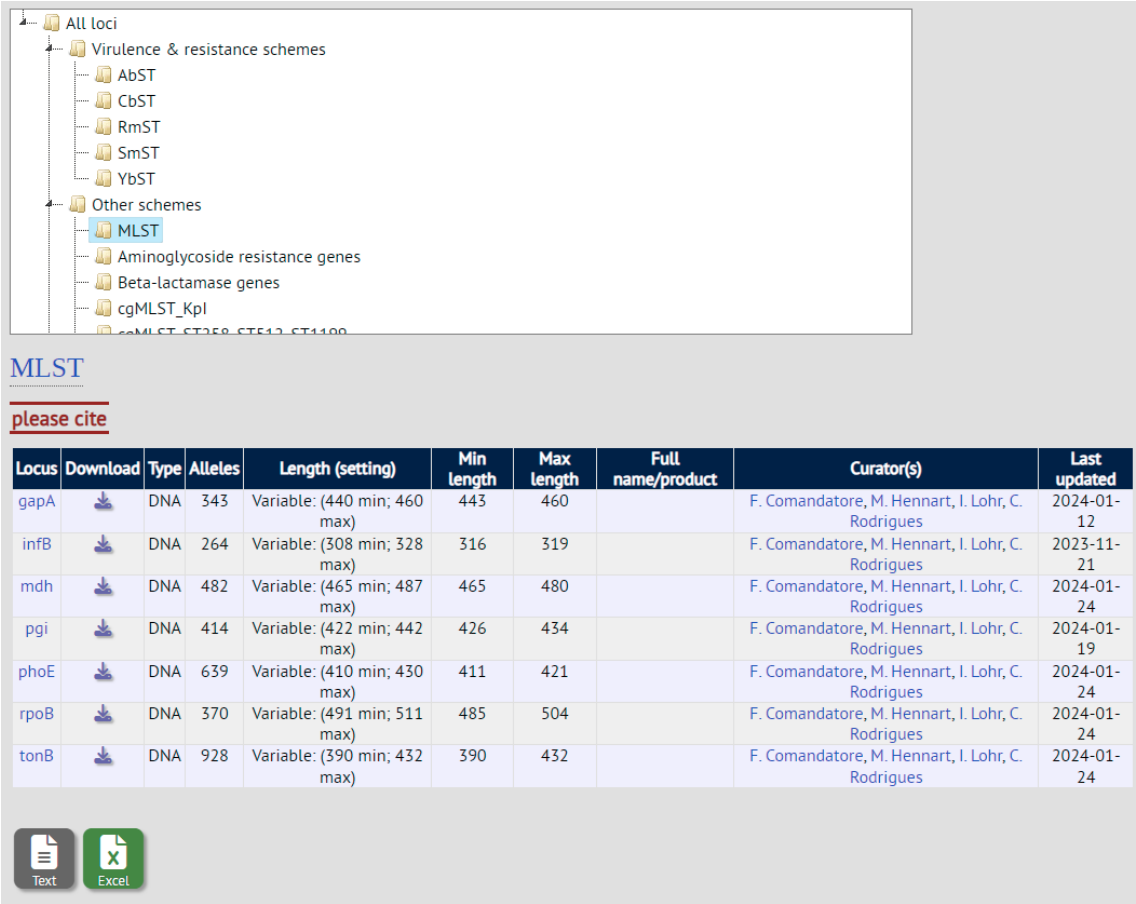


Obr. 2.6: Stránka databáze alel *Klebsielly pneumoniae* na BIGSdbPasteur, [26]



Obr. 2.7: Vyhledávání alel na BIGSdbPasteur a jeho výsledky

Další užitečnou funkcí je možnost stažení sekvencí jednotlivých alel, pole 2 na obrázku 2.6. Dají se stáhnou jednotlivě a lze je vyhledávat podle názvu příslušného lokusu, jejich id, délky a různé kombinace těchto a dalších identifikačních parametrů. Další z možností je zobrazit si rovnou celé schéma a stáhnout alely do něho patřících lokusů [25]. Toto je ukázáno na obrázku 2.8. Sekvence jsou stahovány ve formátu fasta.



Obr. 2.8: Stažení schémat na BIGSdbPasteur

Poslední zmíněnou funkcí je možnost dohledání nebo stažení alelických profilů u jednotlivých schémat. Jedná o pole 3 na obrázku 2.6. Je možnost zobrazit čísla alel, které tvoří konkrétní sekvenční typ, nebo naopak, při znalosti čísla alel se dají vyfiltrovat sekvenční typy, které je obsahují. Další variantou je stažení všech sekvenčních typů s korespondující kombinací čísel alel, což bude blíže ukázáno v praktické části této práce.

3 Frameworky pro tvorbu webové aplikace

Frameworky jsou sady nástrojů a knihoven, které pomáhají vývojářům rychleji a efektivněji vytvářet webové aplikace. Poskytují strukturovaný způsob organizace kódu a nabízí řadu funkcí, které usnadní vývoj webových aplikací. Jedná se například o implementaci formulářů nebo zajišťování přihlašování uživatelů a zabezpečení jejich hesel. Další podstatnou funkcí je možnost šablon pro dynamické generování HTML.

Mezi populární frameworky pro Python patří Django a Flask. Django je vysokoúrovňový webový framework a poskytuje tak mnoho vestavěných funkcí. Proto je vhodný pro komplexnější webové aplikace. [27]

Flask je mikroframework, což znamená, že je jednoduchý, ale lehce rozšiřitelný. Nabízí tak větší flexibilitu a jednoduchost pro menší projekty. Jeho další výhodou je potřeba menšího množství kódu. Z tohoto důvodu byl vybrán jako framework pro použití při vývoji nástroje BaGeTo - Bacterial Gentyping Tool. [28]

4 Vlastní webový nástroj BaGeTo

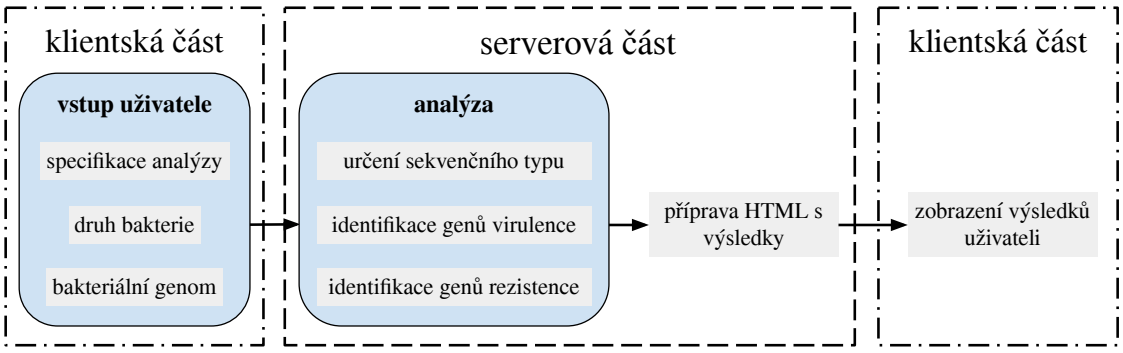
BaGeTo – Bacterial Genotyping Tool je webový nástroj integrující určování ST, hledání genů rezistence a hledání genů virulence pro různé bakterie. V současné podobě zvládne analyzovat bakterie *Klebsiella pneumoniae* a *Staphylococcus aureus*. Je však navržen tak, aby přidávání dalších bakterií bylo snadné.

BaGeTo využívá již vytvořených nástrojů a spojuje je pro analýzu bakteriálního genomu na jednom místě. Konkrétně se jedná o nástroj BLAST, který je použit pro určování, která alela se v analyzovaném genomu nachází a o nástroj ResFinder pro vyhledání genů rezistence vůči antibiotikům a predikci fenotypu.

Tato kapitola se věnuje představení jednotlivých částí nástroje a přiblížení vnitřní logiky algoritmu. Zároveň jsou zde u relevantních sekcí ukázány grafické výstupy nástroje BaGeTo.

