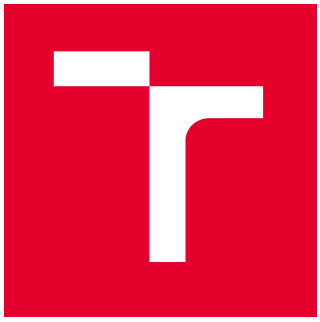
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

BAKALÁŘSKÁ PRÁCE

Brno, 2022 Ema Marta Lorková



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

METODY ANALÝZY PANGENOMU U BAKTERIÁLNÍCH POPULACÍ

BACTERIAL PAN-GENOME ANALYSIS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE

AUTHOR

Ema Marta Lorková

VEDOUCÍ PRÁCE

SUPERVISOR

Ing. Markéta Nykrýnová

BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Ema Marta Lorková ***ID:*** 220941

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

NÁZEV TÉMATU:

Metody analýzy pangenomu u bakteriálních populací

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši na téma analýza pangenomu včetně popisu alespoň tří volně dostupných nástrojů. 2) Vybrané nástroje otestujte na datasetu vybraných bakteriálních genomů. 3) Navrhněte vlastní metodu pro vyhledání a analýzu pangenomu a dílčí části realizujte. 4) Implementujte navrženou metodu pro vyhledání a analýzu pangenomu ve vhodném programovacím prostředí. 5) Implementovanou metodu otestujte na zvoleném datasetu vybraných bakteriálních genomů. 6) Získané výsledky porovnejte s výsledky volně dostupných softwarů a diskutujte je.

DOPORUČENÁ LITERATURA:

[1] KIM, Yeji, Changdai GU, Hyun Uk KIM a Sang Yup LEE. Current status of pan-genome analysis for pathogenic bacteria. Current Opinion in Biotechnology. 2020, 63, 54-62. DOI: 10.1016/j.copbio.2019.12.001.

[2] XIAO, Jingfa, Zhewen ZHANG, Jiayan WU a Jun YU. A Brief Review of Software Tools for Pangenomics. 2015, 13(1), 73-76. DOI: 10.1016/j.gpb.2015.01.007.

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

Vedoucí práce: Ing. Markéta Nykrýnová

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

UPOZORNĚNÍ:

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

ABSTRAKT

Táto bakalárska práca sa zaoberá analýzou pangenómu u bakteriálnych populácií. V prvej časti je opísaný genóm, pseudogény, baktérie vrátane ich genómu, pangenóm a jeho súčasti. Uvedený je popis štyroch vybraných nástrojov na analýzu pangenómu. V praktickej časti je prevedené testovanie týchto nástrojov a porovnanie ich výstupov pre genóm danej baktérie. Nakoniec je navrhnutý algoritmus pre vlastnú metódu analýzy, opísaný je vlastný program a získané výsledky.

KĽÚČOVÉ SLOVÁ

analýza pangenómu, genóm, baktéria

ABSTRACT

This bachelor thesis deals with pan-genome analysis for bacterial populations. The first part specifies genome, pseudogenes, bacteria and its genome, and pan-genome including its components. Following part introduces four tools for pan-genome analysis. Testing of these tools against specific bacterial genome and their comparison is mentioned in practical part. Lastly, an algorithm is implemented for creating new pipeline, and a code and obtained results are described.

KEYWORDS

pan-genome analysis, genome, bacteria

Vysadené pomocou balíčku `thesis` verzie 4.07; `http://latex.feec.vutbr.cz`

LORKOVÁ, Ema Marta. *Metody analýzy pangenomu u bakteriálních populací*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství, 2022, 43 s. Bakalárska práca. Vedúci práce: Ing. Markéta Nykrýnová

Vyhlásenie autora o pôvodnosti diela

Meno a priezvisko autora:	Ema Marta Lorková
VUT ID autora:	220941
Typ práce:	Bakalárska práca
Akademický rok:	2021/22
Téma závěrečnej práce:	Metody analýzy pangenomu u bakteriálních populací

Vyhlasujem, že svoju záverečnú prácu som vypracovala samostatne pod vedením vedúcej/cého záverečnej práce, s využitím odbornej literatúry a ďalších informačných zdrojov, ktoré sú všetky citované v práci a uvedené v zozname literatúry na konci práce.

Ako autorka uvedenej záverečnej práce ďalej vyhlasujem, že v súvislosti s vytvorením tejto záverečnej práce som neporušila autorské práva tretích osôb, najmä som nezasiahla nedovoleným spôsobom do cudzích autorských práv osobnostných a/alebo majetkových a som si plne vedomá následkov porušenia ustanovenia § 11 a nasledujúcich autorského zákona Českej republiky č. 121/2000 Sb., o práve autorskom, o právach súvisiacich s právom autorským a o zmene niektorých zákonov (autorský zákon), v znení neskorších predpisov, vrátane možných trestnoprávnych dôsledkov vyplývajúcich z ustanovenia časti druhej, hlavy VI. diel 4 Trestného zákonníka Českej republiky č. 40/2009 Sb.

Brno ....

podpis autorky*

*Autor podpisuje iba v tlačenej verzii.

POĎAKOVANIE

Rada by som sa poďakovala vedúcej bakalárskej práce, pani Ing. Markéte Nykrýnovej, za odborné vedenie, konzultácie, trpezlivosť a podnetné návrhy k práci.

Obsah

Zoznam obrázkov

Úvod

Táto bakalárska práca sa zaoberá metódami analýzy pangenómu. Pokrokmi v sekvenovaní dát bolo u bakteriálnych populácii experimentálne preukázané, že v jednotlivých bakteriálnych izolátoch sa gény vyskytujú v rôznom počte. Pangenóm je preto svojou stavbou veľmi užitočný na definovanie a popis bakteriálnych druhov.

Od definovania pojmu pangenóm bolo vytvorené množstvo nástrojov na jeho analýzu. Tie sa rôzne odlišujú vstupnými a výstupnými dátami, procesom spracovania dát, ovládaním, a nemusia byť jednoznačne definované. Pre účel tejto práce boli porovnané štyri nástroje.

Pangenómová analýza nám ponúka lepšie pochopenie kmeňov hlavne patogénnych baktérií na základe ich podobností aj odlišností. Výsledky analýz ponúkajú nové poznatky a môžu slúžiť na dôkladnejšie štúdium taxonómií a genetických odlišností. V neposlednom rade je analýza pangenómu dôležitá pre terapeutické postupy v medicínskej sfére, v rôznych biotechnologických a genetických aplikáciách.

V teoretickej časti tejto práce je cieľom oboznámiť čitateľa s genómom, pangenómom a baktériami. Následne sú opísané metódy analýzy pangenómu pomocou nástrojov, ktorými sú panX, PGAdb-builder, BPGA a Roary.

V praktickej časti je prevedené otestovanie daných nástrojov a opísané sú vstupy, výstupy a ovládanie jednotlivých nástrojov. Výsledky testovania sú opísané pre genóm *Streptococcus pneumoniae*. Obsahom praktickej časti je tiež návrh vlastnej metódy programu na analýzu pangenómu a popis jeho funkcií. Program je nakoniec otestovaný na upravenom datasete.

1 Teoretická časť

1.1 Genóm

Genóm je kompletný súbor genetickej informácie organizmu. Obsahuje všetky informácie potrebné k jeho správnemu fungovaniu. Termín genóm bol prvýkrát opísaný v roku 1920 Hansom Winklerom, ktorý genóm opisuje ako súbor haploidných chromozómov, ktoré spolu s príslušnou protoplazmou špecifikuje materiálne základy druhu [1].

Genóm sa rozdeľuje na segmenty DNA nazývané exóny, ktoré definujú funkčné gény, a na nekódujúce úseky DNA nazývané intróny. Rozoznávame eukaryotický genóm a prokaryotický genóm. U toho prvého je jeho priemerná dĺžka 3 200 Mbp a tvorí lineárne molekuly DNA [2]. O prokaryotickom genóme si viac povieme v kap. 1.2.1.

1.1.1 Ortológne a paralógne gény

V analýze pangenómu je dôležité rozumieť pojmu ortológne gény. Ortológne aj paralógne gény patria pod homológne gény, teda sa vyvinuli zo spoločného (ancestrálneho) génu.

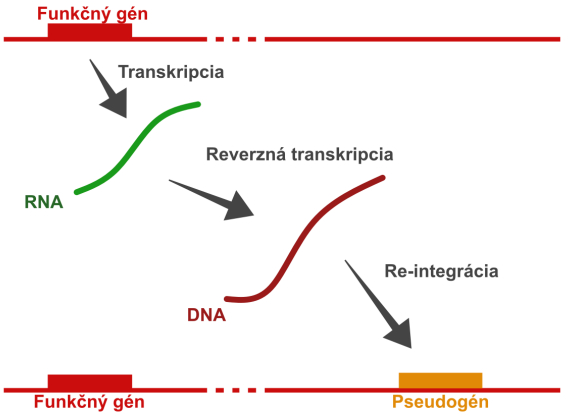
Rozdiel je ale v tom, že ortológne gény sú gény rôznych kmeňov, ktoré vznikli špeciáciou ancestrálneho génu a paralógne gény sú výsledkom duplikácie ancestrálneho génu. Ortológne gény sa využívajú na analýzu funkčných génov, pretože sú viac spoľahlivé ako paralógne, ktorých funkcia sa duplikáciou často mení [3].

1.1.2 Pseudogény

Pseudogény sa považujú za úseky genómu, ktoré sú podobné ozajstným génom, ale nedochádza v nich k transkripcii alebo translácii, pretože nukleotidová sekvencia bola pozmenená mutáciou. Niektoré mutácie majú len malý vplyv na aktivitu génu, pričom iné môžu zmenou len jedného nukleotidu spôsobiť nefunkčnosť génu. Tieto pseudogény sa nazývajú konvenčné (z angl. conventional).

Druhou skupinou pseudogénov sú spracované (z angl. processed). Tieto vznikajú počas génovej expresie z mRNA kópie génu transkribovanej z funkčnej DNA, ktorá sa následne opäť vloží do genómu, grafické znázornenie jeho vzniku je na obr. 1.1. Keďže je kópiou mRNA, spracovaný pseudogén neobsahuje intróny a tiež nukleotidové sekvencie zodpovedné za zapínanie expresie, čo definuje jeho neaktivitu [2].

Pseudogény sa v bakteriálnych genómoch vyskytujú pravidelne, dokonca aj v tých najmenších. Ich počet je v rozmedzí od 27 do 337 [4]. Uvádza sa, že viac ako



Obr. 1.1: Pôvod spracovaného pseudogénu, upravené a prevzaté z [2].

polovica pseudogénov vznikla z génov anotovaných ako hypotetické alebo neznáme. Vyskytujú sa vo veľkom počte v genómoch patogénnych baktérií, ktoré sa zoskupujú s hostiteľom, v porovnaní s ich voľne žijúcimi príbuznými [4], [5].

1.2 Baktérie

Baktérie sú jednobunkové prokaryotické organizmy. Na rozdiel od eukaryotov neobsahujú jadro a iné membránové organely. Na povrchu bakteriálnej bunky sa môžu nachádzať rôzne útvary, ako napríklad bičíky, fimbrie alebo ochranná vrstva nazývaná kapsula. Samotný povrch je tvorený bunkovou stenou z peptidoglykánu, ktorý hraje úlohu pri rozlišovaní baktérií na grampozitívne a gramnegatívne. Genetická informácia baktérií je obsiahnutá v nukleoide, ktorý sa voľne pohybuje v cytoplazme. Okrem neho sa pod bunkovou stenou a cytoplazmatickou membránou nachádzajú plazmidy, ribozómy a inklúzie. Plazmidy obsahujú dodatočnú genetickú informáciu v kruhovej molekule. Uvádza sa, že plazmidy môžu obsahovať gény, ktoré sú zodpovedné za rezistenciu baktérií voči antibiotikám [6], [7].

