Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Detekcia bakteriálnych plazmidov pomocou grafových neurónových sietí

Diplomová práca

2024
Bc. Veronika Tordová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Detekcia bakteriálnych plazmidov pomocou grafových neurónových sietí

Diplomová práca

Študijný program:	Informatika
Študijný odbor:	Informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	doc. Mgr. Bronislava Brejová, PhD.

Bratislava, 2024
Bc. Veronika Tordová

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Veronika Tordová
Študijný program: informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Detekcia bakteriálnych plazmidov pomocou grafových neurónových sietí
Detection of bacterial plasmids using graph neural networks

Anotácia: Plazmidy sú krátke DNA molekuly, ktoré sú často zodpovedné za šírenie antibiotickej rezistencie v baktériách. Program plASgraph2 využíva grafové neurónové siete na detekciu plazmidov v grafoch, ktoré vznikli spracovaním dát zo sekvenovania DNA bakteriálnej vzorky. Vrcholy grafu zodpovedajú úsekom sekvencie DNA a hrany možným následnostiam medzi nimi. Cieľom práce je rozšíriť tento program zakomponovaním zdrojov informácií, ktoré v ňom doteraz neboli využívané. Prvý zdroj je informácia o tom, ku ktorému koncu sekvencie sa viažu jednotlivé hrany. Druhý zdroj je informácia o podobnosti sekvencií v grafe so známymi proteínmi, ktoré sa môžu preferenčne vyskytovať v plazmidoch alebo mimo nich.

Vedúci: doc. Mgr. Bronislava Brejová, PhD.
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: prof. RNDr. Martin Škoviera, PhD.

Dátum zadania: 13.12.2021

Dátum schválenia: 03.01.2022 prof. RNDr. Rastislav Kráľovič, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie: Ďakujem školiteľke mojej diplomovej práce doc. Mgr. Bronislave Brejovej, PhD. za jej cenné rady, vedenie práce, trpezlivosť, konzultácie a čas.

Abstrakt

Plazmidy sú krátke molekuly DNA zodpovedné predovšetkým za šírenie rezistencie voči antibiotikám medzi baktériami. V našej práci sme rozšírili vstupné parametre nástroja plASgraph2, ktorý slúži na detekciu plazmidov v grafoch zostavenia genómu pochádzajúcich zo sekvenovania bakteriálnych izolátov. V prvej časti práce sme pripravili niekoľko možností architektúry programu založených na šírení informácie o tom, ktoré konce DNA sekvencií spolu v grafe zostavenia genómu susedia. V druhej časti sme k parametrom programu pridali vlastnosti založené na podobnosti vstupných sekvencií a proteínových domén. Výsledky oboch prístupov sme vyhodnotili na dátach pochádzajúcich z patogénu *E. faecium* a skupiny patogénov ESKAPEE. Oba prístupy pomohli zlepšiť výsledky klasifikácie oproti pôvodnej verzii programu, ale prístup založený na homológii priniesol výraznejšie zlepšenie.

Kľúčové slová: plazmid, grafové neurónové siete, klasifikácia, extremity, homológia, plASgraph2

Abstract

Plasmids are small DNA molecules primarily responsible for the spread of antibiotic resistance among bacteria. In our work, we extended the input parameters of the plASgraph2 tool designed to detect plasmids in assembly graphs derived from sequencing bacterial isolates. In the first part of the work we explored several options for extending the program architecture based on the propagation of information about which ends of the DNA sequences are adjacent to each other in the genome assembly graph. In the second part, we added features based on the similarity of input sequences and protein domains. We evaluated the results of both approaches on data derived from pathogen *E. faecium* and the ESKAPEE group of pathogens. Both approaches helped to improve the results of the classification in comparison with the original version of the program, but the homology-based approach achieved more significant improvement.

Keywords: plasmid, graph neural networks, classification, extremities, homology, plASgraph2

x

Obsah

Úvod 1

1 Súčasný stav problematiky 3
1.1 Chromozómy a plazmidy . . . 3
1.1.1 Význam plazmidov . . . 4
1.2 Zostavenie genómu . . . 4
1.3 Klasifikácia bakteriálnych a plazmidových kontigov . . . 5
1.3.1 Klasifikátory založené na homológii . . . 6
1.3.2 Klasifikátory používajúce podreťazce dĺžky k . . . 9
1.3.3 Klasifikátory používajúce informáciu zo susedných vrcholov . . . 10
1.4 Grafové neurónové siete . . . 10
1.4.1 Konvolučné grafové neurónové siete . . . 11
1.5 plASgraph2 . . . 12
1.5.1 Vyhodnocovanie úspešnosti . . . 14

2 Extremity 17
2.1 Graf zostavenia genómu . . . 17
2.2 Návrh architektúry . . . 18
2.3 Testovanie architektúry na syntetických dátach . . . 20
2.4 Výsledky . . . 23
2.4.1 Štatistika extremít v trénovacích dátach . . . 26

3 Homológia 29
3.1 Log-odd skóre . . . 29
3.2 Naša implementácia . . . 30
3.3 Výsledky . . . 31
3.4 Spojenie extremít a homológie . . . 36

Záver 37

Príloha 47

Úvod

Plazmid je krátka molekula DNA oddelená od chromozómovej DNA vyskytujúca sa najčastejšie v baktériách, ktorá sa dokáže šíriť medzi bunkami. Vďaka plazmidom sa môžu šíriť gény, ktoré môžu poskytnúť baktériám výhody pre prežitie v meniacom sa prostredí a pri adaptácii na nové podmienky. Medzi najznámejšie gény patrí skupina génov nesúcich rezistencie voči antibiotikám [13, 42].

Existuje niekoľko nástrojov na identifikáciu plazmidov v dátach pochádzajúcich z bakteriálnej alebo metagenomickej vzorky [6, 46, 4, 52, 47, 48]. Vstupom pre tieto nástroje sú grafy zostavenia genómu, ktoré vznikli poskladaním sekvenačných dát. Vrcholy v týchto grafoch sú sekvencie a hrany predstavujú možnú následnosť medzi nimi.

Väčšina z týchto nástrojov klasifikuje vrcholy iba ako samostatné jednotky a nepozerá sa na ostatné vrcholy v grafe. Nástroj plASgraph2 [56] je založený na použití grafovej konvolučnej vrstvy v neurónovej sieti, pomocou ktorej je možné šíriť informácie medzi susednými vrcholmi v grafe.

Cieľom práce je rozšíriť program plASgraph2 tak, aby pracoval s ďalšími dostupnými informáciami, ktoré v ňom pôvodne neboli využité. Program chceme rozšíriť zakomponovaním informácie o tom, ku ktorému koncu sekvencie sa viažu jednotlivé hrany. Ďalším zdrojom je informácia o podobnosti sekvencií v grafe so známymi proteínmi. Tieto proteíny sa môžu preferenčne vyskytovať v plazmidoch alebo mimo nich.

V prvej kapitole sme definovali základné pojmy ako je plazmid, zostavenie genómu a vysvetlili sme grafové konvolučné siete. Venovali sme sa existujúcim nástrojom na klasifikáciu plazmidov a popísali sme ich metódy. Popísali sme architektúru nástroja plASgraph2 a jeho výsledky v porovnaní s existujúcimi štúdiami. V druhej kapitole sme popísali zapracovanie informácie z toho, ku ktorým koncom sekvencií sa viažu jednotlivé hrany. Našu implementáciu sme otestovali na syntetických dátach, ktoré ilustrujú fungovanie našej modifikácie aj na reálnych biologických sekvenciách. V tretej kapitole sme sa venovali podobnosti sekvencií s proteínovými doménami a ako sme tieto podobnosti implementovali do programu plASgraph2. Pozreli sme sa aj na charakteristiky kontigov, ktorým homológia pomohla. Oba prístupy sme vyhodnotili na grafoch zostavenia genómu patogénnej baktérie *Enterococcus faecium* a skupiny patogénov ESKAPEE.

Kapitola 1

Súčasný stav problematiky

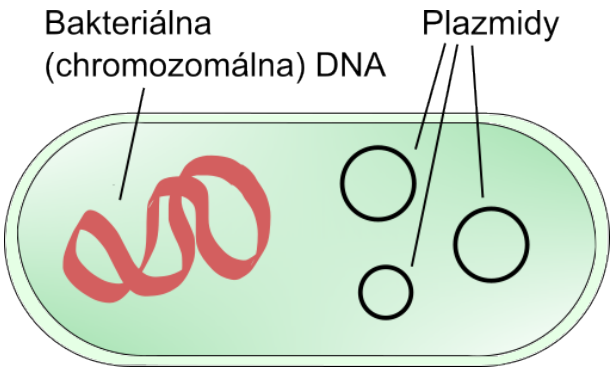
V tejto kapitole zavedieme základné biologické pojmy, ktoré budeme v práci používať. Najprv sa budeme venovať vysvetleniu plazmidu a jeho významu. Definujeme pojem skladania genómov. Popíšeme si existujúce metódy na klasifikáciu kontigov. Ďalej sa pozrieme na grafové neurónové siete a konvolučné grafové neurónové siete. Na záver kapitoly popíšeme program plASgraph2 [56], ktorého parametre budeme v práci rozširovať.