4.1 Serverová část

Serverová část nástroje zajišťuje provedení analýzy, uchovávání výsledků v interních databázích a přípravu výsledků k zobrazení uživateli. Ten nemůže ovlivňovat server jinak než prostřednictvím klientské části, což zaručuje větší bezpečnost a reprodukti-bilitu výsledků. Obrázek 4.1 znázorňuje interakci mezi klientem a serverem. Nástroj je schopen analyzovat bakteriální genom nahraný v jednom souboru fasta formátu. Genom musí být složen alespoň do podoby kontigů.

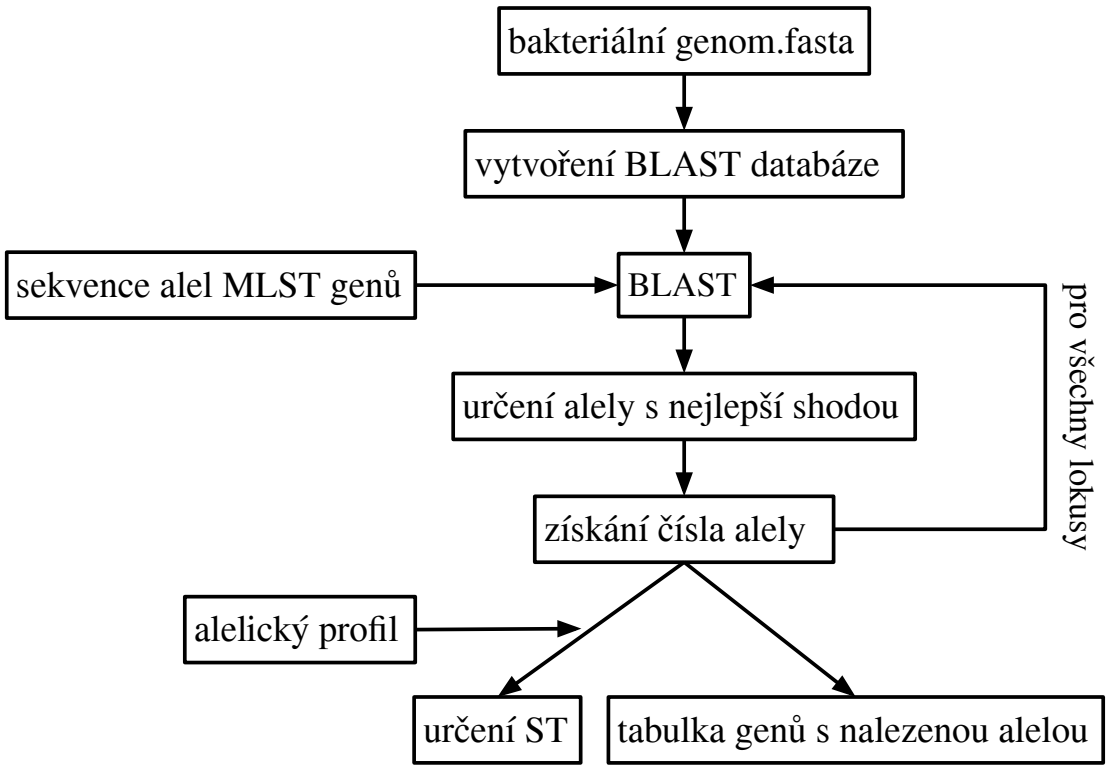


Obr. 4.1: Přehled vytvořeného algoritmu

4.1.1 Určení sekvenčního typu

Tato část algoritmu se věnuje určování sekvenčního typu MLST z nahraného bakteriálního genomu. Blokové schéma lze najít na obrázku 4.2. Algoritmus využívá lokálně staženého nástroje BLAST verze 2.14.1+, sekvence všech alel pro jednotlivé

lokusy a alelický profil pro MLST. Pro *Klebsiellu pneumoniae* byly alely a schéma staženy z BIGSdbPasteur a pro *Staphylococcus aureus* z databáze PubMLST.



Obr. 4.2: Blokové schéma určování sekvenčního typu

V prvním kroku dojde k vytvoření databáze z nahraného genomu tak, aby bylo možno hledat zarovnání pomocí BLAST. Následně dojde k vybrání správného typizačního schématu na základě druhu bakterie, který uživatel poskytl. Pro *Klebsiellu pneumoniae* je v MLST sedm lokusů. Pro každý z nich dojde k vyhledávání všech příslušných alel ve vytvořené databázi pomocí BLAST. Z důvodu následné filtrace a dalšího zpracování jsou získané výsledky načteny jako datový typ `DataFrame`. Jako alela, která se na daném lokusu nachází, je vybrána ta se 100% shodou. V případě, že existuje více takových alel, je zvolena ta s nejvyšším skóre zarovnání. Toto proběhne pro všechny lokusy v daném schématu.

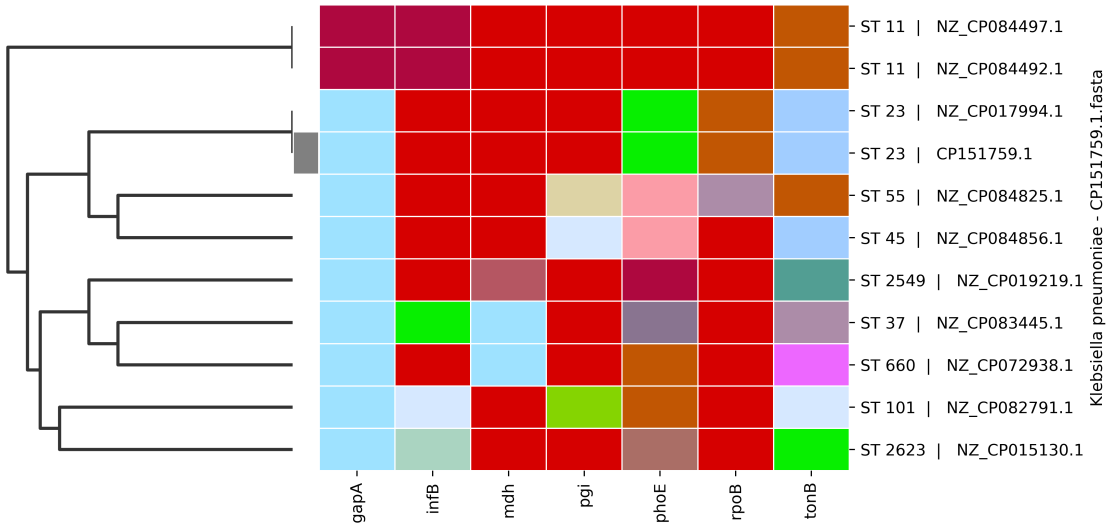
Alely byly staženy z veřejných databází ve formátu fasta, kdy hlavička je ve formátu `lokus_číslo-alely`. Pomocí regulárních výrazů se extrahuje číslo dané alely, načež dojde k uložení těchto čísel ve specifickém pořadí. Každý sekvenční typ má toto pořadí unikátní. Na obrázku 4.3 jde vidět, jak vypadá stažený alelický profil MLST z BIGSdbPasteur. Dojde k prohledání tohoto souboru a určení sekvenčního typu bakterie. Toto, spolu s čísly jednotlivých alel je uloženo do souboru `druh_MLST_results.csv`, kde každý řádek odpovídá jedné proběhlé analýze. Tento

soubor je použit k porovnání vzorku s předchozími analyzovanými. Pokud nebyly nalezeny alely pro všechny lokusy v typizačním schématu, je vrácena tabulka s alelami, které nalezené byly a informace, že nebylo možné ST určit. Tato analýza se neukládá do souboru `druh_MLST_results.csv`.