Baktérie majú rôzne tvary bunkového tela, ako napríklad guľovité (koky), tyčinkovité (bacily), špirálovité (spirilly, spirochéty), vlnité (vibriá) či dokonca hviezdicovité [8].

Výskyt baktérií je rôznorodý. Nachádzajú sa v každom biotope na zemi a obývajú aj iné organizmy, vrátane ľudí. Mnohé vytvárajú komenzalické, či dokonca symbiotické vzťahy. Relatívne len malý počet baktérií je patogénnych, zapríčiňujúci choroby či infekcie [9].

Baktérie sa rozmnožujú nepohlavne, a to najčastejšie binárnym delením. Tento proces začína zväčšením materskej bunky, ktorá zdvojnásobí svoj obsah. Dochádza

k replikácii bakteriálnej DNA, následne sa v strede vytvorí septum, ktoré bunku rozdelí na dve identické dcérske bunky [10].

1.2.1 Bakteriálny genóm

Hlavná časť bakteriálneho genómu je obsiahnutá v nukleoide. Ten je u väčšiny baktérií tvorený jedným kruhovým chromozómom, aj keď existujú výnimky v podobe lineárneho chromozómu u druhov *Borrelia* a *Streptomyces* [11]. V porovnaní veľkosti je bakteriálny genóm menší než eukaryotický a dosahuje veľkostí do 10 Mbp. U niektorých baktérií genóm zahŕňa nielen gény uložené v nukleoide, ale aj v plazmide [5].

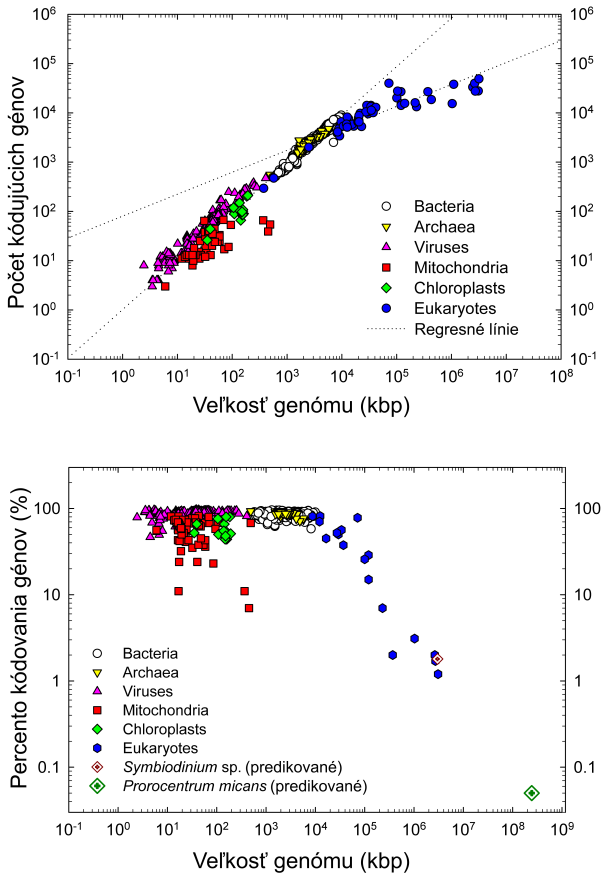
Bakteriálny genóm obsahuje len veľmi málo nekódujúcej DNA a pozostáva hlavne z funkčných častí, ktoré kódujú proteíny. V priemere tvoria funkčné gény 85-90 % genómu, pričom u niektorých genómov môže proteín-kódujúca hustota dosahovať menej než 40 %. Mnohé z týchto baktérií sú patogénne alebo obsahujú veľký počet pseudogénov [12]. Bakteriálne genómy teda vykazujú veľkú koreláciu medzi veľkosťou genómu a počtom funkčných génov, pre porovnanie s inými organizmami je uvedený obrázok 1.2.

1.2.2 *Streptococcus pneumoniae*

S. pneumoniae je grampozitívna, anaeróbna, patogénna baktéria. Tvorí dvojice (diplokoky) s okrúhlym, ale zašpicateným tvarom. Veľkosť genómu sa pohybuje v rozmedzí od 0,36 Mbp do 9,62 Mbp, s priemernou dĺžkou 2,12 Mbp. Počet génov je priemerne 2 181 a počet pseudogénov približne 144. Pneumokoky obývajú sliznicu horných dýchacích ciest ale môžu migrovať aj do pľúc. *S. pneumoniae* spôsobuje infekcie dýchacích ciest, ktorým zvyčajne predchádzajú iné ochorenia zapríčiňujúce oslabenú imunitu. Za najčastejšie ochorenia je považovaný zápal pľúc, meningitída, bakterémia, sinusitída a zápal stredného ucha [14]. Baktéria sa prenáša kvapôčkami vo vzduchu.

V prevencii proti pneumokokovým infekciám je využívaná hlavne terapia antibiotikami a vakcinácia. Nevýhodou antibiotík sú mutácie baktérie, ktoré spôsobujú rezistenciu voči rôznym druhom antibiotík. Očkovanie proti pneumokokom je vhodné najmä pre deti, starších ľudí a pre ohrozené skupiny osôb.

Streptococcus pneumoniae bol v tejto práci použitý na testovanie vybraných nástrojov pre analýzu pangenómu.



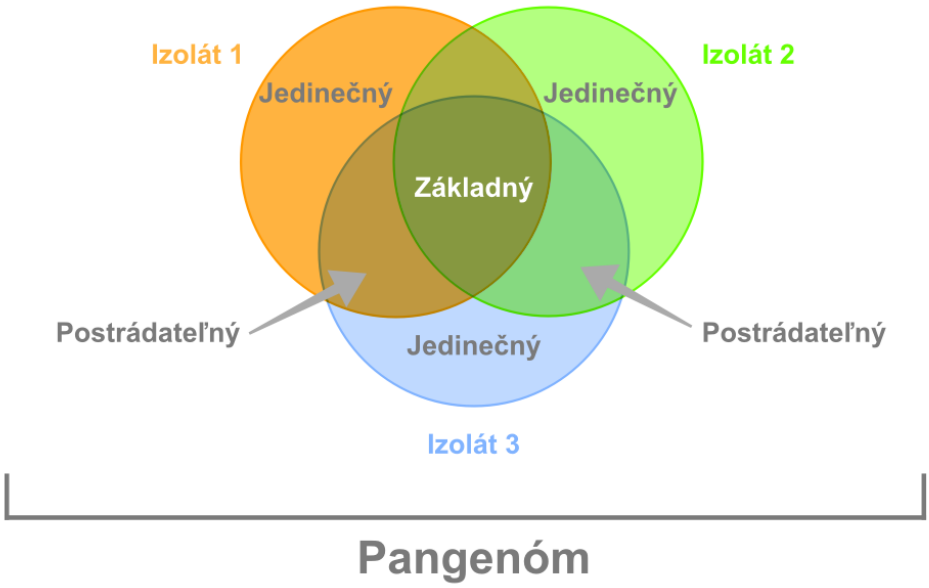
Obr. 1.2: Jednotlivé závislosti medzi veľkosťou genómu a počtom funkčných génov, upravené a prevzaté z [13].

1.3 Pangenóm a jeho súčasti

Termín pangenóm (pan, pôvodom z gréčtiny, čo znamená celý, úplný) bol prvýkrát použitý v štúdii z roku 2005 od Tettelin a kol. [15]. Pangenóm je v ňom definovaný ako spôsob opísania bakteriálnych druhov a skladá sa zo základného (z angl. core), postrádateľného (z angl. dispensable, často uvádzaného aj ako accessory), a jedinečného (z angl. unique) genómu. Grafické znázornenie pangenómu je na obr. 1.3.

Vo všeobecnosti sa dá povedať, že pangenóm definuje celý genómový súbor daného fylogenetického kladu, čiže reprezentuje všetky DNA sekvencie daného kladu [16].

V odbornej literatúre sa rozchádzajú definície rozdelenia pangenómu, v tejto práci sa ale budem držať rozdeleniu podľa Tettelina, pričom postrádateľný genóm ešte rozlíšim na shell a cloud podľa Livingstona [17].



Obr. 1.3: Prelínanie jednotlivých častí pangenómu.

1.3.1 Základný genóm

Základný genóm obsahuje gény vyskytujúce sa vo všetkých kmeňoch v rámci kladu. Väčšina týchto génov nesie informácie o základnej biologickej podstate kmeňa a patria do skupín prevádzkových (z angl. housekeeping) génov, ktoré sú zodpovedné za bunkový obal, regulačné funkcie, transport a väzbu proteínov, a jeho hlavné fenotypové znaky.

Niektoré zdroje základný genóm rozdeľujú ešte na hard core gény, ktoré je možné nájsť vo viac ako 99 % genómoch a soft core gény, ktoré tvoria 95 % - 99 % genómov [16], [17].

1.3.2 Postrádateľný a jedinečný genóm

Postrádateľný genóm, ktorý je vo vedeckých publikáciách často uvádzaný aj ako dodatočný genóm, teda podľa pomenovania zahŕňa doplnkové gény, prítomné v dvoch alebo viacerých kmeňoch, ale nie vo všetkých. Táto časť genómu nie je podstatná pre prežitie, keďže môže chýbať aspoň u jedného kmeňa v rámci druhu, podieľa sa ale na druhovej rozmanitosti kmeňa [18], [19].

Shell časť postrádateľného genómu je reprezentovaná génmi nájdenými vo väčšine genómov, ich výskyt sa pohybuje v rozmedzí 15 % - 95 %.

Cloud časť je zložená z génov obsiahnutých len v menšine genómov, čo predstavuje menej ako 15 % genómov [17].

Jedinečný genóm je jedinečný pre každý kmeň. Gény mu patriace tvoria približne 1,5 % genómu [19].

1.4 Analýza pangenómu

Rozvojom počítačových metód a nahromadením sekvenačných dát v posledných dekádach sa stala analýza pangenómu dôležitým článkom pre lepšie spracovanie množstva genómových dát. V neposlednom rade je pangenomová analýza užitočným nástrojom pre pochopenie vývoja a vzťahov bakteriálnych kladov. To môže v medicínskych aplikáciách nájsť uplatnenie napríklad v cielenej liečbe bakteriálnych onemocnení.

Od vytvorenia prvých nástrojov na pangenómovú analýzu v roku 2010 (Panseq a PanCGHweb) bolo predstavených cez tucet balíčkov a nástrojov. Všetky majú podobné funkcie, ale každý z nich má svoje charakteristické znaky a hranice. Sú schopné zhlukovania ortológnych génov, identifikácie jednonukleotidového polymorfizmu (SNP), konštruovania fylogenetických stromov a profilovanie génov [20].

1.4.1 panX

panX je voľne dostupné webové prostredie na štúdium a vizualizáciu bakteriálnych pangenómových dát. V postupe programu sú ako hlavné body uvedené:

1. Rozdelenie veľkých anotovaných genómov na gény.
2. Zhlukovanie génov do ortológnych skupín.
3. Určenie základného (core) genómu.

panX využíva ako vstup anotované genómy v GenBank formáte. Na identifikovanie homológnych skupín génov využíva panX algoritmus na hľadanie podobností DIAMOND. Jeho výstupom je súbor so všetkými pármi génov s významným skóre podobnosti, ktoré slúži na vstup pre Markovov klastrovací algoritmus (MCL). MCL následne zhlukuje gény do ortológnych klastrov. Keďže naivý algoritmus DIAMOND by mohol byť pre skupiny tisícov genómov nerealizovateľný, aplikuje sa stratégia rozdeľuj a panuj na podskupiny 50 genómov, z ktorých sa získava redukovaná reprezentatívna sekvencia, a nakoniec sa z týchto pseudogenómov skonštruujú kompletné zhluky.