1.1 Chromozómy a plazmidy

Genóm je súbor všetkého genetického materiálu konkrétneho organizmu v bunke [30]. Je zložený z DNA (z ang. deoxyribonucleic acid), prípadne z RNA (z ang. ribonucleic acid) u RNA vírusov. Genóm je zväčša organizovaný do štruktúr nazývaných chromozómy.

Chromozóm je vláknitá štruktúra pozostávajúca z proteínov a dvojvláknovej molekuly DNA [29]. Pomocou DNA je genetická informácia prenášaná do dcérskych buniek. V práci sa budeme venovať baktériám. V baktériách má chromozóm cirkulárny tvar. Okrem chromozómov môže byť nositeľom genetickej informácie aj molekula nazývaná plazmid.

Plazmid je malá, do kruhu uzavretá dvojvláknová molekula DNA. Nachádza sa vo vnútri bunky, je oddelená od chromozómovej DNA a vie sa od nej nezávisle rozmnožovať [31]. Ilustráciu môžeme vidieť na obrázku 1.1. Plazmid je často kratší ako chromozómová DNA [16]. Môže dosahovať dĺžku od 10 000 do 400 000 bázových párov [8], kým bakteriálne chromozómy majú zvyčajne dĺžku okolo 5 miliónov bázových párov [36]. Vyskytuje sa najmä u baktérií, ale je možné ju nájsť aj v iných prokaryotických organizmoch alebo v niektorých eukaryotických organizmoch. V jednej baktérii sa môže naraz nachádzať viacero kópií rôznych plazmidov. Počas rozmnožovania baktérií sa plazmid kopíruje do oboch dcérskych buniek [23].



Obr. 1.1: Zjednodušená štruktúra bakteriálnej bunky

1.1.1 Význam plazmidov

Plazmid je nositeľom génov, ktoré sa prenášajú medzi organizmami. Nový plazmid môže baktéria získať z prostredia pomocou procesu transformácie [27] alebo od inej baktérie prostredníctvom konjugácie [55].

Prenášané gény nie sú esenciálne pre každodenné prežitie hostiteľa. Sú zdrojom výhod pre prežitie v meniacom sa prostredí a pri adaptácii na nové podmienky. Medzi najznámejšie gény patrí skupina génov nesúcich rezistencie voči antibiotikám [13, 42], napríklad beta-laktám, chloramfenikol, aminoglykozidy, tetracyklín [16]. Ďalej sú to gény nesúce virulenčné faktory [49], tolerancie ťažkých kovov [43] alebo fixáciu dusíka [38]. Plazmidové gény môžu vplývať aj na okolité prostredie a susedné organizmy produkciou rôznych chemických látok [50]. Plazmidy môžu slúžiť aj ako nástroj na génové manipulácie v molekulárnej biológii či mikrobiológii [20, 57].

1.2 Zostavenie genómu

DNA môžeme definovať ako reťazec $S = s_1 s_2 ... s_n$, kde $s_i \in \{A, C, G, T\}$ reprezentujú jednotlivé nukleotidy a n je dĺžka reťazca. Na identifikáciu a určenie poradia nukleotidov v DNA slúži sekvenovanie. Poznáme rôzne druhy metód a postupov sekvenovania. Dáta, ktoré budeme v práci používať, boli sekvenované Druhou generáciou sekvenovania (Next Generation Sequencing). Výsledkom je množina DNA sekvencií – čítaní. V prípade druhej generácie sekvenovania sú tieto čítania pomerne krátke, okolo 35 až 700 bp [21].

Zostavenie genómu (ang. genome assembly) je proces skladania krátkych DNA úsekov do väčších celkov – kontigov, ktorého cieľom je zostaviť pôvodnú DNA. Vstupom pre zostavenie genómu sú krátke úseky sekvenovanej DNA – čítania. Zostavený genóm je možné vizualizovať pomocou grafu zostavenia genómu (ang. assembly graph). Vrcholmi v grafe sú úseky DNA. Hrany predstavujú možnú následnosť medzi jednotlivými

úsekmi DNA.

V ideálnom prípade je výsledkom niekoľko izolovaných vrcholov predstavujúcich celé chromozómy alebo plazmidy. V praxi je bežné, že sa nepodarí poskladať celú molekulu do jedného kontigu a teda jej v grafe zodpovedá sled pozostávajúci z viacerých kontigov. Avšak graf môže obsahovať aj nadbytočné vrcholy a hrany [56].

Metódy nástrojov na zostavenie genómu možno rozdeliť do troch kategórií: prekryv-usporiadanie-konsenzus (OLC z ang. Overlap-Layout-Consensus), de Bruijnove grafy a pažravé algoritmy [39].

Metóda OLC pozostáva z troch fáz. V prvej fáze sa určia prekryvy medzi čítaniami, z ktorých sa vytvorí graf prekryvov, kde vrcholy predstavujú čítania a hrany sú určené prekryvy [18]. Cieľom druhej fázy je zjednodušiť graf prekryvov a nájsť cesty zodpovedajúce dlhším celkom. Tieto cesty zodpovedajú kontigom. V poslednej fáze sa na základe väčšiny čítaní – konsenzu, zostaví sekvencia zodpovedajúca kontigom.

V metóde DeBruijnových grafov sú čítania rozbité do kratších prekrývajúcich sa podreťazcov dĺžky k. Z nich je vytvorený de Bruijnov graf, kde vrcholy sú podreťazce dĺžky k a orientované hrany predstavujú podreťazce dĺžky $k-1$ zdieľané medzi koncom reťazca jedného vrcholu a začiatkom druhého. Za predpokladu, že k je dostatočne veľké, aby sa žiaden podreťazec dĺžky k neopakoval a že genóm pozostáva iba z jedného chromozómu, tak nájdením Eulerovskej cesty zostavíme genóm. V praxi Eulerovskú cestu nájsť nemusíme a výsledné kontigy sú opäť odvodené ako cesty bez vetvenia v tomto grafe.

Pažravé algoritmy opakujú jednu základnú operáciu: na základe prekryvu spája dve čítania alebo kontigy [39]. Každá operácia použije prekrytie s najvyšším skóre na vytvorenie spojenia. Skórovacia funkcia môže napríklad uvažovať počet zhodných báz v prekrytí. Potom kontigy rastú pažravým spôsobom – vždy sa vyberie dvojica čítaní, ktorá má najväčšie prekrytie.

V práci budeme pracovať s grafmi zostavenia genómu poskladanými nástrojmi Unicycler [63] a Skesa [58]. Oba nástroje boli použité na zostavenie všetkých genómov z krátkych čítaní – pre každý vstupný genóm budeme mať dva grafy zostavenia. Oba nástroje sú založené na metóde DeBruinových grafov.

1.3 Klasifikácia bakteriálnych a plazmidových kontigov

Ak máme na vstupe kontigy pochádzajúce z bakteriálnych izolátov alebo metagenómu, cieľom klasifikácie je určenie pôvodu kontigov – či pochádzajú z plazmidov, baktérií prípadne ďalších organizmov z metagenomickej vzorky. Metagenóm je súbor genómov viacerých organizmov, väčšinou baktérií alebo vírusov. Klasifikačné algoritmy môžeme

rozdeliť do dvoch skupín, druhovo špecifické a druhovo nezávislé [56]. Druhovo špecifické sú natrénované iba na konkrétnom druhu baktérií a slúžia na klasifikáciu tohto druhu. Druhovo nezávislé sú trénované na rôznych druhoch baktérií a často obsahujú metódy, ktoré zovšeobecňujú klasifikáciu pre rôzne druhy.

Jednou z často používaných charakteristík na rozlíšenie plazmidov a chromozómov je sú podreťazce DNA dĺžky k. Z nich je väčšinou vypočítaná frekvencia podreťazcov pre rôzne hodnoty k. Tieto frekvencie možno použiť ako jeden z parametrov pre kontig alebo z nich možno určiť ďalšie vlastnosti, ako je napríklad profil podreťazcov konkrétnej dĺžky. Nevýhodou je, že krátke kontigy môžu obsahovať väčší šum a z toho dôvodu sú sekvencie kratšie ako 1 kbp často vylúčené z klasifikácie. Medzi nástroje používajúce podreťazce patrí napríklad cBar [68], PlasFlow [35] alebo mlplasmids [6]. Ďalšie často používané charakteristiky kontigov sú dĺžka sekvencie alebo obsah nukleotidov guanínu a cytozínu (GC).