ST	gapA	infB	mdh	pgi	phoE	rpoB	tonB
1	4	4	1	1	7	4	10
2	3	4	1	1	9	4	17
3	5	5	1	1	9	6	11
6785	2	1	2	1	10	4	916
6786	6	3	1	1	632	1	4
6787	3	266	2	4	1	46	4

Obr. 4.3: Ukázka alelického profilu MLST

Dále je vygenerována heatmapa, ve které je daný vzorek porovnán se záznamy z interní databáze, která je reprezentována výše zmíněným souborem. Na obrázku4.4 je znázorněna podobnost jednotlivých vzorků pomocí dendrogramu. Současný vzorek je označen šedým polem lokalizovaným mezi dendogramem a samotnou heatmapou. Popisky jednotlivých záznamů se sestávají ze sekvenčního typu a názvu nahraného souboru. Na okraji obrázku se nachází nápis informující o názvu souboru nahraného k analýze a jaký druh bakterie uživatel zvolil.



Obr. 4.4: Vygenerovaný obrázek pro porovnání sekvenčního typu

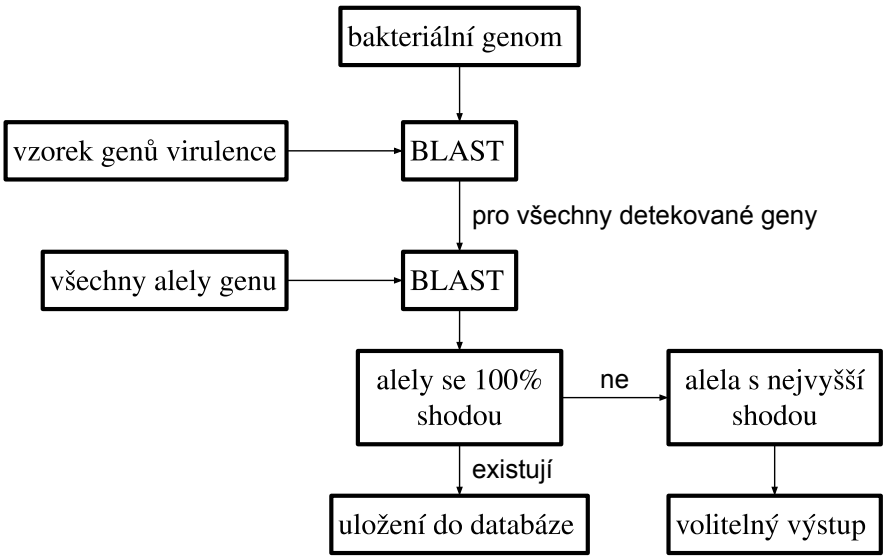
4.1.2 Identifikace genů rezistence

Pro identifikaci genů rezistence a následnou predikci fenotypové rezistence k antibiotikům byl využit nástroj ResFinder verze 4.4.2 od CGE, dostupný pod licencí Apache2.0. [26]

Pro vyhledávání jsou používány parametry o stejné hodnotě jako ty přednastavené u webové verze. Konkrétně se jedná o Threshold for %ID alespoň 90 % a parametr Minimum length rovno 60 %. Dále dojde k vytvoření složky resfinder_results, do které se uloží výsledky vyhledávání. Více informací k výsledkům a jednotlivým parametrům ResFinderu se nachází v kapitole 2.2.

4.1.3 Identifikace genů virulence

Identifikace genů virulence probíhá podobně jako určování alel při MLST viz obrázek 4.5. Sekvence alel hledaných genů ve formátu fasta byly pro *Klebsiellu pneumoniae* staženy z BIGSdbPasteura, a pro *Staphylococus aureus* z databáze nástroje VirulenceFinder. Byl vytvořen pomocný soubor obsahující právě jednu alelu každého genu, který byl použit pro prvotní určení, zda se gen v daném vzorku nachází nebo ne. Pro všechny geny, u kterých BLAST našel shodu s daným genomem, dochází k dalšímu vyhledávání. Zde již do BLAST vstupuje soubor se všemi alelami daného genu. Následně jsou generovány dvě tabulky, jedna s geny u nichž bylo zjištěno číslo alely a druhá pro geny, kde byla nalezena shoda vyšší než shoda požadována uživatelem, ale menší než 100 %.



Obr. 4.5: Blokové schéma zjišťování přítomnosti genů virulence

V další části dojde k uložení nalezených genů do souboru `druh_virulence.csv`. Hlavička souboru obsahuje názvy jednotlivých genů a jejich přítomnost v genomu je značena číslem 1, nepřítomnost číslem 0. Zároveň se ukládá i název souboru s genomem. Soubor `druh_virulence.csv` je využit k porovnávání s předchozími analyzovanými vzorky a podobně jako u určování ST je vygenerována heatmapa s dendogramem. Při tvorbě dendrogramu je v obou případech pro výpočet distanční matice použita Jaccardova vzdálenost.

4.1.4 Vytvoření HTML stránky s výsledky

V této části dochází k úpravě dat a k jejich odeslání do klientské části nástroje. Pro všechny tabulky, které jsou převáděny do HTML, jsou aplikovány kaskádové styly zajišťující příjemnější a přehlednější grafické zobrazení.

U dat z určování ST dojde k oříznutí bílé plochy nad vygenerovanou heatmapou, která vzniká skrytím barevné škály. Zároveň dojde k převedení dat ze souboru `druh_MLST_results.csv` do HTML tabulky a k jejich seřazení, aby odpovídalo pořadí v heatmapě. ST daného genomu je také vrácen.

U genů virulence se jedná o podobný proces, kdy vygenerovaná heatmapa je opět zbavena bílé plochy a jsou vytvořeny dvě HTML tabulky. První obsahuje geny, u kterých byla nalezena alela se 100% přesností a druhá, ve které se nacházejí geny s minimální shodou zadanou uživatelem.

Pro geny rezistence dojde ke zpracování textových souborů výsledků generovaných nástrojem ResFinder. U jednotlivých souborů je z textu extrahována tabulka s predikovaným fenotypem, nalezenými geny rezistence a tabulka s chromozomálními mutacemi. Tabulka fenotypu je následně seřazena tak, ať jsou antibiotika, pro která je rezistence predikována, nahoře. Tyto řádky jsou zároveň při převodu do HTML obarveny červeně. Zobrazení antibiotik, pro které rezistence predikována není, je volitelné. Ostatní tabulky jsou taktéž převedeny do HTML.

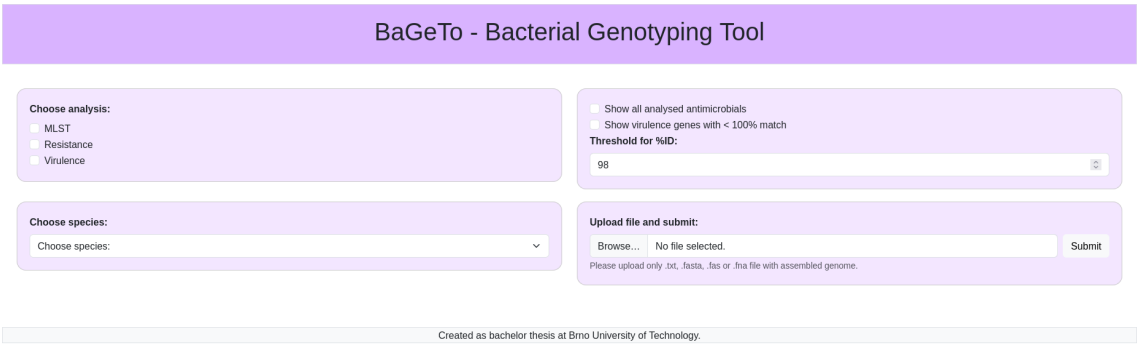
4.2 Klientská část

Tato část nástroje obsahuje kód týkající se interakce s uživatelem a zároveň se jedná o data, ke kterým má uživatel přístup v rámci svého webového prohlížeče. Obsahuje část nezbytnou pro chod programu a část estetickou, jejíž změna funkci BaGeTo neovlivní. Účelem klientské části programu je zjištění, jakou analýzu chce uživatel provést, o jakou bakterii se jedná a nahrání souboru s genomem. Tyto údaje jsou odeslány serveru a uživateli je následně vygenerována stránka s výsledky.