Po určení základného genómu sa ďalej skúmajú ďalšie vlastnosti, ako vytvorenie fylogenézy pomocou SNP, konštruovanie fylogenetických stromov pre jednotlivé gény a mapovanie prítomnosti alebo neprítomnosti génu v strome pre základný genóm.

Výstupom je interaktívne prostredie s rôznymi tabuľkami, grafmi a stromami. Grafy podávajú informácie o základnej štatistike pangenómu a umožňujú filtráciu

získaných výsledkov podľa dĺžky či počtu génov. Jeden graf zobrazuje podiel základného a postrádateľného genómu, pričom každý z nich je možné kliknutím vybrať. Druhý graf predstavuje zhluky zoradené podľa klesajúceho počtu kmeňov v zhluku, posuvným oknom sa dajú nastaviť hranice vymedzujúce základný a postrádateľný genóm.

Tabuľky obsahujú štatistiky a anotácie pre génové zhluky, zobrazenie zarovnania génov a taktiež aj informácie o metadátach. Stromy vyobrazujú základný genóm a fylogenetický strom. Rozloženie stromov je ľahko možné preusporiadať [21].

1.4.2 PGAdb-builder

PGAdb je skratka pre databázu alel pangenómu, (z angl. pan-genome allele database). PGAdb-builder je teda voľne dostupný webový nástroj na skonštruovanie pangenomickej databázy. Server tohto nástoja zahŕňa dva funkčné moduly, a to Build_PGAdb na vytvorenie databázy a Build_wgMLSTtree na zostrojenie wgMLST stromu, ktorý udáva genetickú príbuznosť sekvencií. wgMLST (z angl. whole genome multilocus sequence typing) predstavuje typizáciu multilokusovej sekvencie celého genómu.

Ako prvý krok pri postupe wgMLST sa využíva stanovenie pangenomickej databázy alel. Je to rozšírený princíp roztriedenia dát z celogenómového sekvenovania (WGS) pre porovnávanie a analýzu rozloženia génov.

Vstupom do prvého modulu sú FASTA súbory genómov, z ktorých sú vytvorené anotácie použitím postupu programu Prokka [22]. Nasleduje umiestnenie proteínov do ortológnych zhlukov cez postup programu Roary [23] a paralógne gény sú z datasetu vyňaté. Ortológne zhluky pozostávajú z proteínovej rodiny, označovanej ako lokus, so sekvenčnou identitou nastaviteľnou medzi 90 % a 99%. V ďalšom kroku sú prevedené na nukleotidové sekvencie z dôvodu vytvorenia údajov o alelách pangenómu. Lokusy pangenomického datasetu sú zakódované a jednotlivým alelám v lokuse sú priradené čísla od 1.

Build_wgMLSTtree modul porovnáva vstupné genómové dáta s PGA databázou vytvorenou v prvom module cez BLASTN. Ak je alela prítomná v lokuse, je jej priradené vopred definované číslo. Ak chýba, je označená nulou. Po ukončení je pre daný genóm vytvorená alelická sekvencia, z ktorej je skonštruovaný dendrogram využitím metódy UPGMA, čo predstavuje metódu neváženého párovania s aritmetickým priemerom.

Výstup modulu Build_PGAdb obsahuje:

- sumár nastavení,
- koláčový graf vykresľujúci podiel základných, postrádateľných a jedinečných génov,

- sumárnu tabuľku
- zaškrtávacie políčka na výber definovanej schémy,
- tlačidlá pre vytvorenie wgMLST stromu a stiahnutie výsledku.

Výstup Build_wgMLSTtree modulu obsahuje sumár nastavení, strom genetickej príbuznosti a zoznam súborov na stiahnutie [24].

1.4.3 BPGA

Anglický názov Bacterial Pan Genome Analysis tool pre skratku BPGA, v preklade nástroj pre analýzu bakteriálneho pangenómu je veľmi rýchly postup programu s viacerými modulmi pre komplexné vyhodnotenie pangenómu. BPGA je možné stiahnuť a spustiť na operačných systémoch Windows a Linux aj s potrebnými doplnkami. Vstupom do programu môžu byť tri rôzne súbory:

- GenBank súbor,
- súbor s proteínovou sekvenciou: štandardný fasta formát, alebo
- binárna matica

Spracovaním vstupu sa získavajú súbory na vytvorenie zhlukov ortológnych génov alebo proteínov. Ako v jedinom pangenomickom nástroji si užívateľ môže vybrať z troch dostupných zhlukovacích nástrojov, a to USEARCH, CD-HIT a OrthoMCL, ktoré podávajú takmer rovnaké výsledky. Výstup zhlukovania generuje maticu prítomnosti/neprítomnosti génu, ktorá vstupuje do ďalších analýz.

BPGA dokopy zahŕňa 7 funkčných modulov:

1. Analýza profilu pangenómu.
2. Extrakcia sekvencie pangenómu.
3. Analýza výnimočnej génovej rodiny.
4. Analýza atypického C-G obsahu.
5. Funkčná analýza pangenómu.
6. Fylogenetická analýza druhu.
7. Analýza vybranej menšej podskupiny.

Prvé tri moduly sú zabudované v jednom kroku štandardnej pangenomickej analýzy. Tá zahŕňa rozdelenie ortológnych zhlukov do základného, postrádateľného a jedinečného genómu. Taktiež vytvára grafy častí pangenómu a rozpoznáva génové rodiny výlučne prítomné alebo neprítomné v špecifickom genóme.

Analýza atypického C-G obsahu vyberá sekvencie génov majúce značne vyššie alebo nižšie obsahy C-G oproti priemeru.

Funkčná analýza pangenómu prevádza COG a KEGG (Kyoto Encyklopedia od genes and Genomes) prirovnania pre reprezentatívne sekvencie všetkých rodín ortológnych génov a funkčne porovnáva jednotlivé časti pangenómu.

Modul fylogenetickej analýzy druhu vytvára fylogenetické stromy a modul analýzy vybranej menšej podskupiny spracováva užívateľom vybrané podskupiny genómov, ktoré si môže vybrať na základe akýchkoľvek vlastností organizmov [26].

1.4.4 Roary

Roary je voľne dostupný nástroj vyznačujúci sa rýchlou analýzou pangenómu aj pre tisícky vstupných izolátov. Na spustenie nástroja je potrebné si stiahnuť dostupné balíčky a samotná analýza sa potom spúšťa zadaním príkazu v príkazovom riadku s odkazom na vstupujúce súbory.

Vstupom do postupu programu sú anotované .gff súbory pre jednotlivé genómy. Zo súborov sú vyňaté kódujúce sekvencie, ktoré sú následne prevedené na proteínové sekvencie a filtrované, aby sa odstránili parciálne sekvencie a ďalej sa postupuje pred-zhlukovaním podľa CD-HIT.

Z redukovaných sekvencií je prevedené porovnanie každé-proti-každému cez BLASTP s voliteľnou identitou sekvencií. Tie sú potom zhlukované MCL metódou a výsledky sú zlúčené s pred-zhlukovaním z CD-HIT. Homológne skupiny sú roztriedené do skupín pravých ortológnych génov vďaka informáciám o konzervovaných génoch.

Vytvorený je aj graf opisujúci vzťahy medzi zhlukmi na základe poradia výskytu vo vstupných dátach, čím sa zhluky môžu usporiadať. Izoláty sú potom zoskupené podľa prítomnosti génu v postrádateľnom genóme [23].

2 Praktická časť

2.1 Testovanie nástrojov pre analýzu pangenómu

Na testovanie nástrojov bola vybraná baktéria *Streptococcus pneumoniae* a genómy jej kmeňov získané z databázy NCBI. Na analýzu bolo vybraných 52 genómov jej kmeňov, ktoré ponúka aplikácia panX. Genómy boli z NCBI stiahnuté ako Assembly so všetkými potrebnými súbormi z GenBank databázy, alebo v 11 prípadoch z RefSeq databázy, ich podrobnejší popis je uvedený v tabuľkách A.1, A.2 a A.3 v prílohe. V jednom genóme pod identifikátorom NC_021004 sa ale vyskytol problém, pretože genóm bol z databázy vymazaný, a preto sa v aplikáciách PGAdb-builder, BPGA a Roary pracovalo so zvyšnými 51 genómami.

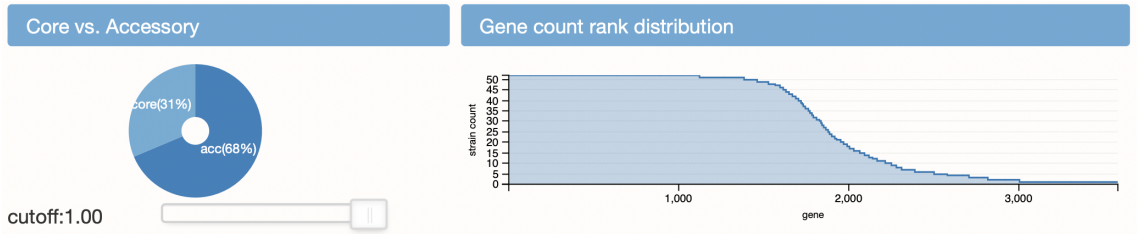
Priemerná veľkosť genómu je 2,12 Mbp a priemerný počet génov je 2 162.

2.1.1 panX

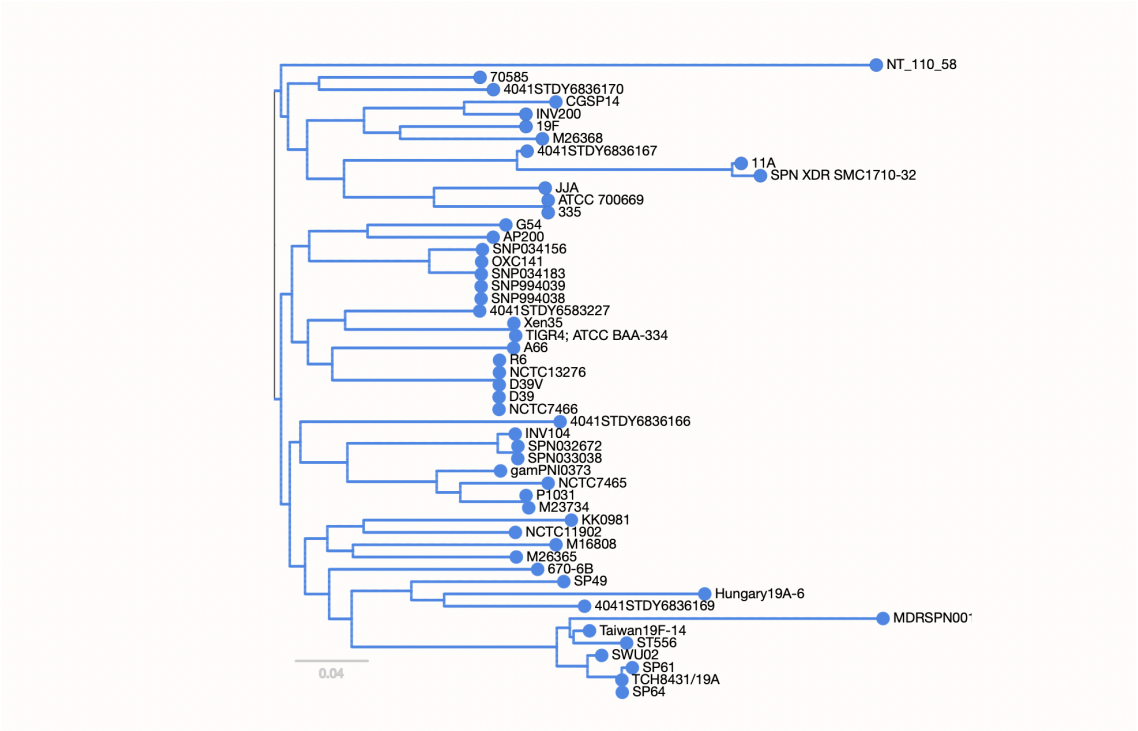
Výstupom testovania webovej aplikácie panX je interaktívne webové okno s grafmi a tabuľkami. Z koláčového grafu na obr. 2.1 bolo zistené, že z celkového počtu 3 582 génov, 31 % pangenómu tvorí základný genóm, to predstavuje 1 122 génov, ktoré sa vyskytujú v každom kmeni. 68 %, čo je 2 459 génov, tvorí postrádateľný genóm.