Klasifikačné nástroje sú založené aj na použití metód hlbokého učenia. Príkladom je nástroj Deeplasmid [4], ktorý je založený na dlhej krátkodobej pamäti (LSTM z ang. Long short-term memory). Ďalším nástrojom je PPR-Meta [19] používajúci konvolučné siete. Okrem konvolučných sietí používa tzv. "one-hot" kódovanie. Každú sekvenciu dĺžky n pretransformovali na maticu $n \times 4$ a nukleotidy adenín, cytozín, guanín a tymín reprezentovali vektormi $[0, 0, 0, 1]$, $0, 0, 1, 0]$, $[0, 1, 0, 0]$ a $[1, 0, 0, 0]$.

Na klasifikovanie plazmidov je možné použiť informáciu zo susedných vrcholov. Na tomto spôsobe sú založené napríklad nástroje plASgraph2 [56], 3CAC [47] a 4CAC [48]. Ďalším spôsobom je použitie homológie, ktorá je založená na hľadaní podobnosti. Princíp homológie používajú nástroje Platon [52], PlasForest [46] a Deeplasmid [4].

V ďalšej časti sa zameriame na niektoré zo spomenutých nástrojov. Nástroj plASgraph2 podrobnejšie popíšeme v závere kapitoly, nakoľko v práci budeme rozširovať jeho vstupné parametre a pred tým popíšeme grafové neurónové siete, na ktorých je plASgraph2 založený.

1.3.1 Klasifikátory založené na homológii

Homológia funguje na princípe hľadania podobnosti medzi sekvenciami, ktoré chceme klasifikovať a známymi sekvenciami z konkrétnej databázy. Na základe nájdených podobností priradíme kontigom parametre, ktoré môžu pomôcť pri klasifikácii. Jednou z nevýhod je potreba rozsiahlej databázy, ďalšou je široké spektrum druhov, ktoré môžu plazmid obsahovať a ktoré nemusia byť zahrnuté v databáze.

Lokálna databáza je častokrát pripravená z proteínových domén. Proteínové domény sú stavebnou, funkčnou a evolučnou jednotkou proteínovej štruktúry [62]. Sú zodpovedné za určité funkcie proteínu [11]. Domény vedia nadobudnúť štruktúru nezávisle od zvyšku proteínu. Všeobecne sa považujú za vysoko konzervované oblasti

proteínovej sekvencie [7].

Platon

V štúdii Platon [52] skúmali zastúpenie konkrétnych génov kódujúcich proteíny v chromozómoch a plazmidoch. Jednotlivé rozdelenia skombinovali do jednej metriky *skóre rozdelenia replikónov* (RDS). Toto skóre sa vzťahuje na jednu proteínovú markerovú sekvenciu (MPS).

Na vytvorenie databázy proteínových marker sekvencií (MPS) použili všetky reprezentácie bakteriálnych sekvencií z databázy UniRef90 proteínových klastrov, spolu takmer 70 miliónov sekvencií. UniRef90 obsahuje rodiny združujúce sekvencie proteínov s podobnosťou 90 % a viac v rámci rodiny, pričom pre každú rodinu je zvolená jedna reprezentatívna sekvencia.

Na výpočet skóre bolo potrebné spustiť homologické vyhľadávanie všetkých MPS nástrojom Diamond [10] voči kódujúcim sekvenciám (CDSs), ktoré sú výsledkom predikcie nástrojom Prodigal [17].

Vstupom pre Prodigal boli všetky plazmidové sekvencie z bakteriálnej plazmidovej sekcie databázy *NCBI Genomes*. Druhým vstupom boli chromozómy zo všetkých kompletných bakteriálnych genómov z databázy *NCBI RefSeq*. Na predídenie plazmidovej kontaminácie v chromozómovej množine dát odstránili sekvencie s dĺžkou kratšou ako 100 kbp. Po predspracovaní použili 17 430 chromozómových sekvencií a 17 369 plazmidových.

Výsledkom hľadania homológií bol pre každú sekvenciu z množiny MPS počet zarovnaní ku kódujúcim sekvenciám s pokrytím sekvencie aspoň 80 % a identitou sekvencie aspoň 90 % pre chromozómy A_c aj plazmidy A_p. Tieto počty normalizovali celkovým počtom plazmidov R_p a chromozómov R_c, čím dostali pomery $F_p = \frac{A_p}{R_p}$ a $F_c = \frac{A_c}{R_c}$. Tieto hodnoty transformovali do metriky RDS:

$$RDS = 2 \cdot \left(\frac{F_p}{F_p + F_c} - 0,5) \right) \cdot \frac{|F_p - F_c|}{\varphi} \cdot (1 - P_{val}),$$

kde $\varphi = \frac{\sum_{i=1}^{n} |F_{p,j} - F_{p,j}|}{n}$, n je počet prvkov v databáze MPS a $F_{p,j}$ a $F_{c,j}$ sú frekvencie zarovnaní MPS j v plazmidoch a chromozómoch. P_{val} označuje p-hodnotu dvojstranného Fisherovho exaktného testu použitím kontingenčnej tabuľky počtu zarovnaní danej sekvencie MPS ku kódujúcim sekvenciám pre chromozómy aj plazmidy v porovnaní so sekvenciami, v ktorých táto MPS nebola nájdená.

Hodnoty RDS sú na intervale $[-1, 1]$. Záporné hodnoty reprezentujú chromozómové proteíny a kladné hodnoty reprezentujú plazmidové proteíny. Hodnoty RDS pre MPS so štatisticky nevýznamnými počtami výskytov sa blížia k nule vďaka výsledku Fisherovho testu. Predikované kódujúce sekvencie, pre ktoré sa nedá identifikovať MPS, majú priradenú neutrálnu hodnotu RDS=0.

Na klasifikovanie kontigov sa spočíta priemerné RDS všetkých proteínových sekvencií z daného kontigu. Výsledná hodnota sa porovná voči vopred určenému prahu. Okrem toho systém využíva ďalšie ručne vytvorené pravidlá, napríklad všetky kontigy, ktoré majú dĺžku kratšiu ako 1 kbp alebo dlhšiu ako 500 kbp sú klasifikované ako chromozóm. Sekvencie, ktoré sú kratšie, častokrát neobsahujú kódujúcu sekvenciu alebo informácie, ktoré by sa dali použiť na spoľahlivú klasifikáciu. Naopak sekvencie dlhšie ako 500 kbp málokedy pochádzajú z plazmidov. Systém tiež kontroluje, či sa kontig vie uzatvoriť do kruhu, obsahuje aspoň jeden replikačný faktor, mobilizačný faktor a podobne.

PlasForest

Nástroj PlasForest [46] je druhovo nezávislý klasifikátor. Jeho metódy sú tiež založené na homológii. Klasifikuje kontigy do dvoch tried ako plazmidy alebo chromozómy v čiastočne zostavenom genóme.

Homológiu aplikuje cez výpočet podobností vstupných sekvencií voči lokálnej databáze plazmidov, ktorá obsahuje 36 450 sekvencií zo všetkých bakteriálnych plazmidových sekvencií z NCBI RefSeq Genomes. Na zarovnanie (hľadanie podobností) používa nástroj BLASTn [12]. Zarovnania boli ponechané, ak ich e-hodnota bole menej ako 10^{-3}.

Základným prvkom bolo použitie prekryvu medzi vstupným kontigom a sekvenciami z plazmidovej databázy. Pre každú nájdenú podobnosť je prekryv vypočítaný ako percentuálna časť vstupného kontigu ktorá sa zarovnala k sekvencii z databázy plazmidov. Ako parametre kontigu pre klasifikátor nástroj používa počet prekryvov, maximálny prekryv, priemerný prekryv, medián prekryvov a rozptyl prekryvov ako aj veľkosť kontigu a obsah guanínu a cytozínu v sekvencii kontigu. Na klasifikáciu zvolili náhodný les s 500 nezávislými rozhodovacími stromami.

Deeplasmid

Nástroj Deeplasmid [4] je druhovo nezávislý klasifikátor postavený na hlbokom učení. Vstupom pre neurónovú sieť sú dva vektory, x_{seq} a x_f. Vektor x_{seq} je súvislý podreťazec dĺžky 300 bp náhodne vybraný z pôvodného kontigu, zakódovaný pomocou one-hot kódovania do binárneho poľa veľkosti 300×4. Počet podreťazcov vybraných zo sekvencie je vypočítaný z odmocniny dĺžky vstupnej sekvencie. To zaistí objektívnu reprezentáciu krátkych aj dlhých sekvencií.

Vektor x_f obsahuje 16 parametrov normalizovaných na interval $[-1, 1]$ – obsah GC, dĺžku najdlhšieho homopolyméru pre každý nukleotid (homopolymér je postupnosť viacerých kópií jedného nukleotidu za sebou), počet homopolymérov dĺžky viac ako 5 pre všetky nukleotidy, počet zarovnaní k chromozómovým proteínom, počet zarov-

naní k plazmidovým proteínom, počet zarovnaní k počiatkom replikácie, počet génov vo vstupnej sekvencii, percento kódovacej časti v sekvencii, priemerná dĺžka sekvencie nukleových kyselín, dĺžka sekvencie. Ďalej je v tomto vektore ešte 1 538 binárnych hodnôt reprezentujúcich Pfam domény, ktoré označujú výskyt proteínovej domény v sekvencii. Tieto parametre sú vypočítané z celej pôvodnej sekvencie, nie z podreťazcov.