4.2.1 Zadávání údajů

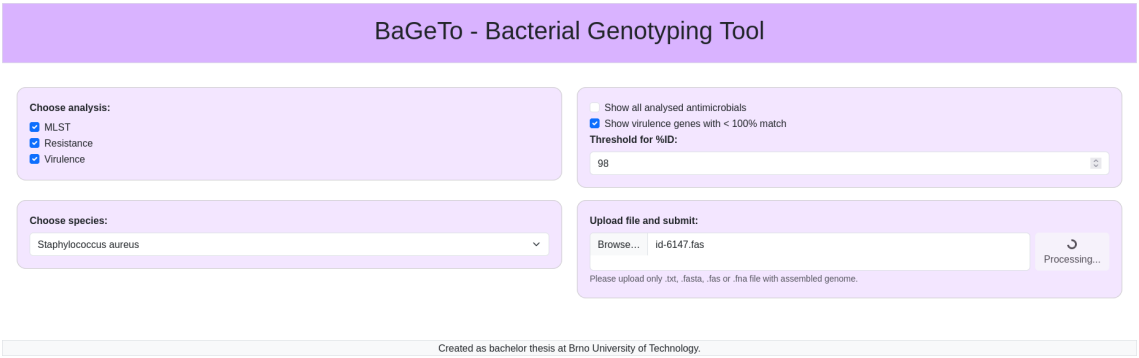
Jak je vidět na obrázku 4.6 úvodní stránka je minimalistická, tak aby používání bylo co nejrychlejší a jednoduché také pro lidi, kteří neovládají angličtinu.

Uživatel si může vybrat a zaškrtnou pouze ty typy analýzy, které chce provést. Toto umožňuje rychlejší zobrazení výsledků, kdy není třeba vždy analyzovat vše. Zobrazení antibiotik, pro která nebyla predikována rezistence, je dalším z volitelných parametrů. Stejně tak zobrazení genů virulence s méně než 100% shodou. Spodní hranice je volitelná. Další pole jsou již povinná a uživateli není umožněno odeslat data k analýze bez jejich vyplnění. Jedná se o pole rozbalovacího seznamu, kde se vybírá druh bakterie k analýze. Po kliknutí na pole `Browse...` je uživateli zobrazeno okno k vybrání a nahrání souboru z jeho počítače. Po zmáčknutí tlačítka `Upload` dojde k odeslání dat serveru a zahájení analýzy.



Obr. 4.6: Úvodní strana BaGeTo

Následně se zobrazí stránka, viz 4.7. Podstatná je změna tlačítka `Upload` na tlačítko `Processing...`, což zabraňuje několikanásobnému odeslání stejného formuláře. Zároveň lze vidět, jaké typy analýzy jsou zaškrtnuty, zvolená bakterie a název nahraného souboru.

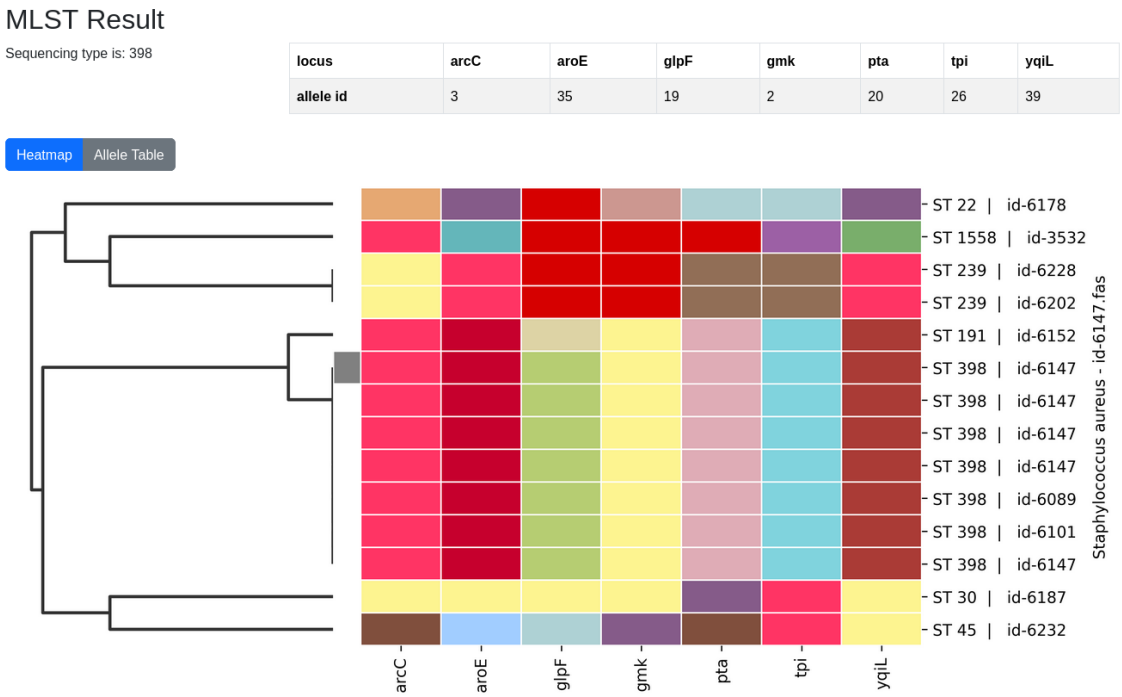


Obr. 4.7: BaGeTo po odeslání požadavku

4.2.2 Zobrazení výsledků

Výsledky určování ST lze vidět na obrázku 4.8. Vzhledem k tomu, že číslo alely nenese žádnou informaci, byly barvy v heatmapě zvoleny s ohledem na maximalizaci diskriminační schopnosti v závislosti na variabilitě alel v datasetu. Barvy nenesou žádnou informaci a pro každý dataset může barevná škála odpovídat jiným číslům alel.

Existuje možnost překliknout na tabulku, která čísla alel, stejně jako sekvenční typ, zobrazuje. Záznamy jsou řazeny ve stejném pořadí jako v heatmapě. Přechod mezi těmito dvěma zobrazeními zajišťují tlačítka `Heatmap` a `Allele Table`, která jsou na obrázku 4.8. Aktivní tlačítko je označeno modře a neaktivní šedě.



Obr. 4.8: Zobrazení výsledku při určování ST pomocí BaGeTo

Nalezené geny rezistence jsou prezentovány ve třech tabulkách, které zobrazují predikovaný fenotyp, nalezené geny a detekované chromozomální mutace.

Na obrázku 4.9 je tabulka s predikovanou AMR. Podle volby uživatele jsou zobrazeny buď všechna antibiotika, nebo pouze ta s predikovanou rezistencí. V další tabulce na obrázku 4.10, jsou vidět nalezené geny rezistence spolu s informacemi o jejich vyhledávání. Obrázek 4.11 zobrazuje část poslední tabulky, která obsahuje detekované bodové mutace. Tabulka informuje o názvu mutace, jaké nukleotidy jsou zaměněny, a jaká aminokyselina je nyní kódována.