Z tabuľky génových zhlukov (gene cluster table) boli zistené ďalšie informácie o pangenóme. Najdlhšiu dĺžku génu má nepomenovaný gén s anotáciou „accessory Sec-dependent LPXTG-anchored adhesin" s dĺžkou takmer 21 kbp vyskytujúci sa len v jednom kmeni. Najkratší gén, „hypothetical protein" je o dĺžke 72 bp, vyskytujúci sa v piatich kmeňoch. Najväčší počet zmien v sekvencii (gene gain/loss event) génu, 18, má NS_transporter, bez zmien je 1 122 génov. Nevýhodou je, že táto tabuľka sa nedá exportovať pre ďalšiu analýzu a taktiež nie je možné určiť, presne ktoré gény tvoria základný genóm.

Výstupom je taktiež fylogenetický strom, znázornený na obr. 2.2.



Obr. 2.1: Ukážka grafického výstupu webovej aplikácie panX.



Obr. 2.2: Fylogenetický strom znázorňujúci SNP v základnom genóme.

2.1.2 PGAdb-builder

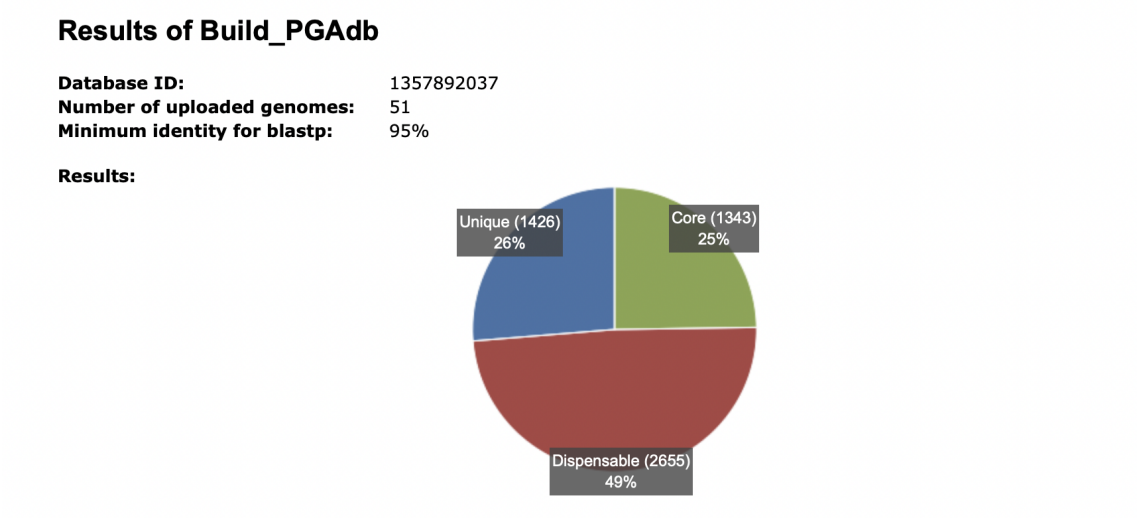
Po nahraní 51 genómov *S. pneumoniae* na server PGAdb-builer bol požadovaný čas na spracovanie odhadnutý na 2,75 hodiny. Po prebehnutí analýzy na module Build_PGAdb bol v koláčovom grafe zobrazený výsledok, znázornený na obr. 2.3. *S. pneumoniae* podľa PGAdb-builder obsahoval 5 424 lokusov, z ktorých 25 % (1 343) patrí základnému genómu, 49 % (2 655) postrádateľnému a 26 % (1 426) jedinečnému genómu.

Výsledkom je tiež sumárna tabuľka s nájdenými génmi, nie je z nej ale možné určiť, v ktorých konkrétnych genómoch sa jednotlivé gény nachádzajú, iba ich percentuálny výskyt.

Modulom Build_wgMLSTtree bol vytvorený dendrogram znázornený na obr. 2.4. Označenie koncových vetiev si nástroj generuje sám, a preto ich nie je možné jednoznačne priradiť k samostatným genómom kmeňov baktérie.

2.1.3 BPGA

Vstupom do programu BPGA boli fasta súbory 51 genómov *S. pneumoniae*. Spustená bola one-click analysis pre rýchle spracovanie a identita sekvencií bola zvolená na 95 %. Vo výsledku boli získané rôzne grafy a tabuľky, taktiež aj xcl súbory s



Obr. 2.3: Výstupný koláčový graf s číselnými hodnotami početností pre jednotlivé časti pangenómu, získaný z programu PGAdb-builder.

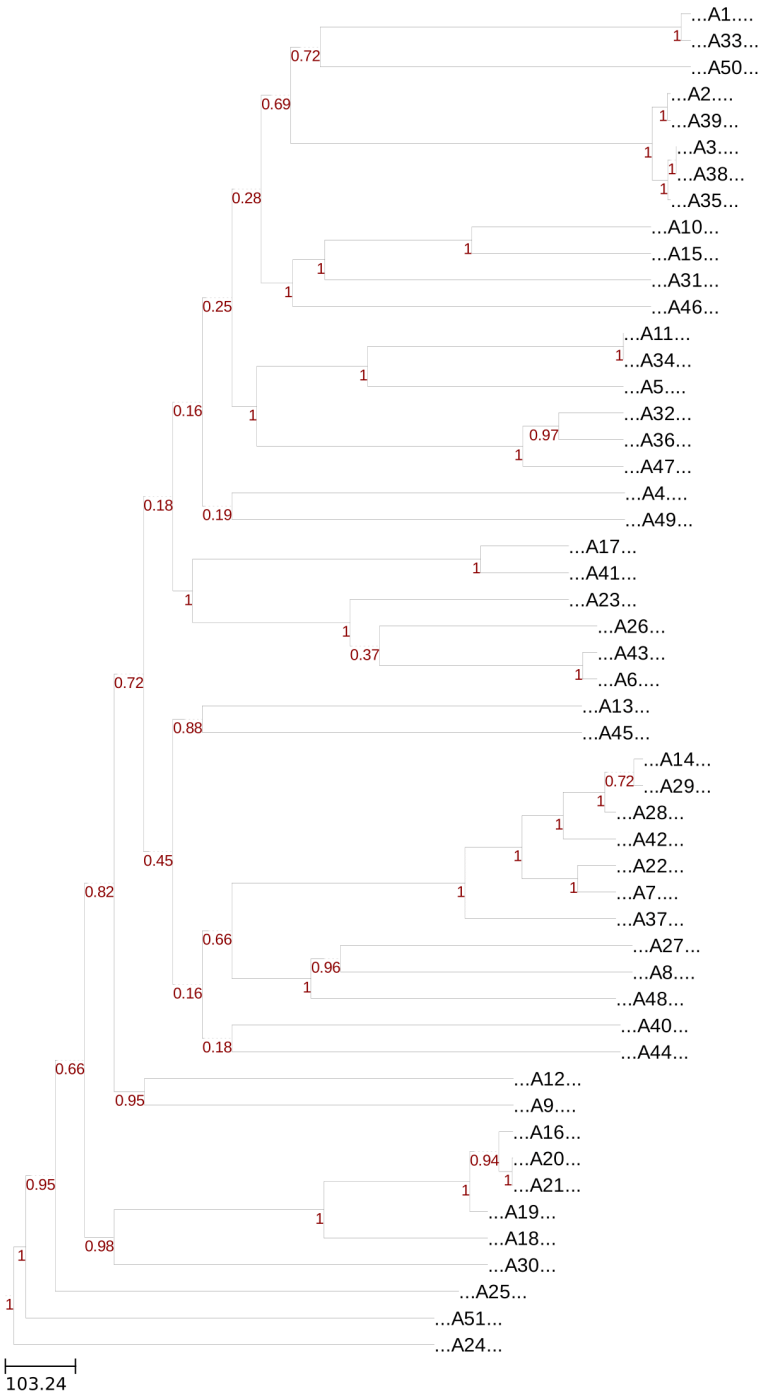
popisom výstupu. Z celkového pangenómu s počtom génov 1 996, je počet základných génov 981, počet postrádateľných génov priemerne 968 a jedinečných génov je v priemere 43. Vypočítaný je aj počet výnimočne chýbajúcich génov, a to sú 4 gény. Získané boli fylogenetické stromy, napr. pre celkový pangenóm, na obr. 2.5. Zaujímavý je graf COG distribúcie, na obr. 2.6. Histogram ukazuje rôzne kategórie a triedy funkcií zhlukov ortológnych génov. Môžeme z neho usúdiť, že najviac génov zodpovedá za transláciu, rekombináciu a opravu bunky a najmenej ich je zodpovedných za pohyblivosť bunky.

2.1.4 Roary

Po spustení analýzy na 51 anotovaných súboroch *S. pneumoniae* boli vygenerované dva súbory. Za zmienku stojí tabuľka prítomnosti alebo neprítomnosti génu, v ktorej sú obsiahnuté informácie o názve génu, počte izolátov, v ktorých sa gén vyskytuje a iné číselné dáta. Tabuľka tiež ukazuje presne v ktorých izolátoch boli jednotlivé gény nájdené. Druhý súbor je vo formáte newick. Oba výstupné súbory sú vstupom do skriptu v jazyku Python pre grafické znázornenie výsledkov.

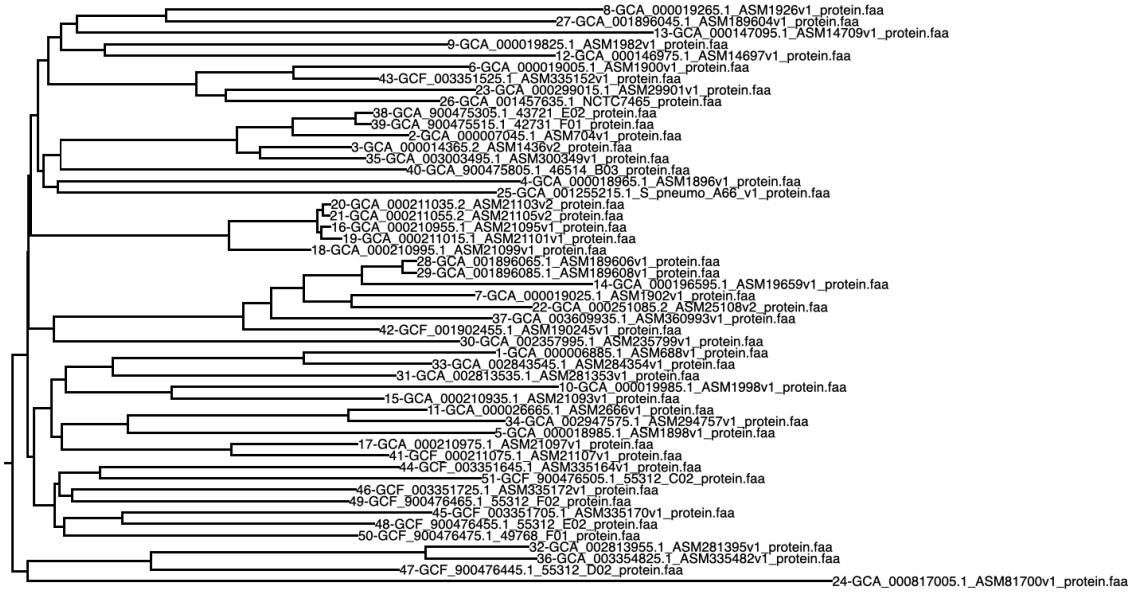
Najdôležitejším výstupom je koláčový graf na obr. 2.7 obsahujúci informácie o počte génov tvoriacich pangenóm a taktiež informáciu o tom, v koľkých izolátoch baktérie sa nachádzajú. Základný genóm tvorí 1 022 génov, pričom sa rozlišuje aj soft-core, ktorý zahŕňa 252 génov. Postrádateľný genóm Roary rozdeľuje na jeho cloud a shell časť. V shell časti bolo nájdených 1 359 génov a cloud pangenóm tvorí až 3 474 génov.