Pfam domény vyhľadali v bakteriálnych sekvenciách dlhších ako 1 Mbp z databázy IMG [15]. Sekvencie dlhšie ako 1 Mbp považujú za chromozomálne, preto výskyty domén v týchto sekvenciách označujú za chromozómovo špecifické. K sekvenciám plazmidov z PLSDB vyhľadali v databáze NCBI príslušné proteíny a tie čo mali Pfam anotáciu, tak boli použité ako plazmidovo špecifické výskyty domén. Náhodne vybrali 70-tisíc výskytov domén z oboch skupín. Vypočítali a normalizovali frekvencie každej domény v chromozómoch a plazmidoch a porovnali pomery frekvencií. Do vektoru použili 800 plazmidovo špecifických domén vyskytujúcich sa 10-krát častejšie v plazmidoch ako v chromozómoch a 738 domén špecifických pre chromozómy vyskytujúcich sa 20-krát častejšie v chromozómoch ako v plazmidoch.

Celý vektor x_f sekvencie S je skopírovaný každému podreťazcu z príslušnej sekvencie S. Výsledkom je skóre z intervalu $[0, 1]$. Čím vyššie skóre, tým pravdepodobnejšie sekvencia pochádza z plazmidu.

1.3.2 Klasifikátory používajúce podreťazce dĺžky k

mlplasmids

Nástroj mlplasmids [6] patrí do kategórie druhovo špecifických klasifikátorov. Je trénovaný na klasifikáciu plazmidových a bakteriálnych kontigov na troch druhoch baktérií *E. faecium*, *K. pneumoniae* a *E. coli*. Pre každý druh z týchto baktérií natrénovali 5 algoritmov učenia s učiteľom – logistickú regresiu, bayesovský klasifikátor, rozhodovacie stromy (ang. decision trees), náhodný les (ang. random forest) a metódu podporných vektorov (ang. support vector machines). Vstupné dáta pre klasifikátory tvorili vektory frekvencií podreťazcov dĺžky 5 pre každý kontig.

Hyperparametre rozhodovacích stromov, náhodného lesu a metódy podporných vektorov boli optimalizované použitím náhodného vyhľadávania v preddefinovanom priestore. Na vyhodnotenie úspešnosti jednotlivých kombinácií použili desaťzložkovú krížovú validáciu (ang. ten-fold cross validation), pri ktorej chybovosť modelu bola použitá ako metrika úspešnosti. Avšak pri modeli pre *E. coli* použili ako metriku počet správne predikovaných pozitívnych hodnôt, aby predišli problému nízkej frekvencii plazmidov. Predikovaná trieda bola pre každý kontig pridelená na základe vypočítanej posteriórnej pravdepodobnosti. Na základe F1 skóre vybrali najlepší model pre každý bakteriálny druh. Pre všetky tri druhy baktérií zvolili metódu podporných vektorov.

1.3.3 Klasifikátory používajúce informáciu zo susedných vrcholov

3CAC a 4CAC

Jedným z nástrojov, v ktorom sa pri klasifikácii kontigov používa informácia zo susedných vrcholov v grafe zostavenia genómu, je nástroj 3CAC [47], na ktorý nadväzuje nástroj 4CAC [48]. Oba nástroje klasifikujú metagenómy.

3CAC [47] je klasifikátor pre tri triedy: bakteriofágy, plazmidy a chromozómy. Bakteriofág je vírus, ktorý infikuje baktérie a dokáže sa v nich ďalej replikovať [32]. Nástroj najprv klasifikuje jednotlivé kontigy pomocou klasifikátorov PPR-Meta [19] a viralVerify [5]. ViralVerify klasifikuje kontigy ako bakteriofágy, plazmidy, chromozómy a neklasifikované. PPR-Meta určí skóre pre príslušnosť k baktériám, plazmidom a chromozómom a trieda s najvyšším skóre je výsledným označením pre daný vrchol. Ak žiadne skóre nie je vyššie alebo rovné hodnote 0,7, tak je vrchol označený ako neklasifikovaný. Potom 3CAC použije nasledovnú heuristiku na rozšírenie informácie medzi vrcholmi.

- Ak má klasifikovaný vrchol v aspoň dvoch susedov a oba majú rovnakú klasifikáciu, ale opačnú ako klasifikovaný vrchol, tak opraví klasifikáciu klasifikovaného vrcholu v na takú, ako majú jeho susedia.

- Ak má neklasifikovaný vrchol v aspoň jedného alebo viac klasifikovaných susedov a všetci majú rovnakú klasifikáciu, tak klasifikuje aj vrchol v do tejto triedy. Môže mať aj neklasifikovaných susedov.

Použitie tejto heuristiky zvýšilo F1-skóre PPR-Meta o 10 až 60 percentuálnych bodov.

Klasifikátor 4CAC [48] je rozšírený až na 4 triedy bakteriofágy, plazmidy, mikroeukaryoty a prokaryoty. Rozdielom oproti 3CAC je použitie xgboost algoritmov na hlavnú klasifikáciu a vytvorenie 5 modelov pre rôzne veľkosti sekvencií. Vstupom pre model je zoznam kontigov a pre každý z nich vektor frekvencií podreťazcov dĺžky 3 až 7. Po klasifikácii je opäť použitá heuristika ako pri 3CAC.

1.4 Grafové neurónové siete

Grafové neurónové siete (ďalej GNN z ang. graph neural network) patria medzi metódy založené na hlbokom učení. Počiatky grafových neurónových sietí siahajú do deväťdesiatych rokov minulého storočia, kedy boli prvýkrát neurónové siete aplikované na orientované acyklické grafy [59]. Ďalším významným príspevkom v oblasti bola štúdia, v ktorej sa zaoberali rekurzívnymi GNN [51].

Pre učenie na grafoch možno definovať úlohy na úrovni vrcholov, hrán a celých grafov [69]. Na úrovni vrcholov je to napríklad regresia alebo klasifikácia konkrétneho vrcholu. Na úrovni hrán je to klasifikácia alebo predikcia spojení. Na celých grafoch to je hlavne klasifikácia.

Grafové neurónové siete možno rozdeliť do 4 kategórií: rekurentné grafové neurónové siete, konvolučné neurónové siete, grafové autoenkódery a priestorovo-časové neurónové siete [64].

Cieľom rekurentných GNN (RecGNN z ang. recurrent GNN) je naučiť sa reprezentáciu vrcholov pomocou rekurentných neurónových sietí [64]. Predpokladom je, že vrchol si v grafe neustále vymieňa informáciu s jeho susedmi, kým nie je dosiahnutá rovnováha. RecGNN boli základom pre výskum konvolučných grafových neurónových sietí, ktorým sa budeme venovať neskôr.

Grafové autoenkódery (GAE z ang. graph autoencoder) patria medzi architektúry učenia bez učiteľa [64]. Kódujú vrcholy a grafy do latentného vektorového priestoru a následne rekonštruujú grafové dáta zo zakódovanej informácie. Používajú sa na naučenie vnorenia (angl. embedding) grafu do vektorového priestoru a generovanie nových grafov.

Priestorovo-časové grafové neurónové siete (STGNN z ang. spatial-temporal graph neural networks) zachytávajú dynamiku grafov [64]. Ich úlohou je naučiť sa skrytý vzor z priestorovo-časových grafov a predpovedať napríklad budúce hodnoty alebo označenia vrcholov. Používajú sa na dopravné predpovede [66] alebo v rozpoznávaní ľudského konania [65]

1.4.1 Konvolučné grafové neurónové siete

Konvolučné grafové neurónové siete (ConvGNN z ang. Convolutional graph neural network) zovšeobecňujú operáciu konvolúcie z mriežkových dát (napr. rastrových obrázkov) na grafové dáta. Hlavnou myšlienkou je generovať reprezentáciu vrchola v agregáciou jeho parametrov x_v a parametrov jeho susedov x_u, kde u je susedom vrchola v. Väčšinou nasleduje za sebou niekoľko grafových konvolučných vrstiev. Vďaka tomu získa každý vrchol informácie zo širšieho okolia [64]. Môžeme ich rozdeliť do dvoch kategórií na spektrálne a priestorové [67]. Spektrálne robia konvolúciu transformáciou reprezentácií vrcholov do spektrálnej domény použitím grafovej Fourierovej transformácie. Priestorové berú do úvahy susedov vrchola.