Resfinder Results

Antimicrobial	Class	WGS-predicted phenotype	Genetic background
fluoroquinolone	quinolone	Resistant	-
cephalosporins	under_development	Resistant	-
carbapenem	under_development	Resistant	-
ciprofloxacin	quinolone	Resistant	OqxB (OqxB_EU370913), OqxA (OqxA_EU370913)
unknown beta-lactam	beta-lactam	Resistant	blaSHV-190 (blaSHV-190_KP868753)
trimethoprim	folate pathway antagonist	Resistant	OqxB (OqxB_EU370913), OqxA (OqxA_EU370913)
chloramphenicol	amphenicol	Resistant	OqxB (OqxB_EU370913), OqxA (OqxA_EU370913)
fosfomycin	fosfomycin	Resistant	fosA6 (fosA6_KU254579)
nalidixic acid	quinolone	Resistant	OqxB (OqxB_EU370913), OqxA (OqxA_EU370913)
clindamycin	lincosamide	No resistance	-
carbomycin	macrolide	No resistance	-

Obr. 4.9: Zobrazení predikce rezistence vůči antibiotikům z BaGeTo

Acquired AMR gene hits

Resistance gene	Identity	Alignment Length/Gene Length	Coverage	Position in reference	Contig	Position in contig	Phenotype	Accession no.
blaSHV-190	99.88	861/861	100.0	1..861	CP151759.1 Klebsiella pneumoniae strain Kp067-M1 chromosome, complete genome	158896..159756	Unknown Beta-lactam	KP868753
fosA6	98.33	420/420	100.0	1..420	CP151759.1 Klebsiella pneumoniae strain Kp067-M1 chromosome, complete genome	1963136..1963555	Fosfomycin	KU254579
OqxB	98.76	3153/3153	100.0	1..3153	CP151759.1 Klebsiella pneumoniae strain Kp067-M1 chromosome, complete genome	3876239..3879391	Chloramphenicol, Nalidixic acid, Ciprofloxacin, Trimethoprim	EU370913
OqxA	99.15	1176/1176	100.0	1..1176	CP151759.1 Klebsiella pneumoniae strain Kp067-M1 chromosome, complete genome	3879415..3880590	Chloramphenicol, Nalidixic acid, Ciprofloxacin, Trimethoprim	EU370913

Obr. 4.10: Zobrazení nalezených genů rezistence z BaGeTo

Chromosomal mutations mediating AMR

Mutation	Nucleotide change	Amino acid change	Resistance	PMID
acrR p.P161R	CCG -> CGG	P -> R	Fluoroquinolone	12936981
acrR p.G164A	GGC -> GCC	G -> A	Fluoroquinolone	12936981
acrR p.F172S	TTC -> TCC	F -> S	Fluoroquinolone	12936981
acrR p.R173G	CGA -> GGG	R -> G	Fluoroquinolone	12936981
acrR p.L195V	CTC -> GTC	L -> V	Fluoroquinolone	12936981

Obr. 4.11: Zobrazení nalezených chromozomálních mutací z BaGeTo

Geny virulence, u kterých byla alela nalezená se 100% shodou, jsou zobrazeny v tabulce, která obsahuje kromě názvu genu a čísla alely také parametry výsledků

BLAST vyhledávání. Tato tabulka je ukázána na obrázku 4.12. Stejná tabulka je generována také pro geny se shodou nižší než 100 %, konkrétní práh si může zvolit uživatel. Přítomnost této tabulky je volitelná.

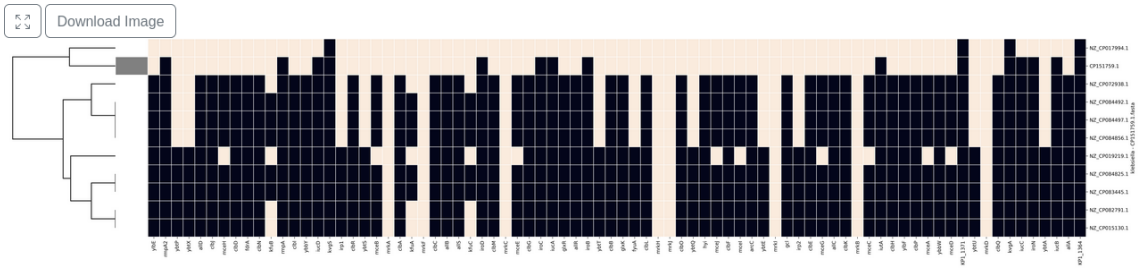
Virulence Results

Virulence genes with 100% match

gene_name	id	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore
ylbE	1	CP151759.1	100.00	1260	0	0	1	1260	1333341	1332082	0.00×10^{0}	2327
ybtP	2	CP151759.1	100.00	1713	0	0	1	1713	4628862	4630574	0.00×10^{0}	3164
ybtX	2	CP151759.1	100.00	1281	0	0	1	1281	4632356	4633636	0.00×10^{0}	2366

Obr. 4.12: Zobrazení genů virulence z BaGeTo

Poslední částí je heatmapa zobrazující přítomnost genů virulence pro porovnání s předchozími vzorky z databáze. Vzhledem k velkému množství genů je obtížně čitelná, jak dokazuje obrázek 4.13. Proto je možné ji zobrazit v plné velikosti a k procházení využít posuvník. Zároveň je možné ji v plné velikosti stáhnout.



Obr. 4.13: Zobrazení genů virulence a jejich vizuální indikace z BaGeTo

5 Výsledky a diskuze

Webový nástroj BaGeTo umožňuje v současné podobě bioinformatickou analýzu dvou klinicky významných druhů bakterií: *Klebsiella pneumoniae* a *Staphyllococcus aureus*. Účelem BaGeTo je zrychlení procesu analýzy genomu a to tak, že integruje určování ST, detekci genů virulence a detekci genů rezistence. Zároveň u nahraného genomu určí podobnost se záznamy v databázi a výsledek prezentuje formou dendrogramu.

Správnost fungování nástroje byla pro *Klebsiellu pneumoniae* ověřena pomocí webového nástroje stránek BIGSdb-Pasteur, zmiňovaném v kapitole 2.3. U všech jedenácti sekvencí stažených z veřejně dostupné databáze GenBank byl ST určen správně. Stejně tak se shodují geny virulence nalezené se 100% shodou. Soubory s nalezenými geny pro každou sekvenci obsahuje elektronická příloha. Výsledky byly identické také při použití deseti složených genomů poskytnutých Centrem molekulární biologie a genetiky Fakultní nemocnice Brno. Nepodařilo se ověřit, zda nástroj BaGeTo správně identifikuje geny, které mají shodu menší než 100 %. Stránky BIGSdb-Pasteur tuto možnost nenabízejí a bylo by nutné hledat každý gen zvlášť.

Správnost určení ST pro *Staphylococcus aureus* byla ověřena stažením deseti sekvencí o známém ST z databáze PubMLST. Všechny byly určeny správně. Nalezené geny virulence byly porovnány s výsledky nástroje VirulenceFinder 2.0.5, při zadaných parametrech threshold for %ID 90 % a minimum length 100 %. Tímto způsobem bylo možné ověřit, zda nástroj BaGeTo správně detekuje geny virulence se shodou menší než 100 %.

U devíti genomů byly správně detekovány všechny geny virulence. U genů s menší než 100% shodou se detekovaná hodnota shody mezi nástroji lišila o zaokrouhlovací chybu. Nástroj VirulenceFinder poskytuje výsledky se dvěma desetinnými místy, proto byly pro porovnání takto zaokrouhleny také výsledky z BaGeTo. U desátého genomu byl nástrojem BaGeTo detekován také gen **sec3**, který však Virulencefinder nenalezl.

genome	gene name	id	pident	length	mismatch	evalue	bitscore
id-3532	sec3	2:M28364.1	98.002	801	16	0.0	1391
id-6187	sen	2:AP014653.1	97.426	777	20	0.0	1325