V grafickom výstupe je tiež obsiahnutý fylogenetický strom a matica výskytu



Obr. 2.4: Dendrogram pre všetkých 51 izolátov *S. pneumoniae*, získaný z programu PGAdb-builder.

génov v jednotlivých izolátoch baktérie, na obr. 2.8, z ktorej môžeme usúdiť, ako sa izoláty zhodujú.



Obr. 2.5: Fylogenetický strom pangenómu, získaný z BPGA.

2.2 Porovnanie testovaných nástrojov

Všetky nástroje sa v číselných hodnotách líšia vo veľkosti celkového pangenómu. Najväčší počet génov, až 6 107 našlo Roary a najviac sa od neho líši BPGA s počtom 1 996 génov.

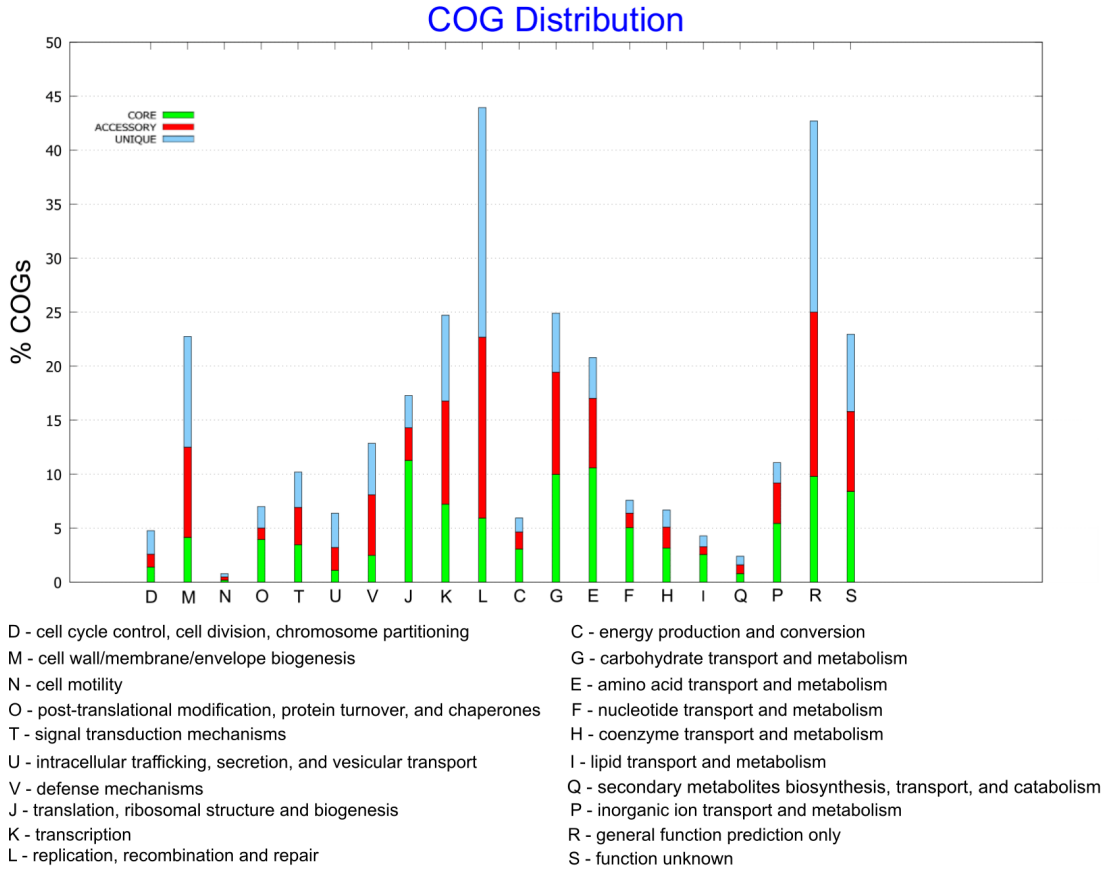
V určení veľkosti základného pangenómu sú nástroje celkom podobné, s priemerným počtom génov 1 117. Najviac zhodné sú BPGA a Roary, ktoré sa odlišujú len o 41 génov, najmenej podobný výsledok ukazujú PGAdb-builder a BPGA, ktoré sa líšia o 362 základných génov.

V rozdelení postrádateľného pangenómu sú najviac zhodné výsledky z panX a PGAdb-builder, ktoré v ňom našli 2 459 a 2 655 génov jednotlivo. Najviac ich našlo Roary s počtom 4 822 a najmenej BPGA s 968 postrádateľnými génmi.

Všetky nástroje teda vedia vyčísliť počet génov v základnom a postrádateľnom pangenóme. Vyčíslenie pre jedinečný pangenóm zahŕňajú PGAdb-builder a BPGA, ktoré sa ale v tejto hodnote veľmi líšia. Roary rozdeľuje postrádateľný pangenóm na jeho shell a cloud časť a taktiež určuje hodnotu pre soft-core pangenóm.

Vďaka možnosti získania informácií o génoch boli porovnané stopercentné výskyty génov z PGAdb-builder a Roary. Oba sa zhodujú v 512 génoch, vypísaných v prílohe B, pričom Roary zaradilo do základného genómu ďalších 510 génov, ktoré PGAdb-builder ako základné nepozná, a naopak, PGAdb-builder tak učinil v 311 prípadoch, znázornené na obr. 2.9.

Z grafických výstupov sa každý nástroj orientuje na niečo iné. Spoločné majú



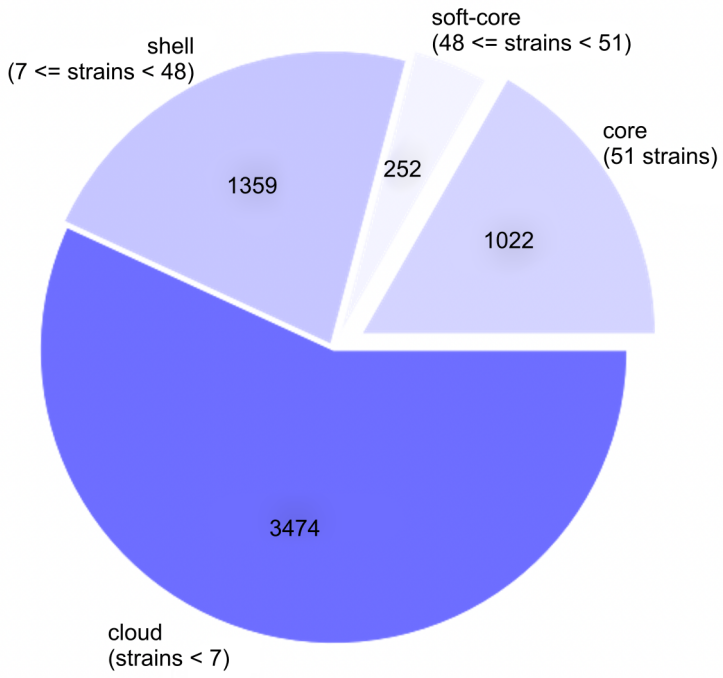
Obr. 2.6: Histogram COG distribúcie, získaný z BPGA.

vykreslenie fylogenetického stromu a nejakú formu koláčového grafu, ktorý reprezentuje rozloženie pangenómu, ten sa neobjavuje len u BPGA.

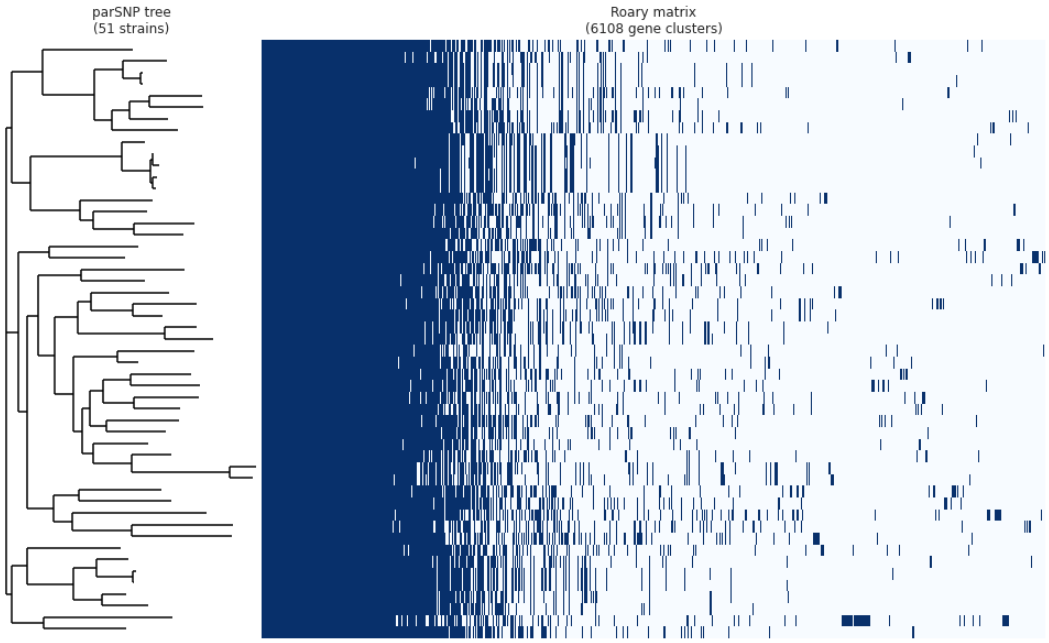
Úplne základnú analýzu vykonáva PGAdb-builder, ktorý vo výstupe ponúka len tieto zmienené grafické výstupy a tabuľku percentuálneho výskytu jednotlivých génov. Za nevýhodu považujem dlhšie spracovávanie dát než u ostatných nástrojov, čo môže robiť problémy pri analýzach veľkého počtu genómov.

panX okrem spomenutých grafov ponúka aj tabuľku s informáciami o nájdených génoch, ako napríklad v koľkých genómoch izolátov baktérie sú gény najdené alebo ich dĺžku. Nevýhodou tejto aplikácie je nemožnosť tieto dáta ľahko exportovať, pretože je potrebné sa v tabuľke preklikávať, a taktiež nemožnosť analyzovať iné genómy než tie, ktoré ponúka samotný nástroj.