Ďalej sa budeme venovať konvolučnej GNN zo štúdie z roku 2017 [34]. Má dva vstupné parametre. Prvým je zoznam parametrov x_i pre každý vrchol i v podobe matice parametrov X typu $N \times D$, kde N je počet vrcholov a D počet vstupných parametrov. Druhým vstupným parametrom je reprezentácia štruktúry grafu v podobe matice $N \times N$, väčšinou matice susednosti A [33]. Výstupom je pre úlohy na úrovni

vrcholov matica Z typu $N \times F$ pričom F je počet výstupných parametrov pre každý vrchol. Pre úlohy na úrovni grafov to je vektor alebo číslo z. Každá vrstva neurónovej siete môže byť reprezentovaná ako nelineárna funkcia

$$H^{(l+1)} = f(H^{(l)}, A)$$

kde $H^{(0)} = X$, $H^{(L)} = Z$ a L je počet vrstiev. $H^{(l)} \in \mathbb{R}^{N \times D}$ je matica aktivácii vo vrstve číslo l. Jednotlivé modely sa líšia v tom, akú funkciu f použijeme a ako nastavíme jej parametre. V tomto modeli je použitá nasledovná funkcia:

$$H^{(l+1)} = \sigma(\bar{D}^{-\frac{1}{2}}\bar{A}\bar{D}^{-\frac{1}{2}})H^{(l)}W^{(l)}, \tag{1.1}$$

kde $\bar{A} = A + I$ je matica susednosti neorientovaného grafu G s pridaním slučiek, I je matica identity, $\bar{D}_{ii} = \Sigma_j \bar{A}_{ij}$ je diagonálna matica so stupňami jednotlivých vrcholov, $W^{(l)}$ je matica váh pre danú vrstvu, ktorú je možné trénovať a $\sigma(\cdot)$ reprezentuje aktivačnú funkciu, napríklad $ReLU(x) = max(0, x)$.

Matica A prináša niekoľko obmedzení [33]. Ak by sme vynásobili maticu A iba s H, tak pre každý vrchol dostaneme súčet parametrov všetkých susedných vrcholov, ale bez parametrov daného vrchola. Z toho dôvodu je k matici susednosti A pripočítaná matica identity I. Ďalším limitom je, že matica A zvyčajne nie je normalizovaná a ak by sme s ňou násobili, tak by sme zmenili škálu hodnôt vektorov parametrov. Ak by sme maticu A vynásobili inverznou diagonálnou maticou D^{-1}, súčet prvkov v každom riadku bude rovný jedna. Súčin takto upravenej matice s H zodpovedá výpočtu priemeru parametrov susedných vrcholov. Z praktických dôvodov je v metóde použitá symetrická normalizácia $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$.

1.5 plASgraph2

Najdôležitejším štúdiou pre našu prácu je štúdia o nástroji plASgraph2 [56]. Slúži na klasifikáciu kontigov z krátkych čítaní. Nástroj funguje *de novo*, teda nepotrebuje referenčné genómy a môže byť trénovaný ako druhovo nezávislý, ale aj druhovo špecifický. Každý kontig má niekoľko parametrov:

- Stupeň vrchola v grafe zostavenia genómu
- Dĺžka kontigu vydelená dvomi miliónmi
- Logaritmus dĺžky kontigu
- Relatívne pokrytie – hĺbka čítania kontigu predelená váženým mediánom hĺbok čítaní v celom grafe, v ktorom sa kontig nachádza. Medián je ováhovaný dĺžkou kontigov

- Relatívny obsah GC – od obsahu GC v kontigu je odčítaná priemerná hodnota GC celého grafu, v ktorom sa kontig nachádza
- Relatívny obsah podreťazcov dĺžky $k = 5$ – definovaný ako skalárny súčin medzi vektormi reprezentujúcimi profily podreťazcov daného kontigu a celého grafu zostavenia

Relatívne parametre umožňujú, aby plASgraph2 mohol byť druhovo nezávislým nástrojom – neučí sa vlastnosti, ktoré sú špecifické pre jednotlivé druhy.

Vstupom je graf zostavenia genómu zostavený z krátkych čítaní bakteriálneho izolátu. Kontig môže byť označený ako chromozóm, plazmid, nejednoznačný (chromozóm aj plazmid) a neoznačený. Cieľom je klasifikovať pôvod jednotlivých kontigov – či pochádzajú z chromozómu, plazmidu, z oboch, teda je to nejednoznačný kontig alebo ich typ nie je možné určiť.

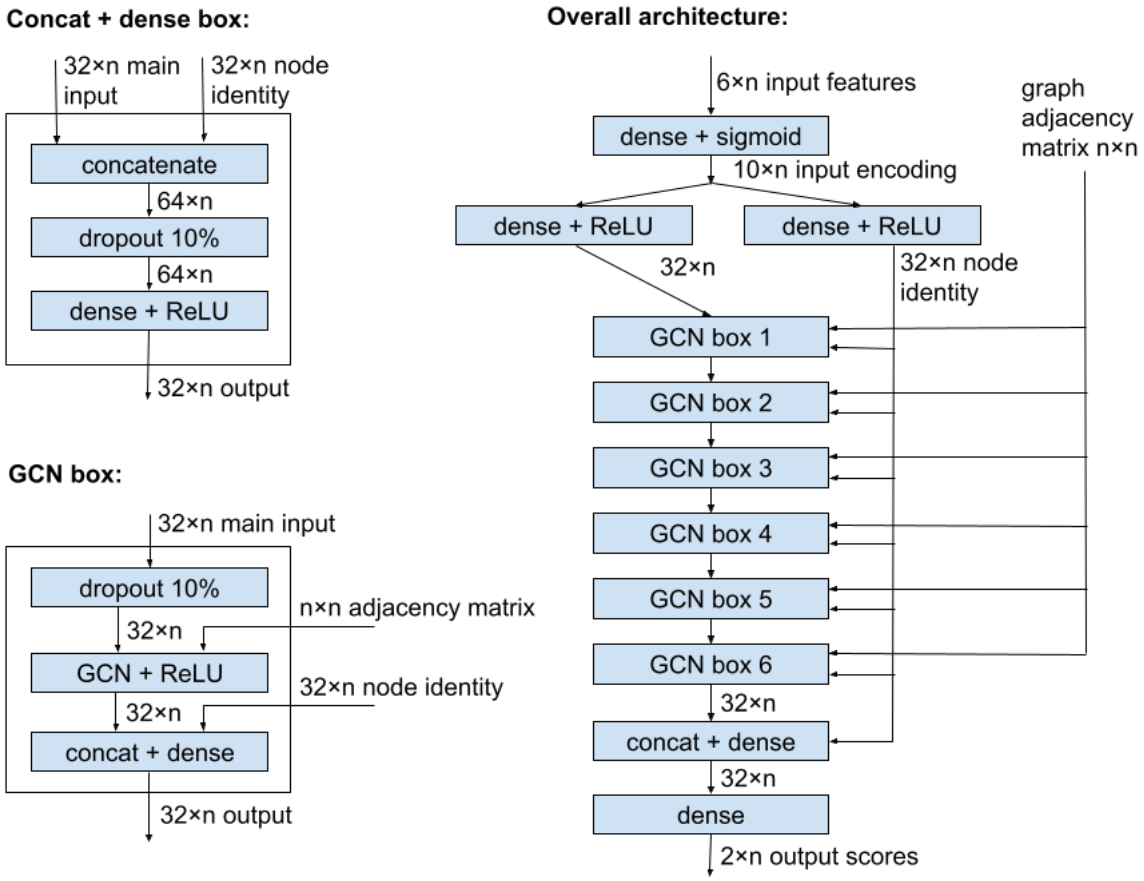
Z dôvodu výskytu nejednoznačných vrcholov je táto úloha rozdelená na dve klasifikačné úlohy – výpočet skóre zvlášť pre pôvod z chromozómu aj plazmidu. Výstupom nástroja sú dve čísla pre každý vrchol, jedno pre pôvod z chromozómu a druhé pre pôvod z plazmidu. Ak obe čísla prekročia vopred danú hranicu, tak plASgraph2 vyhodnotí daný kontig ako nejednoznačný. Naopak, ak obe skóre sú nižšie ako nastavený prah, tak typ kontigu nie je možné určiť.

Na trénovanie programu boli použité dáta zo skupiny patogénov ESKAPEE [6, 28, 37, 14, 45, 41, 54, 2, 9]. Na testovanie a porovnanie s ďalšími nástrojmi boli použité tiež dáta z ESKAPEE a dáta z jednotlivých druhov nepochádzajúcich z ESKAPEE skupiny.

Architektúra

Program stavia na tom, že informácia zo susedných vrcholov môže zlepšiť presnosť, nakoľko susedné vrcholy často pochádzajú z tej iste molekuly. K tomu ako jeden z prvých používa konvolučné GNN, ktoré propagujú informáciu medzi susednými vrcholmi.