Tab. 5.1: Část výsledku z BaGeTo pro detekovaný gen **sec3**

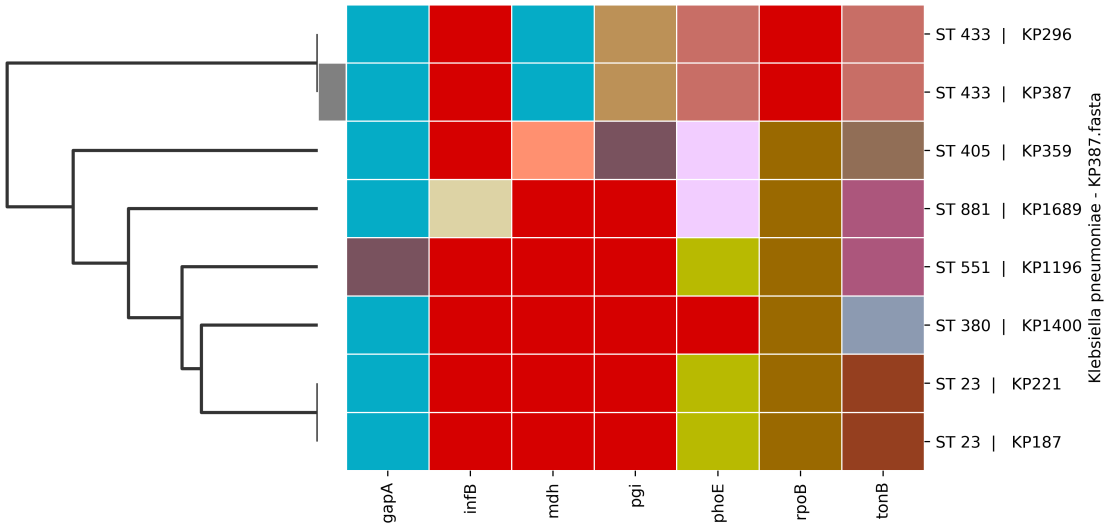
Jak je vidět v tabulce 5.1, gen nebyl nalezen se 100 % shodou. Předpoklad, že by nástroj VirulenceFinder obsahoval omezení na počet záměn nukleotidů, byl vyvrácen po kontrole výsledků dalších analyzovaných genomů. Z tabulky 5.1 lze vidět, že gen **sen**, který byl v genomu `id-6187.fasta` detekován oběma nástroji,

má o 4 záměny více než chybně detekovaný gen **sec3**. Když je tento gen nalezen se 100% shodou, jeho přítomnost určí oba nástroje. Není tak jasné, proč tato chyba u nástroje BaGeTo existuje. Přehled genomů, které byly použity k testování lze najít v tabulce B.1.

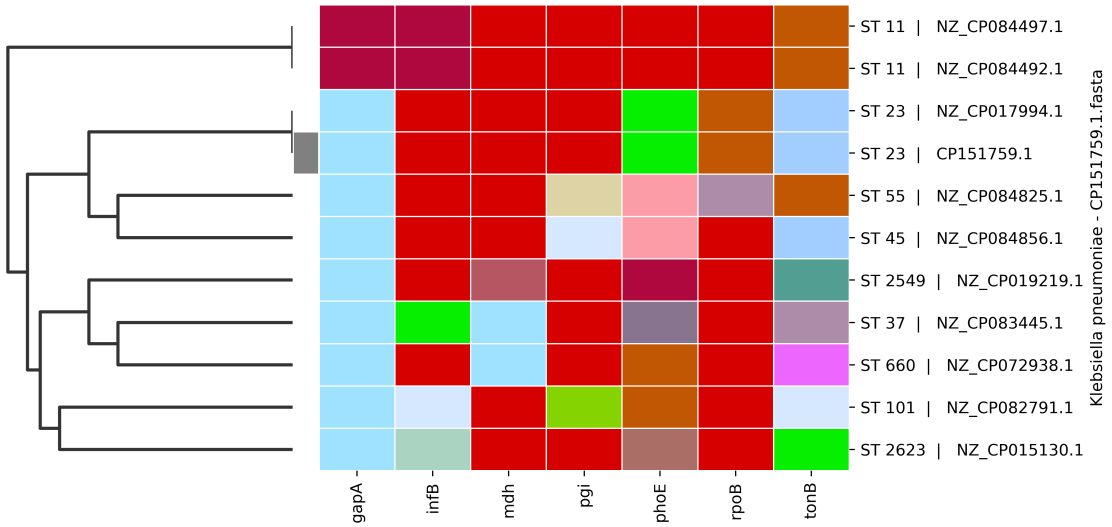
K detekci genů rezistence a predikce fenotypu byl použit již zavedený nástroj ResFinder a výsledky z něj byly upraveny pouze po vizuální stránce. Proto správnost výsledků ověřována nebyla.

5.1 Porovnání analyzovaných genomů

Složené genomy *Klebsielly pneumoniae* poskytnuté Centrem molekulární biologie a genetiky Fakultní nemocnice Brno byly sekvenovány pomocí platformy Oxford Nanopore. U dvou genomů nebylo možné určit sekvenční typ, a proto nejsou ukázány na obrázku 5.1. Sekvenční typ nebylo možno určit z důvodu nedetekování alely u jednoho z lokusů v typizačním schématu. Je možné, že se tak projevila vyšší chybovost platformy Oxford Nanopore při sekvenování, což v kombinaci s nutností detekovat alelu se 100% přesností, vede k neschopnosti ji určit.



Obr. 5.1: Výsledky určování ST u genomů bakterií z nemocnice



Obr. 5.2: Výsledky určování ST u genomů bakterií z databáze

Při porovnání dendrogramů na obrázcích 5.1 a 5.2 si lze všimnout, že u sekvenčních typů získaných z nemocnice jsou větve kratší a více se řetězí. V obou případech byla pro vytvoření dendogramu využita metoda UPGMA. Kratší větve znamenají větší podobnost mezi ST. Pozorování je v souladu s předpokladem, že bakterie vyskytující se v jedné nemocnici budou příbuzné, zatímco genomy bakterií náhodně vybraných z databáze ne. Bylo by vhodné toto tvrzení objektivně ověřit analýzou většího množství genomů a například porovnáním průměrné vzdálenosti mezi sekvencemi v jednotlivých skupinách. Graf A.1 v příloze poskytuje přehled genů virulence nalezených se 100% shodou.

5.2 Limitace

BaGeTo poskytuje pouze omezené možnosti nastavení jednotlivých parametrů vyhledávání. Úprava vstupních parametrů nástroje ResFinder je uživateli nepřístupná. Stejně tak nemá možnost editovat databázi, se kterou je analyzovaný genom porovnáván. A to může být problém z důvodu automatického ukládání analyzovaných genomů do této databáze.

Jednotlivé genomy jsou identifikovány podle názvu souboru, který uživatel nahraje. V případě nahrání více souborů se stejným názvem není ze strany uživatele možnost jednoznačně přiřadit záznam konkrétnímu genomu. Interně jsou záznamy do databáze ukládány chronologicky, některé omyly by tedy bylo možno opravit.

Závěr

Hlavním cílem bakalářské práce je vytvoření nástroje, který detekuje geny virulence, rezistence a určí ST nahraného bakteriálního genomu. Porovnání zjištěného genotypu v rámci lokální databáze a vytvoření webového rozhraní tvoří další cíl.

Vznikl nástroj BaGeTo, který k určování genů rezistence a následné predikci fenotypu využívá již zavedený nástroj ResFinder. Detekce genů virulence se provádí pomocí nově vytvořeného algoritmu založeném na BLAST. Určování sekvenčního typu probíhá podobně jako detekce genů virulence.

Jednotlivé úspěšně určené ST a detekované geny virulence se ukládají do souborů ve formátu CSV. Ty slouží jako lokální databáze, v rámci kterých je analyzovaný genom porovnáván. Distanční matice je vypočítána pomocí Jaccardovy vzdálenosti a dendrogram sestaven metodou UPGMA.

Webový nástroj BaGeTo pracuje s genomem klinicky významných bakterií *Klebsiella pneumoniae* a *Staphylococcus aureus*. Sekvenační data musí být poskytnuta jako jeden soubor ve fasta formátu a složena alespoň do podoby kontigů.

Výsledky detekce genů virulence u jedenácti genomů *Klebsielly pneumoniae* stažených z databáze GenBank a deseti poskytnutých Centrem molekulární biologie a genetiky Fakultní nemocnice Brno jsou k dispozici v elektronické příloze. Nalezené geny a jejich alely jsou identické jako u nástroje poskytovaného BIGSdb-Pasteur. Stejně tak ST je určen identicky.