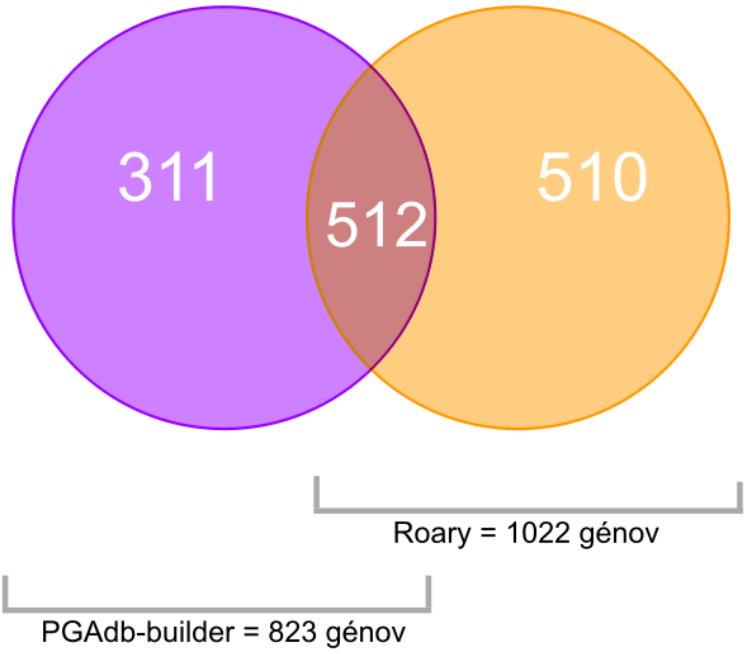
BPGA sa na rozdiel od ostatných nástrojov zameriava na funkčnú analýzu pangenómu. V grafických výstupoch ponúka COG a KEGG distribúcie aj pre jednotlivé časti pangenómu a rozdeľuje gény do skupín podľa ich funkcií. Číselné rozdelenie pangenómu je obsiahnuté v jednoduchej .xls tabuľke a pre jeho úplnosť je potrebné



Obr. 2.7: Koláčový graf znázorňujúci rozloženie génov v pangenóme podľa Roary.



Obr. 2.8: Fylogenetický strom a matica zhody izolátov, získané z Roary.



Obr. 2.9: Porovnanie počtu génov so stopercentným výskytom u PGAdb-builder a Roary.

ho dopočítať.

Roary ponúka tabuľku výskytu génov s rôznymi údajmi, ktoré vďaka ľahkej dostupnosti môžu slúžiť na ďalšiu analýzu. Veľkou výhodnou je, že pre každý gén je určené, v ktorom izoláte sa nachádza. Koláčový graf zase poskytuje informácie aj o cloud, shell a soft-core pangenóme, čo ostatné nástroje neanalyzujú. Zaujímavá je tiež matica vykresľujúca výskyt génov v izolátoch, ktorá ponúka doplňujúce informácie o ich podobnosti. Výhodou je rýchlosť analýzy.

2.3 Vlastný program get_pangenome

Vlastný program pre analýzu pangenómu bol vytvorený v programovacom jazyku Python. Pri tvorbe boli využité moduly os, re a Bio z ktorého boli importované vybrané položky SeqIO a pairwise2. Medzi hlavné body vlastného programu pre analýzu pangenómu patrí nahranie vhodných vstupných dát. Ďalej nasleduje zarovnanie vložených sekvencií podľa vybraného algoritmu. Pre vyhodnotenie zarovnania je vypočítaná hodnota p-distance, ktorá nám poskytuje informácie o zhode alebo nezhode génov. Gény alebo sekvencie považujeme za zhodné, ak je výsledkom p-distance hodnota nižšia ako 0,05, teda ak sú gény zhodné na 95%, naopak, hodnota menšia ako 95% predstavuje nerovnaké gény. Táto hodnota bola zvolená preto, že aj

voľne dostupné nástroje, ktoré boli analyzované v tejto práci pracujú pri zarovnávaní sekvencií práve s touto hodnotou zhody.

Výstupom programu je textový súbor vo formáte 'pangenome_file.txt', ktorý obsahuje informácie o počet génov v pangenóme a ich názvy. Príklad výstupného súboru je zobrazený na obrázku 2.10.

```
Výsledok: v pangenóme sa nachádza 1366 génov, tieto to sú:
ABC transporter, ATP-binding protein
conserved hypothetical protein
conserved hypothetical protein
transport protein
integrase core domain, putative
aminodeoxychorismate lyase
membrane protein, putative
hypothetical protein SPH_0658
pyruvate formate-lyase 3
neopullulanase
hypothetical protein SPH_1452
protein DltB
peptidase, M20/M25/M40 family
3-dehydroquinate synthase
conserved hypothetical protein
ATP synthase (C/AC39) subunit
UDP-N-acetylglucosamine--N-acetylmuramyl-(pentapeptide)pyrophosphoryl-undecaprenol N-acetylglucosamine transferase
transcriptional activator
galactoside O-acetyltransferase
conserved hypothetical protein
chlorohydrolase
phosphate-binding protein PstS 1 (PBP 1)
dihydroorotate dehydrogenase
ABC transporter, ATP-binding protein
hypothetical protein SPH_0604
cell wall surface anchor family protein
4-alpha-glucanotransferase
nitroreductase family protein
```

Obr. 2.10: Výstup programu zobrazený v textovom editore.

Bližší popis programu a jeho funkcií je uvedený v nasledujúcich kapitolách. Jednotlivé funkcie sú v závere volané a spúšťané v *main* funkcii. Vývojový diagram navrhnutého programu je v prílohe C.

2.3.1 Načítanie genómov

Funkcia *get_num_of_files* berie na vstup cestu k súboru s genómami vo FASTA formáte, ktoré obsahujú proteínové sekvencie prokaryotického organizmu.

Funkcia vo výstupe vráti zoznam s genómami na analýzu a ich počet, príklad je na obr. 2.11.

```
Genómy na analýzu sú tieto: ['GCA_000019265.1_ASM1926v1_protein.faa', 'GCA_000019825.1_ASM1982v1_protein.faa',
 'GCA_000019005.1_ASM1900v1_protein.faa', 'GCA_000019025.1_ASM1902v1_protein.faa', 'GCA_000018985.1_ASM1898v1_protein
 .faa', 'GCA_000019985.1_ASM1998v1_protein.faa', 'GCA_000006885.1_ASM688v1_protein.faa', 'GCA_000014365
 .2_ASM1436v2_protein.faa', 'GCA_000007045.1_ASM704v1_protein.faa', 'GCA_000018965.1_ASM1896v1_protein.faa']
Počet genómov na analýzu je: 10
```

Obr. 2.11: Príklad výstupu funkcie *get_num_of_files*.

2.3.2 Zarovnanie a výpočet p-distance

Pomocou funkcie *get_alignment* je zo zarovnania dvoch sekvencií získaná hodnota p-distance. Vstupom do funkcie sú 2 sekvencie, ktoré porovnávame.

V prvom kroku je podľa modulu Bio.pairwise prevedené globálne zarovnanie 2 sekvencií. Ako parametre sú zvolené skóre zhody 1 a penalizácia medzery 0. Tým pádom je výsledné skóre zarovnania rovné počtu zhodných znakov v zarovnaných sekvenciách. Z prevedeného globálneho zarovnania je získané celkové skóre a dĺžka zarovnania. Tieto premenné sú použité pri výpočte hodnoty p-distance, ktorá predstavuje podiel rozdielnych znakov, alebo mutácií k dĺžke sekvencie. V našom prípade sme p-distance nadefinovali ako

$$p - distance = (d - s)/d,$$

kde d = dĺžka zarovnania a s = skóre zarovnania.

Vo výstupe získame logickú hodnotu True, ak je hodnota p-distance nižšia alebo rovná 0,05, alebo logickú hodnotu False, ak je p-distance vyššia než 0,05. Hodnota 0,05 predstavuje 95% zhodu medzi sekvenciami a bola zvolená preto, že aj voľne dostupné nástroje, ktoré boli analyzované v tejto práci pracujú pri zarovnávaní sekvencií práve s touto hodnotou.

2.3.3 Hlavná analýza

Funkcia *analyses* do vstupu berie počet genómov a ich zoznam. Vytvorí prázdny zoznam, ktorý funguje ako vnorený a do ktorého sa následne vložia všetky sekvencie génov pre každý genóm, ktorý analyzujeme. Pred podrobnejším popisom funkcie je dôležité vysvetliť, že pri porovnávaní 2 sekvencií je pri zhodnom náleze záznam druhej sekvencie označený v zozname ako None.

Prechádza sa jednotlivými genómami v zozname a vyberie sa daný gén. Ak ten je označený ako None, prechádza sa na ďalší gén. Ak gén nie je označený ako None, vyberie sa z neho samotná sekvencia a prechádza sa ďalším genómom, z ktorého sa vyberie gén, a opäť sa kontroluje či je záznam označený ako None. Ak áno, prechádza sa na ďalší gén v druhom genóme, ak nie, vyberie sa z génu z druhého genómu samotná sekvencia a zavolá sa funkcia *get_alignment*, ktorá prevedie zarovnanie sekvencií a vypočíta hodnotu p-distance. Ak je na výstupe tejto funkcie logická hodnota True, teda sekvencie sú zhodné, na mieste génu druhého genómu bude záznam zamenený za None. A pokračuje sa na ďalší gén v druhom genóme. Takto sa postupne prejdú všetky gény z druhého genómu a po ukončení tohoto cyklu sa prechádza na druhý gén v prvom genóme a opäť sa porovnáva so všetkými génmi v druhom genóme. Postupne sa týmto zarovnajú gény metódou každý-proti-každému.

Po prebehnutí všetkých génov vo všetkých genómoch zoznam genes obsahuje aj hodnoty None, ktoré sú v ďalšom kroku odstránené a vo výsledku sú na výstupe funkcie *analyses* získané záznamy génov, ktoré tvoria samotný pangenóm.

2.3.4 Úprava názvov

Keďže názvy génov obsahujú okrem samostatného názvu aj identifikátory jedinečné pre každý genóm, funkcia *pangenome* tieto identifikátory spred aj za názvom génu odstráni, znázornené v príklade na obr. 2.12.

```
Pred úpravou: ['AAK74194.1 chromosomal replication initiator protein DnaA [Streptococcus pneumoniae TIGR4]', 'AAK74195.1
 DNA polymerase III, beta subunit [Streptococcus pneumoniae TIGR4]', 'AAK74196.1 hypothetical protein SP_0003
 [Streptococcus pneumoniae TIGR4]']
Po úprave: ['chromosomal replication initiator protein DnaA', 'DNA polymerase III, beta subunit', 'hypothetical protein
 SP_0003']
```

Obr. 2.12: Príklad výstupu funkcie *pangenome*.

2.3.5 Vytvorenie výstupného súboru

Pre viac prehľadný výsledok je funkciou *output_file* vytvorený textový súbor, ktorý obsahuje na začiatku správu, koľko génov tvorí pangenóm a následne sú všetky názvy génov vypísané.

2.4 Testovanie programu get_pangenome

Z dôvodu veľkého množstva dát na analýzu a komplikovanosť vlastného algoritmu, ktorý je výpočtovo veľmi náročný, nebolo možné vlastnou metódou otestovať program na tom istom datasete, ktorý sa použil pri testovaní jednotlivých nástrojov. Preto bol tento dataset redukovaný, aby bolo možné ukázať aspoň nejaký výstup, ktorý program *get_pangenome* ponúka. Zredukovaný dataset obsahoval prvých 25 genómov z pôvodného datasetu. Genómy obsahovali prvých 100 génov z ich záznamov. Po prevedení analýzy sme vo výstupnom textovom súbore získali informácie o počte génov v pangenóme aj ich výpis. Pangenóm z redukovaného datasetu tvorí celkovo 1366 génov, ukážka výstupného súboru je na obr. 2.10.

Vlastný algoritmus by sa dal vylepšiť napríklad rozdelením celého datasetu na menšie časti a aplikáciou metódy rozdeľuj a panuj. V menších častiach by prebiehala analýza paralelne a následne by výsledky z menších častí boli porovnávané a analyzované medzi sebou. Program by sa dal doplniť aj o vyčíslenie jednotlivých častí pangenómu, vykreslenie koláčových grafov alebo fylogenetických stromov.

Záver

Obsahom tejto bakalárskej práce je popis nástrojov na analýzu pangenómu pre bakteriálne populácie a návrh vlastnej metódy.