Na obrázku 1.2 môžeme vidieť diagram architektúry nástroja plASgraph2. Vstupom je n vrcholov s 6 parametrami popísanými vyššie. Tie sú transformované dvomi plne prepojenými vrstvami do vektorov dĺžky 32. Potom nasleduje 6 konvolučných vrstiev (GCN box). Vstupom pre túto vrstvu je okrem matice vrcholov a ich parametrov aj matica susednosti. Vrstva najprv kombinuje vektory parametrov prislúchajúcich vrcholom a ich susedom podľa vzťahu (1.1). Jedna konvolučná vrstva tak získa informácie od priamych susedov. Ak použijeme L vrstiev, tak získa a rozšíri informáciu do vrcholov vo vzdialenosti L v grafe. ProgramplASgraph2 používa 6 konvolučných vrstiev, ktoré používajú jednu maticu váh W. Vektor parametrov každého vrcholu je po každej konvolučnej vrstve transformovaný plne prepojenou vrstvou s nelineárnou aktivačnou



Obr. 1.2: Architektúra programu plASgraph2 [56]

funkciou. Konvolučné vrstvy kombinujú informácie zo susedných vrcholov a to môže oslabiť vplyv pôvodných parametrov. Preto je vstupom pre plne prepojené vrstvy aj identita pôvodných parametrov.

Posledné dve plne prepojené vrstvy pracujú s každým kontigom samostatne. Ich výstupom sú dve čísla v intervale od 0 po 1, ktorých súčet nemusí byť nutne 1. Označujú príslušnosť vrchola k chromozómom a plazmidom. Kontig je klasifikovaný ako chromozóm, ak výstupné číslo pre chromozóm je aspoň 0,5. Podobne, ak je jeho skóre pre plazmid aspoň 0,5, tak je kontig klasifikovaný ako plazmid.

1.5.1 Vyhodnocovanie úspešnosti

Zo správnych a predikovaných skóre boli vypočítané počty skutočne pozitívnych (TP), skutočne negatívnych (TF), falošne pozitívnych (FP) a falošne negatívnych (FN) kontigov pre každú klasifikačnú úlohu. Každý kontig je započítaný ako jedna jednotka nezávisle od jeho dĺžky. Kontigy bez označenia nie sú počítané do evaluácie. Obe úlohy klasifikačné úlohy – klasifikácia kontigu ako plazmid a klasifikácia kontigu ako chromo-

zóm sú vyhodnotené zvlášť. Nasledovné metriky sú použité na vyhodnotenie:

- precíznosť (ang. precision) – $p = \frac{TP}{TP+FP}$
- návratnosť (ang. recall) – $r = \frac{TP}{TP+FN}$
- presnosť (ang. accuracy) – $a = \frac{TP+TN}{TP+FP+FN+TN}$
- F1 skóre – predstavuje harmonický priemer presnosti a návratnosti, $f = 2\frac{p \times r}{p+r}$
- AUC (ang. Area Under Receiver Operating Characteristic) – obsah plochy pod grafom, ktorý znázorňuje binárnu klasifikáciu a popisuje kvalitu príslušného klasifikátora v závislosti od nastavenia klasifikačného prahu [26].

V porovnaní s nástrojmi mlplasmids [6], PlasClass [44], PlasForest [46], Platon [52], Deeplasmid [4] a RFPlasmid [60] dosiahol najlepšie F1 skóre, presnosť a AUC v klasifikácii plazmidov. V klasifikácii chromozómov bol najúspešnejší nástroj Platon pričom plASgraph2 mal druhé najlepšie F1 skóre. V oboch prípadoch boli klasifikované kontigy s dĺžkou nad 100 báz.

Ak boli klasifikované kontigy s dĺžkou nad 1000 báz, tak lepšie výsledky vykazujú nástroje založené na homológii, keďže od tohto prístupu sa očakáva lepší výkon na dlhších sekvenciách. Nástroj plASgraph2 má vyššiu presnosť na grafoch zostavenia genómu s nižším počtom kontigov.

Kapitola 2

Extremity

V tejto kapitole sa budeme venovať rozšíreniu systému plASgraph2, ktoré podrobnejšie využíva informáciu o prepojeniach kontigov v grafe zostavenia genómu. Najprv sa pozrieme na grafy zostavenia genómu a prepojenie susedných kontigov v nich. Ďalej sa budeme venovať návrhom architektúry a implementácii. Popíšeme testovanie našej architektúry na syntetických dátach a vyhodnotíme výsledky architektúr.

2.1 Graf zostavenia genómu

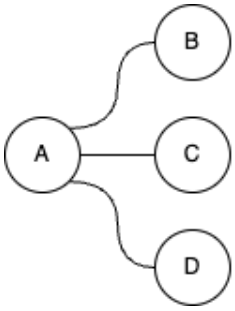
Grafy zostavenia genómu možno reprezentovať pomocou súboru GFA (ang. Graphical Fragment Assembly) [24]. V ňom je zapísaný zoznam vrcholov grafu s identifikátorom kontigu a jeho sekvenciou. Ďalej obsahuje zoznam hrán, v ktorom je pre každú hranu uvedené, ktoré konce kontigov sú susedné a koľko ich nukleotidov sa prekrýva. Na obrázku 2.1 môžeme vidieť príklad súboru GFA. Riadky začínajúce písmenom S označujú kontig (segment) a riadky začínajúce písmenom L susedné segmenty.

Program plASgraph2 zo súboru na obrázku 2.1 vytvorí graf na obrázku 2.2. Pri vytvorení grafu program nerozlišuje, či je susedný kontig pripojený z ľavej alebo z pravej strany kontigu, pracuje iba s celkovým počtom susedov. Našou myšlienkou je, že ak má kontig na jednom konci jedného suseda typu T a na druhom konci má susedov dvoch a viac typov $T_1, T_2, ...$, tak je pravdepodobné, že daný vrchol bude toho istého typu ako je jeho sused T. Preto by sme chceli v našej architektúre vedieť pracovať s grafom, kde

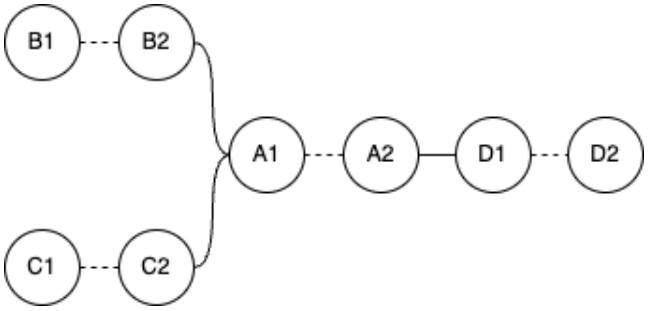
```
S       A       ATGTCTCGGAAAGGGAACTGCTTAGATAATTCACTTATGGAGAATTTTTTTG
S       B       CAAAAAAATTCTCCATAAGTGAATTATCTAAGCAGTTCCCTTTCCGAGACAG
S       C       TGATTCGTTTTTGGCTGCCGTAGCAACTTTTATATCATATCAGAGAGACG
S       D       TCGCCGTTTGTATAATTAAGCGAGACAAATAAAAAAGCCATTTATGT
L       B       +       A       +       0M
L       C       +       A       +       0M
L       A       +       D       +       0M
```

Obr. 2.1: Príklad súboru GFA

sa táto informácia nestratí. Príklad želaného grafu môžeme vidieť na obrázku 2.3.



Obr. 2.2: Vizualizácia vstupného grafu pre program plASgraph2



Obr. 2.3: Vizualizácia vstupného grafu pre program plASgraph2 s použitím extremít. Plné čiary reprezentujú maticu susednosti A a čiarkované čiary hrany maticu susednosti A_{extr}.

Zvolili sme použitie grafu susednosti intervalov (ang. Interval Adjacency Graph) [3], ktorý môžeme formálne definovať nasledovne. Segment $S = [s^z, s^k]$ je súvislá časť molekuly so začiatkom s^z a koncom s^k, ktoré budeme nazývať *extremity*. Susednosť $a = p, q$ spájajúca extremity p a q dvoch sekvencií určuje prechod medzi susednými segmentmi v molekule. Potom intervalový graf susednosti $G(S, A) = (V, E)$ je graf vytvorený na základe množiny segmentov S a množiny priľahlostí A. Množina vrcholov $V = \{s^z, s^k | s \in S\}$ reprezentuje extremity segmentov v množine S. Množina hrán E pozostáva z dvoch typov neorientovaných hrán. Prvým typom sú hrany spájajúce extremity v segmente $E_S = \{\{s^z, s^k\} | s \in S\}$ a druhým sú susedné hrany $E_A = \{\{p, q\} | \{p, q\} \in A\}$ spájajúce extremity rôznych segmentov.

2.2 Návrh architektúry

Grafová neurónová sieť systému plASgraph2 reprezentuje každý kontig jedným vrcholom. V našom modeli sme každý vrchol (kontig, segment) rozdelili na dva, ktoré reprezentujú dva rôzne konce jedného kontigu (extremity). Oba vrcholy budú mať nakopírované takmer všetky vlastnosti pôvodného vrchola, ako napríklad dĺžka sekvencie, logaritmus dĺžky sekvencie, obsah guanínu a cytozínu, relatívne pokrytie kontigu čítaniami a relatívny obsah podreťazcov dĺžky $k = 5$. Jednou výnimkou je počet susedov, ktorý sa určí ako reálny počet susedov danej extremity v našom grafe. Ďalej má graf dve matice susedností, maticu A, ktorá reprezentuje susednosti z množiny E_A a maticu A_{extr}, ktorá reprezentuje susednosti z množiny E_S. Na obrázkoch 2.2 a 2.3 je znázornené porovnanie pôvodného vstupného grafu pre program plASgraph2 a nového s použitím extremít.