V případě *Staphylococcus aureus* ověření výsledků probíhalo za pomoci dvou nástrojů a deseti genomů stažených z databáze. ST byl ve všech případech úspěšně ověřen na stránkách PubMLST. Úspěšnost detekce genů virulence je porovnávána s nástrojem VirulenceFinder. Ve vzorku `id-3532.fas` je odlišně detekován gen **sec3**. VirulenceFinder ho ve svých výsledcích neuvádí a BaGeTo ho nachází se shodou 98,002 %. Další výsledky si již odpovídají.

Správnost detekce genů rezistence nebyla vzhledem k použití nástroje ResFinder ověřována. Tento nástroj funguje v operačním systému Linux. Z toho důvodu i BaGeTo vyžaduje operační systém Linux, u Windows není podporován. Ostatní části algoritmu jsou funkční i pro operační systém Windows. V současné době není BaGeTo na žádném veřejně přístupném webovém serveru. Vývojové práce byly prováděny v lokálním prostředí na osobním počítači. To znamená, že aplikace musí být spuštěna manuálně pomocí příkazového řádku nebo terminálu, což omezuje její dostupnost pouze na IP adresu v lokální síti.

Do budoucna je možnost nasadit BaGeTo na veřejný server, tak ať je přístupný všem uživatelům pomocí url adresy. Než k tomu dojde, bylo by vhodné vytvořit registraci a přihlašování jednotlivých uživatelů tak, ať mají přístup pouze k datům která nahráli sami a vytváří se jim jejich lokální databáze výsledků.

Literatura

[1] MAHNER, Martin a KARY, Michael. What Exactly Are Genomes, Genotypes and Phenotypes? And What About Phenomes? Online. *Journal of Theoretical Biology*. 1997, roč. 186, č. 1, s. 55-63. ISSN 00225193. Dostupné z: https://doi.org/10.1006/jtbi.1996.0335. [cit. 2023-12-01].

[2] REN, Zhongqing; LIAO, Qin; KARABOJA, Xheni; BARTON, Ian S.; SCHANTZ, Eli G. et al. Conformation and dynamic interactions of the multipartite genome in Agrobacterium tumefaciens. Online. *Proceedings of the National Academy of Sciences*. 2022, roč. 119, č. 6. ISSN 0027-8424. Dostupné z: https://doi.org/10.1073/pnas.2115854119. [cit. 2023-12-01].

[3] SHARMA, Aditya Kumar; DHASMANA, Neha; DUBEY, Neha; KUMAR, Nishant; GANGWAL, Aakriti et al. Bacterial Virulence Factors: Secreted for Survival. Online. *Indian Journal of Microbiology*. 2017, roč. 57, č. 1, s. 1-10. ISSN 0046-8991. Dostupné z: https://doi.org/10.1007/s12088-016-0625-1. [cit. 2023-12-05].

[4] OLIVE, D. Michael a BEAN, Pamela. Principles and Applications of Methods for DNA-Based Typing of Microbial Organisms. Online. *Journal of Clinical Microbiology*. 1999, roč. 37, č. 6, s. 1661-1669. ISSN 0095-1137. Dostupné z: https://doi.org/10.1128/JCM.37.6.1661-1669.1999. [cit. 2023-11-29].

[5] 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. 2023-12-28].

[6] 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. 2023-12-27].

[7] 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. 2023-12-28].

[8] RAGHAVENDRA, Pongali a PULLAIAH, Thammineni. Pathogen Identification Using Novel Sequencing Methods. Online. In: *Advances in Cell and Molecular Diagnostics*. Elsevier, 2018, s. 161-202. ISBN 9780128136799. Dostupné

z: https://doi.org/10.1016/B978-0-12-813679-9.00007-5. [cit. 2023-12-28].

[9] AMBARDAR, Sheetal; GUPTA, Rikita; TRAKROO, Deepika; LAL, Rup a VAKHLU, Jyoti. High Throughput Sequencing: An Overview of Sequencing Chemistry. Online. *Indian Journal of Microbiology*. 2016, roč. 56, č. 4, s. 394-404. ISSN 0046-8991. Dostupné z: https://doi.org/10.1007/s12088-016-0606-4. [cit. 2023-12-29].

[10] Chapter 2 - Techniques for Oral Microbiology. Online. In: *Atlas of Oral Microbiology*. Elsevier, 2015, s. 15-40. ISBN 9780128022344. Dostupné z: https://doi.org/10.1016/B978-0-12-802234-4.00002-1. [cit. 2024-05-27].

[11] WANG, Yunhao; ZHAO, Yue; BOLLAS, Audrey; WANG, Yuru a AU, Kin Fai. Nanopore sequencing technology, bioinformatics and applications. Online. *Nature Biotechnology*. 2021, roč. 39, č. 11, s. 1348-1365. ISSN 1087-0156. Dostupné z: https://doi.org/10.1038/s41587-021-01108-x. [cit. 2023-12-30].

[12] CHEN, Zhao; ERICKSON, David L. a MENG, Jianghong. Polishing the Oxford Nanopore long-read assemblies of bacterial pathogens with Illumina short reads to improve genomic analyses. Online. *Genomics*. 2021, roč. 113, č. 3, s. 1366-1377. ISSN 08887543. Dostupné z: https://doi.org/10.1016/j.ygeno.2021.03.018. [cit. 2024-05-29].

[13] SIMAR, Shelby R.; HANSON, Blake M. a ARIAS, Cesar A. Techniques in bacterial strain typing: past, present, and future. Online. *Current Opinion in Infectious Diseases*. 2021, roč. 34, č. 4, s. 339-345. ISSN 0951-7375. Dostupné z: https://doi.org/10.1097/QCO.0000000000000743. [cit. 2023-11-29].

[14] IBARZ PAV-N, Ana Belén a MAIDEN, Martin C.J. Multilocus Sequence Typing. Online. In: CAUGANT, Dominique A. (ed.). *Molecular Epidemiology of Microorganisms*. Methods in Molecular Biology. Totowa, NJ: Humana Press, 2009, s. 129-140. ISBN 978-1-60327-998-7. Dostupné z: https://doi.org/10.1007/978-1-60327-999-4_11. [cit. 2024-01-02].

[15] RUPPITSCH, Werner. Molecular typing of bacteria for epidemiological surveillance and outbreak investigation / Molekulare Typisierung von Bakterien für die epidemiologische Überwachung und Ausbruchsabklärung. Online. *Die Bodenkultur: Journal of Land Management, Food and Environment*. 2016, roč. 67, č. 4, s. 199-224. ISSN 0006-5471. Dostupné z: https://doi.org/10.1515/boku-2016-0017. [cit. 2024-01-01].

[16] BEZDICEK, Matej; NYKRYNOVA, Marketa; SEDLAR, Karel; KRALOVA, Stanislava; HANSLIKOVA, Jana et al. Rapid high-resolution melting genotyping scheme for Escherichia coli based on MLST derived single nucleotide polymorphisms. Online. *Scientific Reports*. 2021, roč. 11, č. 1. ISSN 2045-2322. Dostupné z: https://doi.org/10.1038/s41598-021-96148-3. [cit. 2024-01-03].

[17] MADDEN, Thomas. *The NCBI Handbook [Internet]. 2nd edition*. Online. 2. 2013. Dostupné z: https://www.ncbi.nlm.nih.gov/books/NBK153387/. [cit. 2024-01-31].

[18] U.S. NATIONAL LIBRARY OF MEDICINE. *NCBI help manual*. Online. NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (US) . BETHESDA (MD). National library of Medicine: National Center for Biotechnology Information. 2005. Dostupné z: https://www.ncbi.nlm.nih.gov/books/NBK3831/. [cit. 2024-05-26].

[19] 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. 2024-05-26].

[20] *Standard Nucleotide BLAST*. Online. NCBI, National Library of Medicine. Dostupné z: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome. [cit. 2024-01-31].