V prvej polovici teoretickej časti práce sú popísané poznatky o genóme a pangenóme s dôrazom na ich funkciu a zmysel u bakteriálnych populácií.

Druhá polovica sa zameriava na popis nástrojov, ktoré sa využívajú na analýzu bakteriálneho pangenómu. Vybrané boli štyri nástroje, a to panX, PGAdb-builder, BPGA a Roary. Prvé dva nástroje boli vybrané z dôvodu ľahkej dostupnosti na webových serveroch, BPGA a Roary je potrebné nainštalovať do počítača, ale vyznačujú sa rýchlosťou analýzy. Spoločným znakom vybraných nástrojov je jednoduché ovládanie. Ich otestovanie slúžilo na porovnanie ich vlastností, funkcií, vstupov a výstupov z jednotlivých programov.

Praktická časť sa zameriava hlavne na popis výsledkov jednotlivých nástrojov a zobrazené sú aj grafické výstupy. Testovanie nástrojov bolo prevedené na genómoch patogénnej baktérie *Streptococcus pneumoniae*, pričom pre panX bolo použitých 52 genómov jej izolátov a pre zvyšné tri nástroje bolo použitých 51 genómov. Zmena bola vykonaná z dôvodu odstránenia jedného genómu z RefSeq databázy.

Porovnaním výsledkov štyroch vybraných nástrojov na analýzu pangenómu môžeme povedať, že výstupy sa líšia, ale zároveň si sú podobné. Každý nástroj sa vo výstupe zameriava na niečo iné a ponúka rôzne metódy spracovania.

Druhá polovica praktickej časti je zameraná na vlastnú analýzu pangenómu. Algoritmus bol vytvorený v programovacom prostredí Python. V práci sú popísané jednotlivé funkcie, ktoré v skripte figurujú, ich vstupy aj výstupy. Algoritmus spočíva v spracovaní vstupných dát, zarovnaní sekvencií každá-proti-každej, a vo výpočte hodnoty p-distance, podľa ktorej sú gény ukladané do pangenómu. Následne sú výsledky upravené pre lepšie hodnotenie a uložené do výstupného textového súboru.

Vzhľadom na výpočtovú náročnosť algoritmu a veľké množstvo vstupných dát bolo testovanie prevedené na zredukovanom datasete, aby bolo jednoznačné, že program funguje. Tento dataset obsahoval prvých 25 genómov z pôvodného datasetu, ktoré mali po 100 génov. Vo výsledku sme získali textový súbor s názvami génov v pangenóme a ich počet, ktorý program vyčíslil na 1 366 génov.

Na záver sa dá povedať, že výsledky testovaných nástrojov vykazujú najväčšiu podobnosť v číselných hodnotách pre základný pangenóm. V celkovej veľkosti pangenómu a jeho ostatných častí sa výsledky líšia. Grafické výstupy nástrojov sú v základe podobné ale každý má svoje špecifiká. Podľa rýchlosti spracovania dát a grafického či tabuľkového výstupu by som ako najlepší nástroj na analýzu pangenómu vyhodnotila Roary. Nástroj je ľahko ovládateľný, ponúka jednoduché grafické výstupy a zaradenie nájdených génov je najviac presné.

Nejednoznačnosť výsledkov jednotlivých nástrojov svedčí o motivácii vytvárať nové nástroje na analýzu pangenómu, alebo vylepšovať tie už známe.

Literatúra

[1] GOLDMAN, Aaron David, Laura F. LANDWEBER a W. Ford DOOLITTLE. What Is a Genome? *PLOS Genetics* [online]. 2016, **12**(7) [cit. 2021-11-22]. ISSN 1553-7404. Dostupné z: doi:10.1371/journal.pgen.1006181

[2] BROWN, TA, 2002. *Genomes. 2nd edition* [online]. Oxford: Wiley-Liss. Chapter 1, The Human Genome [cit. 2021-12-09]. Dostupné z: https://www.ncbi.nlm.nih.gov/books/NBK21134/

[3] DUFAYARD, J.-F., L. DURET, S. PENEL, M. GOUY, F. RECHENMANN a G. PERRIERE. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* [online]. 2005, **21**(11), 2596-2603 [cit. 2022-01-06]. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/bti325

[4] LERAT, E. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Research* [online]. 2005, **33**(10), 3125-3132 [cit. 2022-01-03]. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gki631

[5] BOBAY, Louis-Marie a Howard OCHMAN. The Evolution of Bacterial Genome Architecture. *Frontiers in Genetics* [online]. 2017, **8** [cit. 2022-01-01]. ISSN 1664-8021. Dostupné z: doi:10.3389/fgene.2017.00072

[6] SALTON, Milton R.J. a KIM Kwang-Shin, 1996. *Medical Microbiology. 4th edition* [online]. Galveston: University of Texas Medical Branch at Galveston. Chapter 2, Structure [cit. 2021-12-10]. Dostupné z: https://www.ncbi.nlm.nih.gov/books/NBK8477/

[7] VIDYASAGAR, Aparna a PAPPAS Stephanie, 2021. What are bacteria?. In: *livescience.com*[online]. 14.10. [cit. 2021-12-10]. Dostupné z: https://www.livescience.com/51641-bacteria.html

[8] YANG, Desirée C., Kris M. BLAIR a Nina R. SALAMA. Staying in Shape: the Impact of Cell Shape on Bacterial Survival in Diverse Environments. *Microbiology and Molecular Biology Reviews* [online]. 2016, **80**(1), 187-203 [cit. 2021-12-10]. ISSN 1092-2172. Dostupné z: doi:10.1128/MMBR.00031-15

[9] Microbiology Society. *Bacteria* [online]. [cit. 2021-12-10]. Dostupné z: https://microbiologysociety.org/why-microbiology-matters/what-is-microbiology/bacteria.html

[10] CHIEN, An-Chun, Norbert S. HILL a Petra Anne LEVIN. Cell Size Control in Bacteria. *Current Biology* [online]. 2012, **22**(9), R340-R349 [cit. 2021-12-10]. ISSN 09609822. Dostupné z: doi:10.1016/j.cub.2012.02.032

[11] CHACONAS, George a Carton W. CHEN. Replication of Linear Bacterial Chromosomes: No Longer Going Around in Circles. HIGGINS, N. Patrick, ed. *The Bacterial Chromosome* [online]. Washington, DC, USA: ASM Press, 2004, 2014-04-30, s. 525-539 [cit. 2021-12-31]. ISBN 9781683672043. Dostupné z: doi:10.1128/9781555817640.ch29

[12] LAND, Miriam, Loren HAUSER, Se-Ran JUN, et al. Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics* [online]. 2015, **15**(2), 141-161 [cit. 2022-01-01]. ISSN 1438-793X. Dostupné z: doi:10.1007/s10142-015-0433-4

[13] HOU, Yubo, Senjie LIN a Rosemary Jeanne REDFIELD. Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes. *PLoS ONE* [online]. 2009, **4**(9) [cit. 2022-01-01]. ISSN 1932-6203. Dostupné z: doi:10.1371/journal.pone.0006978

[14] BROOKS, Lavida R. K. a George I. MIAS. Streptococcus pneumoniae-s Virulence and Host Immunity: Aging, Diagnostics, and Prevention. *Frontiers in Immunology* [online]. 2018, **9** [cit. 2022-01-01]. ISSN 1664-3224. Dostupné z: doi:10.3389/fimmu.2018.01366

[15] TETTELIN, H., V. MASIGNANI, M. J. CIESLEWICZ, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome."*Proceedings of the National Academy of Sciences* [online]. 2005, **102**(39), 13950-13955 [cit. 2021-12-10]. ISSN 0027-8424. Dostupné z: doi:10.1073/pnas.0506758102

[16] VERNIKOS, George, Duccio MEDINI, David R RILEY a Hervé TETTELIN. Ten years of pan-genome analyses. *Current Opinion in Microbiology* [online]. 2015, **23**, 148-154 [cit. 2021-12-10]. ISSN 13695274. Dostupné z: doi:10.1016/j.mib.2014.11.016

[17] LIVINGSTONE, Paul G., Russell M. MORPHEW a David E. WHITWORTH. Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. *Frontiers in Microbiology* [online]. 2018, **9** [cit. 2021-12-10]. ISSN 1664-302X. Dostupné z: doi:10.3389/fmicb.2018.03187

[18] MARRONI, Fabio, Sara PINOSIO a Michele MORGANTE. Structural variation and genome complexity: is dispensable really dispensable? *Current Opinion in Plant Biology* [online]. 2014, **18**, 31-36 [cit. 2021-12-10]. ISSN 13695266. Dostupné z: doi:10.1016/j.pbi.2014.01.003

[19] MEDINI, Duccio, Claudio DONATI, Hervé TETTELIN, Vega MASIGNANI a Rino RAPPUOLI. The microbial pan-genome. *Current Opinion in Genetics & Development* [online]. 2005, **15**(6), 589-594 [cit. 2021-12-10]. ISSN 0959437X. Dostupné z: doi:10.1016/j.gde.2005.09.006

[20] XIAO, Jingfa, Zhewen ZHANG, Jiayan WU a Jun YU. A Brief Review of Software Tools for Pangenomics. *Genomics, Proteomics & Bioinformatics* [online]. 2015, **13**(1), 73-76 [cit. 2022-01-02]. ISSN 16720229. Dostupné z: doi:10.1016/j.gpb.2015.01.007

[21] DING, Wei, Franz BAUMDICKER a Richard A NEHER. PanX: pan-genome analysis and exploration. *Nucleic Acids Research* [online]. 2018, **46**(1), e5-e5 [cit. 2022-01-02]. ISSN 0305-1048. Dostupné z: doi:10.1093/nar/gkx977

[22] SEEMANN, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* [online]. 2014, **30**(14), 2068-2069 [cit. 2022-01-06]. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/btu153

[23] PAGE, Andrew J., Carla A. CUMMINS, Martin HUNT, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* [online]. 2015, **31**(22), 3691-3693 [cit. 2022-01-06]. ISSN 1367-4803. Dostupné z: doi:10.1093/bioinformatics/btv421

[24] LIU, Yen-Yi, Chien-Shun CHIOU a Chih-Chieh CHEN. PGAdb-builder: A web service tool for creating pan-genome allele database for molecular fine typing. *Scientific Reports* [online]. 2016, **6**(1) [cit. 2022-01-04]. ISSN 2045-2322. Dostupné z: doi:10.1038/srep36213

[25] CHEN, Xinyu, Yadong ZHANG, Zhewen ZHANG, et al. PGAweb: A Web Server for Bacterial Pan-Genome Analysis. *Frontiers in Microbiology* [online]. 2018, **9** [cit. 2022-01-05]. ISSN 1664-302X. Dostupné z: doi:10.3389/fmicb.2018.01910

[26] CHAUDHARI, Narendrakumar M., Vinod Kumar GUPTA a Chitra DUTTA. BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports* [online]. 2016, **6**(1) [cit. 2022-01-05]. ISSN 2045-2322. Dostupné z: doi:10.1038/srep24373

Zoznam symbolov a skratiek

bp	párov báz
C-G	cytozín - guanín
COG	zhluky ortológnych génov
DNA	deoxyribonukleová kyselina
KEGG	Kyotská databáza génov a genómov
kbp	kilo párov báz
Mbp	mega párov báz
mRNA	mitochondriálna ribonukleová kyselina
NCBI	National Center for Biotechnology Information
SNP	jednonukleotidový polymorfizmus
UPGMA	metóda neváženého párovania s aritmetickým priemerom
wgMLST	typizácia multilokusovejsekvencie celého genómu
WGS	celogenómové sekvenovanie

A Tabuľky

Tab. A.1: Zoznam informácií o použitých genómoch *S. pneumoniae*.