Ďalším krokom bolo premyslenie, ako prispôsobíme architektúru neurónovej siete grafu s extremitami, pričom sme museli vyriešiť dva problémy. Prvý problém je, že výsledkom výpočtu má byť dvojica skóre (skóre pre chromozóm a plazmid) pre každý vstupný kontig, teda matica $n \times 2$ a pri priamočiarom použití pôvodnej architektúry dostaneme $2n \times 2$, keďže sme každý vrchol rozdelili na dva. Druhým problémom je, ako použiť informáciu z oboch matíc susedností A a A_{extr}.

Počet výsledných skóre sme skúsili zredukovať dvomi spôsobmi. Najprv sme v celej architektúre ponechali $2n$ vrcholov, pričom pri trénovaní sme počítali chybu na každej extremite zvlášť a až následne v testovacej fáze sme spriemerovali skóre oboch extremít do jedného výsledného skóre pre daný kontig.

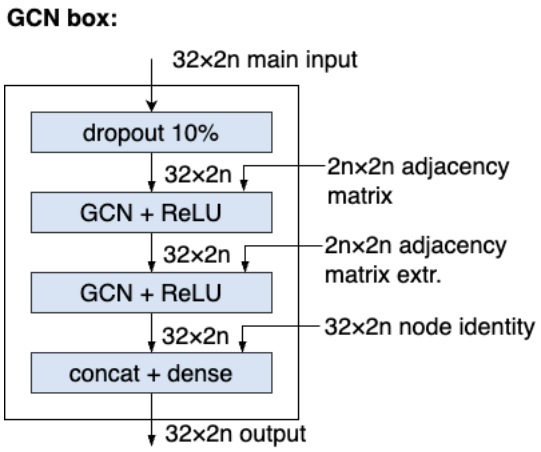
Druhou možnosťou bolo spojenie informácie oboch extremít pred poslednou plne prepojenou vrstvou siete. K tomu sme použili operáciu reshape, čiže sme spojili dva príslušné vektory extremít s dĺžkou m do jedného vektora s dĺžkou $2m$. Z tohto vektora potom plne prepojená vrstva spočítala skóre plazmidu a skóre chromozómu platné pre celý kontig.

Na použitie oboch matíc susedností sme zvolili opäť dva prístupy. V existujúcej architektúre sa najprv použije matica susednosti A v grafovej konvolučnej vrstve a za ňou nasleduje plne prepojená vrstva. V novej architektúre medzi tieto dve vrstvy vložíme ešte jednu grafovú konvolučnú vrstvu, ktorej vstupom bude výstup z prvej konvolučnej vrstvy a matica susednosti A_{extr}. V tomto prípade je možné vyskúšať použiť iba jednu inštanciu konvolučnej vrstvy pre obe matice susedností alebo dve – jednu pre maticu A a druhú pre maticu A_{extr}. Ak použijeme pri inicializácii iba jednu maticu váh, má to za následok, že obidve GCN vrstvy budú používať rovnaké váhy a teda sa budú správať rovnako k informácii od susedov z matice A a z matice A_{extr}. Na obrázku 2.4 môžeme vidieť znázornenú popisovanú architektúru.

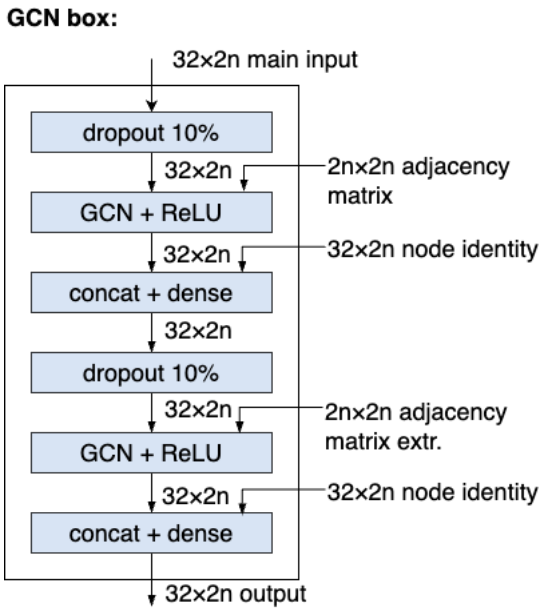
Druhou možnosťou bolo zopakovanie celého tzv. GCN boxu – skupiny vrstiev, ktorý sme uviedli na obrázku 1.2. Vstupom pre druhú skupinu sú výsledné dáta z prvého GCN boxu a matica susednosti A_{extr}. V tomto prípade tiež možno testovať rôznu inicializáciu konvolučnej, ale aj plne prepojenej vrstvy (rovnaké alebo rôzne inštancie). Na obrázku 2.5 môžeme vidieť znázornenú popisovanú architektúru GCN boxu.

Implementácia

Program plASgraph2 sme upravili tak, aby z každého vstupného segmentu z GFA súboru vytvoril dva vrcholy s rovnakými vlastnosťami, ale s počtom susedov príslušným k danému koncu segmentu, ako sme už opísali vyššie. Susediace segmenty sme ukladali do jedného grafu G a spájajúce extremity v segmente do druhého grafu G_{extr} v triedach Graph z knižnice Networkx [25]. Informáciu z oboch grafov sme spojili v upravenej triede SingleLoader z knižnice Spektral [22], kde sme implementovali ukladanie dvoch



Obr. 2.4: Architektúra s dvomi konvolučnými vrstvami



Obr. 2.5: Architektúra so zopakovaným GCN boxom

matíc susedností. Architektúra neurónovej siete je založená na knižniciach Tensorflow [1] a Spektral [22].

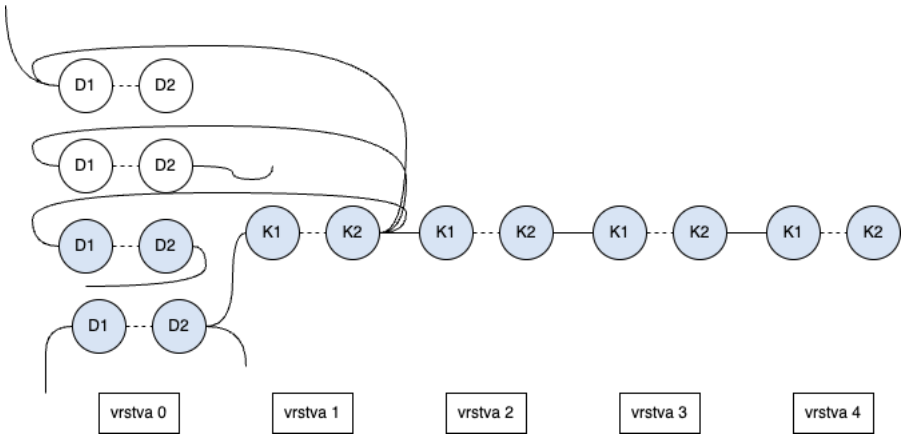
2.3 Testovanie architektúry na syntetických dátach

Na testovanie, či naša architektúra správne šíri informáciu v kontigu od jednej extremity k druhej sme použili umelo vygenerované vstupné dáta, na ktorých sme sledovali, ako sa šíri informácia v grafe. Vytvorili sme graf zostavenia genómu, v ktorom sa nachádzajú dlhé a krátke vrcholy typu chromozóm aj plazmid.

Dlhé vrcholy je jednoduché klasifikovať, pretože dlhý vrchol typu plazmid má obsah guanínu a cytozínu 40 % a dlhý vrchol typu chromozóm má obsah guanínu a cytozínu 60 %. Šírenie informácie overíme na krátkych vrcholoch, pretože oba typy majú obsah guanínu a cytozínu 50 %. Kontigy boli náhodne vygenerované s daným obsahom GC. Krátke vrcholy majú dĺžku 200 bp a dlhé vrcholy 1 000 bp. Všetky kontigy mali rovnaké pokrytie rovné 1.