[21] BORTOLAIA, Valeria; KAAS, Rolf S; RUPPE, Etienne; ROBERTS, Marilyn C; SCHWARZ, Stefan et al. ResFinder 4.0 for predictions of phenotypes from genotypes. Online. *Journal of Antimicrobial Chemotherapy*. 2020, roč. 75, č. 12, s. 3491-3500. ISSN 0305-7453. Dostupné z: https://doi.org/10.1093/jac/dkaa345. [cit. 2024-01-31].

[22] CLAUSEN, Philip T. L. C.; AARESTRUP, Frank M. a LUND, Ole. Rapid and precise alignment of raw reads against redundant databases with KMA. Online. *BMC Bioinformatics*. 2018, roč. 19, č. 1. ISSN 1471-2105. Dostupné z: https://doi.org/10.1186/s12859-018-2336-6. [cit. 2024-05-26].

[23] *ResFinder*. Online. Center for Genomic Epidemiology. 2011. Dostupné z: http://genepi.food.dtu.dk/resfinder. [cit. 2024-01-31].

[24] *Klebsiella PasteurMLST database*. Online. BIGSdb-Pasteur. Dostupné z: https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_isolates. [cit. 2024-01-31].

[25] DIANCOURT, Laure; PASSET, Virginie; VERHOEF, Jan; GRIMONT, Patrick A. D. a BRISSE, Sylvain. Multilocus Sequence Typing of Klebsiella pneumoniae Nosocomial Isolates. Online. *Journal of Clinical Microbiology*. 2005, roč. 43, č. 8, s. 4178-4182. ISSN 0095-1137. Dostupné z: `https://doi.org/10.1128/JCM.43.8.4178-4182.2005`. [cit. 2024-01-31].

[26] *Klebsiella locus/sequence definitions database*. Online. BIGSdb-Pasteur. Dostupné z: `https://bigsdb.pasteur.fr/cgi-bin/bigsdb/bigsdb.pl?db=pubmlst_klebsiella_seqdef`. [cit. 2024-01-31].

[27] DJANGO SOFTWARE FOUNDATION. *Django*. Online. 2005. Dostupné z: `https://www.djangoproject.com/`. [cit. 2024-01-31].

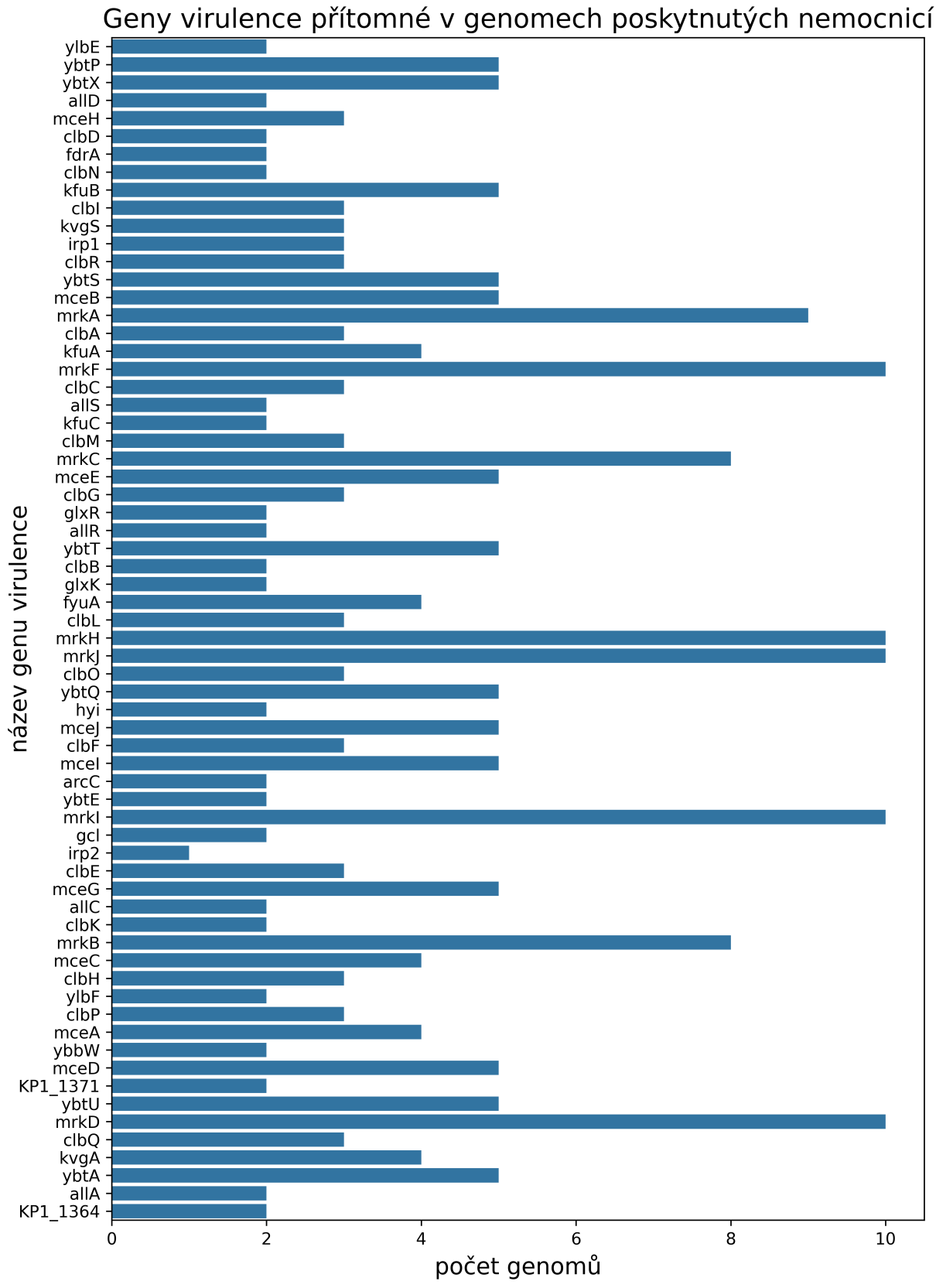
[28] PALLETS. *Flask*. Online. 2010. Dostupné z: `https://flask.palletsprojects.com/en/3.0.x/`. [cit. 2024-01-31].

Seznam symbolů a zkratek

PCR	polymerázová řetězová reakce
WGS	celogenomové sekvenování
MLST	multilokusová sekvenční typizace
PFGE	pulzní gelová elektroforéza
rep-PCR	repetitivní polymerázová řetězová reakce
NGS	sekvenování nové generace
SLST	jednolokusová sekvenční typizace
ST	sekvenční typ
cgMLST	core genome multilocus sequence typing
wgMLST	whole genome multilocus sequence typing
SNP	detekce bodových mutací
HRM	vysokorozlišovací analýza křivek tání
BLAST	basic local alignment search tool
NCBI	National Center for Biotechnology Information
MSP	maximal segment pair
HSP	high scoring pair
CGE	Center for Genomic Epidemiology
AMR geny	geny antimikrobiální rezistence
KMA	k-mer alignment
BIGSdb-Pasteur	Bacterial Isolate Genome Sequence database
BaGeTo	Bacterial Genotyping Tool
UPGMA	unweighted pair group method with arithmetic mean

Seznam příloh

A Detekované geny virulence v genomech



Obr. A.1: Geny virulence detekované v 10 genomech z FN Brno

B Sekvence použité k ověření fungování nástroje

Klebsiella pneumoniae	*Staphylococcus aureus*
GenBank	PubMLST
accession number	ID
CP151759.1	3532
NZ_CP015130.1	6089
NZ_CP017994.1	6101
NZ_CP019219.1	6147
NZ_CP072938.1	6152
NZ_CP082791.1	6178
NZ_CP083445.1	6187
NZ_CP084492.1	6202
NZ_CP084497.1	6228
NZ_CP084825.1	6232
NZ_CP084856.1	

Tab. B.1: Genomy použité při ověřování fungování nástroje BaGeTo