Označenie sekvenačných dát	**Identifikátor**	**Izolát**	**Dĺžka [Mbp]**	**Gény**	**Databáza**
GCA_000006885 .1_ASM688v1	NC_003028	TIGR4; ATCC BAA-334	2,16	2176	GenBank
GCA_000007045 .1_ASM704v1	NC_003098	R6	2,04	2071	GenBank
GCA_000014365 .2_ASM1436v2	NC_008533	D39	2,05	2075	GenBank
GCA_000018965 .1_ASM1896v1	NC_012468	70585	2,18	2247	GenBank
GCA_000018985 .1_ASM1898v1	NC_012466	JJA	2,12	2169	GenBank
GCA_000019005 .1_ASM1900v1	NC_012467	P1031	2,11	2187	GenBank
GCA_000019025 .1_ASM1902v1	NC_012469	Taiwan19F-14	2,11	2153	GenBank
GCA_000019265 .1_ASM1926v1	NC_010380	Hungary19A-6	2,25	2297	GenBank
GCA_000019825 .1_ASM1982v1	NC_011072	G54	2,08	2107	GenBank
GCA_000019985 .1_ASM1998v1	NC_010582	CGSP14	2,21	2220	GenBank
GCA_000146975 .1_ASM14697v1	NC_014494	AP200	2,13	2182	GenBank
GCA_000147095 .1_ASM14709v1	NC_014498	670-6B	2,24	2328	GenBank
GCA_000196595 .1_ASM19659v1	NC_014251	TCH8431/19A	2,09	2137	GenBank
GCA_000210935 .1_ASM21093v1	NC_017593	INV200	2,09	2105	GenBank
GCA_000210955 .1_ASM21095v1	NC_017592	OXC141	2,04	2114	GenBank
GCA_000210975 .1_ASM21097v1	NC_017591	INV104	2,14	2195	GenBank

pokračovanie na ďalšej strane

Tab. A.2: pokračovanie z predchádzajúcej strany

Označenie sekvenačných dát	**Identifikátor**	**Izolát**	**Dĺžka [Mbp]**	**Gény**	**Databáza**
GCA_000210995 .1_ASM21099v1	NC_021006	SNP034156	2,02	2085	GenBank
GCA_000211015 .1_ASM21101v1	NC_021028	SNP034183	2,04	2114	GenBank
GCA_000211035 .2_ASM21103v2	NC_021026	SNP994038	2,03	2106	GenBank
GCA_000211055 .2_ASM21105v2	NC_021005	SNP994039	2,03	2106	GenBank
GCA_000211075 .1_ASM21107v1	NC_021003	SPN032672	2,13	2193	RefSeq
GCA_000211095 .1_ASM21109v1	NC_021004	SPN033038	2,13	2203	vymazaný
GCA_000251085 .2_ASM25108v2	NC_017769	ST556	2,15	2203	GenBank
GCA_000299015 .1_ASM29901v1	NC_018630	gamPNI0373	2,06	2130	GenBank
GCA_000817005 .1_ASM81700v1	NZ_CP007593	NT_110_58	2,29	2282	GenBank
GCA_001255215.1_ S_pneumo_A66_v1	NZ_LN847353	A66	1,98	2068	GenBank
GCA_001457635 .1_NCTC7465	NZ_LN831051	NCTC7465	2,11	2170	GenBank
GCA_001896045 .1_ASM189604v1	NZ_CP018136	SP49	2,21	2287	GenBank
GCA_001896065 .1_ASM189606v1	NZ_CP018137	SP61	2,07	2109	GenBank
GCA_001896085 .1_ASM189608v1	NZ_CP018138	SP64	2,07	2108	GenBank
GCA_001902455 .1_ASM190245v1	NZ_CP018347	SWU02	2,09	2126	RefSeq
GCA_002357995 .1_ASM235799v1	NZ_AP017971	KK0981	2,15	2200	GenBank
GCA_002813535 .1_ASM281353v1	NZ_CP025076	19F	2,13	2147	GenBank
GCA_002813955 .1_ASM281395v1	NZ_CP018838	11A	2,08	2134	GenBank

pokračovanie na ďalšej strane

Tab. A.3: pokračovanie z predchádzajúcich strán

Označenie sekvenačných dát	**Identifikátor**	**Izolát**	**Dĺžka [Mbp]**	**Gény**	**Databáza**
GCA_002843545 .1_ASM284354v1	NZ_CP025256	Xen35	2,17	2175	GenBank
GCA_002947575 .1_ASM294757v1	NZ_CP026670	335	2,22	2254	GenBank
GCA_003003495 .1_ASM300349v1	NZ_CP027540	D39V	2,05	2077	GenBank
GCA_003351525 .1_ASM335152v1	NZ_CP031247	M23734	2,11	2184	RefSeq
GCA_003351645 .1_ASM335164v1	NZ_CP031245	M16808	2,1	2141	RefSeq
GCA_003351705 .1_ASM335170v1	NZ_CP031248	M26365	2,17	2255	RefSeq
GCA_003351725 .1_ASM335172v1	NZ_CP031246	M26368	2,11	2139	RefSeq
GCA_003354825 .1_ASM335482v1	NZ_CP025838	SPN XDR SMC1710-32	2,06	2120	GenBank
GCA_003609935 .1_ASM360993v1	NZ_AP018391	MDRSPN001	2,05	2067	GenBank
GCA_900475305 .1_43721_E02	NZ_LS483374	NCTC7466	2,05	2075	GenBank
GCA_900475515 .1_42731_F01	NZ_LS483390	NCTC13276	2,04	2068	GenBank
GCA_900475805 .1_46514_B03	NZ_LS483417	NCTC11902	2,09	2136	GenBank
GCA_900476445 .1_55312_D02	NZ_LS483448	4041STDY 6836167	2,13	2204	RefSeq
GCA_900476455 .1_55312_E02	NZ_LS483523	4041STDY 6836169	2,15	2170	RefSeq
GCA_900476465 .1_55312_F02	NZ_LS483449	4041STDY 6836170	2,1	2125	RefSeq
GCA_900476475 .1_49768_F01	NZ_LS483450	4041STDY 6583227	2,16	2198	RefSeq
GCA_900476505 .1_55312_C02	NZ_LS483451	4041STDY 6836166	2,2	2245	RefSeq
GCA_000026665 .1_ASM2666v1	NC_011900	ATCC 700669	2,22	2251	GenBank

B Výpis zhodných génov z nástrojov PGAdb-builder a Roary

Pri porovnaní výsledkov z PGAdb-builder a Roary boli ako zhodné gény označené tieto: trpF, thrC, tag, bceB, prmC, niaX, leuD, thiN, purE, purK, bcrC_1, thiM_2, rsmI, coaE, trmK, recX, rsmE, dexB, valS, uvrC, miaA, rpiA, iscS_2, proA, smc, msbA_2, arnB, wcaJ, malX, mngB, glpF, dnaG, hcnC, divIB, mnmA, pbuO, sulD, rnhC, rimP, rbfA, pepV, metG, folD, nifU, dnaE, trmFO, pyrK, mutM, cpoA, rplU, yfnB, truB, crcB, ogt, yvgN, bglK_1, trpA, galK, msmE, ndk, acyP, yfkJ, tcyJ, mtcA1, agaS, tcyA, metP, rplF, rpsM, glmS, fabM, fabZ, yecS_1, ribU, hrcA, rnjB, gatC_2, clpP, gor, yhbU_1, rodA, rpsT, coaA, tmpC, mglC, tmk, frr, glmU, pepT, xylH, rpmA, yjjK, ligA, pyrC, yidA_3, aroE, rlmI, murB, suhB, atpG, patB, murE, plsC, graS, graR, ileS, ytrB, scrB, nadD, yheH, xpt, galT, proV, pheT_2, endA, licR_2, rpsB, rny, trpE, gap, dnaA, purA, tadA, dut, purC, purB, frlR, trkA, tsaD, dapE, ruvA, cls, rplB, rplX, rplR, rpmD, infA, rpsK, ruvB, uppS, rpsG, pepS, rplM, kdgK, ugl, fni, accA, rpoE, ecsA_1, infB, metE, cysE, cysS, rplK, msrAB_1, hflX, murG, ykoE, upp, prfB, rpsP, nrdD_2, bioY2, deoB, punA, plsY, ilvE, ftsK, macB_3, nspC, aguA, yidA_3, pyrH, gloA, ybeY, dgkA, era, rnr, gpmA_1, ybaK, mtnN, dapA, tdk, glyA, prs, gapN, eno, addA, ptsH, lacB, lacA, lepA, pphA, tlyA, udk, gyrA, yycF, panT, carB, pyrR, ffh, crcB, rplJ, pheA, aroB, potB, rpsU, ydaF_5, pdxS, apbE, atpE, metI, ytgP, yugI, rpsF, yghU, dapB, glmM, tpiA, dnaD, metA, apt, pdxK, truA, ywnA, pepO, tpx, ftsA, recR, acpS, nrdR, trpC, malK_2, pnuC, scpB, ydfG, rsmD, ybhL, rnmV, nadC, gmuA, ulaD, ulaB, ulaA_1, rnpA, ackA, nagA, pgi, patB, ykuR, malR, aspS, rpmF, hslO, ctsR, purH, glcK, ftsX, ydhF, parC, ydaF_4, clcA, nagB, queA, yumC, psaA, amiA, pdxT, scpA, mtaD, pepA_1, rplI, rpmB, rplV, rpsS, rplP, hslR, rplN, rpmI, rplE, rpsQ, acpP, mreC, proS, ychF, glpK, tsf, fabG, rpsI, fabH, mfd, groL, rplA, pgsA, hpdA, rplO, rluB, rnz, phoR, ftsW, pyrG, parE, lspA, adhE, ilvB, gyrB, ilvA, ribF, rex, rplQ, pbpX_2, rplD, nrdF, trxB, dnaK, trkA, ftsE, ulaA_2, asnA, rpsH, potA, map, atpA, rpoZ, rplC, aroF, ilvD, rplT, ansB, luxS, relA, galE, yheS_2, sodA, rpsC, atpH, tilS, rpoC, rplS, ppc, rpoA, saeR, pepN, gldA, fmt, ppaC, atpF, dnaN, fabF, yxeN, atpB, rpsJ, pta, pyrB, ycjO, saeS, ugpQ, ftsL, ktrB, atpD, fba, pgk, trpD, hemN, pflB, yjbM, yhbY, ftsH, asd, fabD, pfkA, efp, metK, pepC, guaB, pcrA, murC, exoA, carA, trmB, dltC, accC, artP_2, cmk, gluP, cdsA, yigZ, brnQ, speE, rsmB, pepQ, trmD, rpsE, rpsL, treR, xseB, rplW, rpsO, ethR, fepD_2, purL, lysS, leuS, accB, rexB, ulaC, adk, gnd, metN, tuf, cysB, gatB, purM, nudF, folE, ppnK, pyrE, rpsR, cysK, gatA, htrA, topA, ftsZ, recF, rpmC, purF, rnc, ldh, argS, gdhA, yutF, engB, pth, mutS2, alaS, yesO_2, uvrA, hpt, xseA, livF, trpB, ilvC, rlmL, prfA, sorC, tyrS, tabA_2,

gpsA, polA, rplGA, glyQ, recN, aroC, smpB, thyA, pyrF, clpX, guaC, azr_1, thiD, birA, accD, yclQ, czcD, malQ, manZ_1, yajL, gph, plsX, gla, murD, mecA, metQ, codY, zwf, alr, glyS, speA, gltX, metF, recO, trxA, hprK, ycfH, guaA, greA, mrnC, mscL, purN, murI, aroD, glnQ_1, hmpT, thiI, hisS, dnaJ, recG, ezrA, gmuB, lrp, rarA, gmk, tig, rebM, lgt, feuB, thrS, adhR, tsaB, murF, ybjI, yheS_1.

C Vývojový diagram programu