Na obrázku 2.6 môžeme vidieť ilustráciu syntetického vstupu. Vrcholy rozdelíme do vrstiev. Dlhé vrcholy budú reprezentovať vrstvu 0. Za suseda môžu mať iba krátke vrcholy vrstvy 1. Každý krátky vrchol vrstvy 1 má na extremite K_1 jedného dlhého suseda rovnakého typu a na druhej extremite K_2 jedného dlhého suseda rovnakého typu a 2 dlhých susedov druhého typu. Tieto dlhé susediace kontigy sú vybrané náhodne. Extremita K_2 má ešte na pravej strane jedného krátkeho suseda toho istého typu, ale z vrstvy 2. Takto môže nasledovať postupnosť krátkych kontigov, kde každý krátky

kontig vrstvy $v \geq 2$ môže mať za suseda iba jeden krátky vrchol z vrstvy $v-1$ na ľavej strane a jeden z vrstvy $v+1$ na pravej strane, ak existuje $v+1$ vrstiev krátkych kontigov. Ak nie, tak na pravej strane nemá žiadneho suseda. Kontig vrstvy $v \geq 2$ nesusedí so žiadnym dlhým kontigom. V každej vrstve je rovnaký počet vrcholov.



Obr. 2.6: Umelo vygenerovaný graf na testovanie návrhu architektúry. Rôzne farby označujú rôzne molekuly – chromozóm a plazmid. Dlhé kontigy sú označené písmenom D a krátke K. Čísla 1 a 2 označujú extremity. Plné čiary označujú hrany medzi kontigmi a čiarkované medzi extremitami v kontigoch.

Pôvodný program by nemal vedieť správne určiť označenie krátkych kontigov nakoľko považuje všetkých susedov za rovnocenných. Verzia s extremitami by mala vedieť správne vyhodnotiť všetky krátke kontigy. Extremita $K1$ získa informáciu o označení suseda $D2$, túto informáciu rozšíri použitím matice susednosti A_{extr} k extremite $K2$ vrstvy 1 a ďalej extremite $K1$ vo vrstve 2, atď.

Pre vyhodnotenie hypotézy sme vygenerovali 10 umelých grafov s počtom kontigov každého typu 1 000. V každom grafe sme vygenerovali 1 vrstvu dlhých kontigov a 4 vrstvy krátkych kontigov. Päť grafov sme použili na trénovanie a päť na testovanie.

Porovnali sme výsledky pôvodného programu a verziu s extremitami s použitím konvolučnej vrstvy 2x za sebou. Obe verzie sme spustili najprv s 4 a potom 5 GCN vrstvami.

V tabuľke 2.1 môžeme vidieť priemerné skóre a štandardnú odchýlku pre každý typ kontigu. Ako sme predpokladali, pôvodný program (orig-4GCN a orig-5GCN) správne klasifikoval dlhé kontigy, ale nevedel správne klasifikovať krátke kontigy ani so 4 ani s 5 GCN vrstvami. Priemerné skóre dlhých kontigov je 1 v prospech správneho označenia, ale hodnoty krátkych kontigov sa pohybujú okolo 0,5, čo je aj prah určenia, či daný kontig je považovaný za chromozóm alebo plazmid. Verzia s extremitami so 4 GCN vrstvami (v1-4GCN) správne klasifikovala dlhé kontigy a krátke po vrstvu 3 vrátane, ale vrstvu 4 klasifikovala so skóre okolo 0,5, ako môžeme vidieť v tabuľke. Verzia

s extremitami s 5 GCN vrstvami (v1-5GCN) určila správne všetky kontigy a všetky priemerné skóre kontigov dosiahli hodnotu 1 v prospech daného označenia.

Typ kontigu	Priem. plaz.	Štd. odch. plaz.	Priem. chr.	Štd. odch. chr.
orig-4GCN				
Dlhý chr.	0	0,0001	1,0000	0
Dlhý plazmid	1,0000	0	0	0
Kr. chr. v1	0,4999	0,0010	0,4993	0,0007
Kr. chr. v2	0,4995	0	0,4999	0
Kr. chr. v3	0,4991	0	0,4998	0
Kr. chr. v4	0,4994	0	0,4999	0
Kr. plazmid v1	0,4999	0,0010	0,4993	0,0007
Kr. plazmid v2	0,4995	0	0,4999	0
Kr. plazmid v3	0,4991	0	0,4998	0
Kr. plazmid v4	0,4994	0	0,4999	0
orig-5GCN				
Dlhý chr.	0	0	1,0000	0
Dlhý plazmid	1,0000	0	0	0
Kr. chr. v1	0,5039	0,0018	0,4978	0,0027
Kr. chr. v2	0,5000	0	0,5006	0
Kr. chr. v3	0,4976	0	0,5035	0
Kr. chr. v4	0,4981	0	0,5029	0
Kr. plazmid v1	0,5039	0,0018	0,4978	0,0027
Kr. plazmid v2	0,5000	0	0,5006	0
Kr. plazmid v3	0,4976	0	0,5035	0
Kr. plazmid v4	0,4981	0	0,5029	0
v1-4GCN				
Dlhý chr.	0	0,000	1,0000	0
Dlhý plazmid	1	0	0	0
Kr. chr. v1	0	0	1,0000	0
Kr. chr. v2	0	0	1	0
Kr. chr. v3	0	0	1,0000	0
Kr. chr. v4	0,5056	0	0,4943	0
Kr. plazmid v1	1,0000	0	0	0
Kr. plazmid v2	1,0000	0	0	0
Kr. plazmid v3	1,0000	0	0	0
Kr. plazmid v4	0,5056	0	0,4943	0
v1-5GCN				
Dlhý chr.	0	0	1,0000	0

Dlhý plazmid	1	0	0	0
Kr. chr. v1	0	0	1	0
Kr. chr. v2	0	0	1	0
Kr. chr. v3	0	0	1	0
Kr. chr. v4	0	0	1	0
Kr. plazmid v1	1	0	0	0
Kr. plazmid v2	1,0000	0	0	0
Kr. plazmid v3	1,0000	0	0	0
Kr. plazmid v4	1,0000	0	0	0

Tabuľka 2.1: Porovnanie priemerného skóre kontigov pôvodnej verzie a verzie s extremitami na syntetických vstupných dátach

Verzia s extremitami s piatimi GCN vrstvami dosahuje všetky výsledné metriky popísané v časti 1.5.1 rovné 1. Ako sme videli v tabuľke 2.1 verzia s extremitami so štyrmi GCN vrstvami (v1-4GCN) preferovala triedu plazmid na určenie krátkych kontigov vrstvy 4, čo sa odrazilo aj na výsledkoch návratnosti (0,8 pre chromozómy) a precíznosti (0,8 pre plazmidy). Klasifikovala správne všetky plazmidové kontigy, ale z chromozómových kontigov klasifikovala správne 80 %.

2.4 Výsledky

Enterococcus faecium

Na vývoj a testovanie implementácie sme použili grafy zostavenia genómu patogénnej baktérie *Enterococcus faecium* pripravené autormi plASgraph2. Trénovacia množina obsahuje 46 grafov a spolu 13674 kontigov dlhších ako 100 bp, z toho 1 658 nejednoznačných vrcholov, 8 260 chromozómov, 2 826 plazmidov a 930 neoznačených vrcholov. Z nich plASgraph2 náhodne vyberie 20 % kontigov, ktoré použije ako validačnú množinu a zvyšných 80 % použije na trénovanie. Sekvencie kratšie ako 100 bp neboli na trénovanie použité. Testovacie dáta obsahujú 60 grafov.

Skúsili sme 7 rôznych architektúr, ktoré sme bližšie popísali vyššie a pôvodnú implementáciu programu plASgraph2:

- v-orig – Pôvodná verzia programu plASgraph2
- v0 – Verzia s dvomi konvolučnými vrstvami + spriemerovanie výsledných skóre
- v1 – Verzia s dvomi konvolučnými vrstvami + reshape
- v1.1 – Verzia s dvomi konvolučnými vrstvami rôznej inštancie + reshape

- v2 – Verzia s dvomi GCN boxmi + reshape
- v2.1 – Verzia s dvomi GCN boxmi a konvolučnými vrstvami rôznej inštancie + reshape
- v2.2 – Verzia s dvomi GCN boxmi a plne prepojenými vrstvami nasledujúcimi za konvolučnými vrstvami rôznej inštancie + reshape
- v2.3 Verzia s dvomi GCN boxmi a konvolučnými aj plne prepojenými vrstvami nasledujúcimi za konvolučnými vrstvami rôznej inštancie + reshape

Všetky verzie sme testovali s rovnakou konfiguráciou. Použili sme 1 000 epoch, 6 vrstiev GCN boxov, všetky vrstvy boli inicializované na jednu inštanciu a použili sme rovnaké vlastnosti vyextrahované z grafu. Sumarizáciu výsledkov môžeme vidieť v tabuľke 2.2. V tabuľke uvádzame pre experimenty výsledné metriky, ktoré sme bližšie popísali v časti 1.5.1. Okrem nich uvádzame aj trénovaciu a testovaciu chybu. Pre každú verziu sú v tabuľke dva riadky rozlíšené stĺpcom *Mol.*, ktorý obsahuje výsledné metriky klasifikácie pre danú molekulu – chromozóm a plazmid. Nenachádza sa tam presnosť, keďže výpočet tejto metriky nebol zahrnutý vo verzii programu, do ktorej sme začali implementovať extremity.

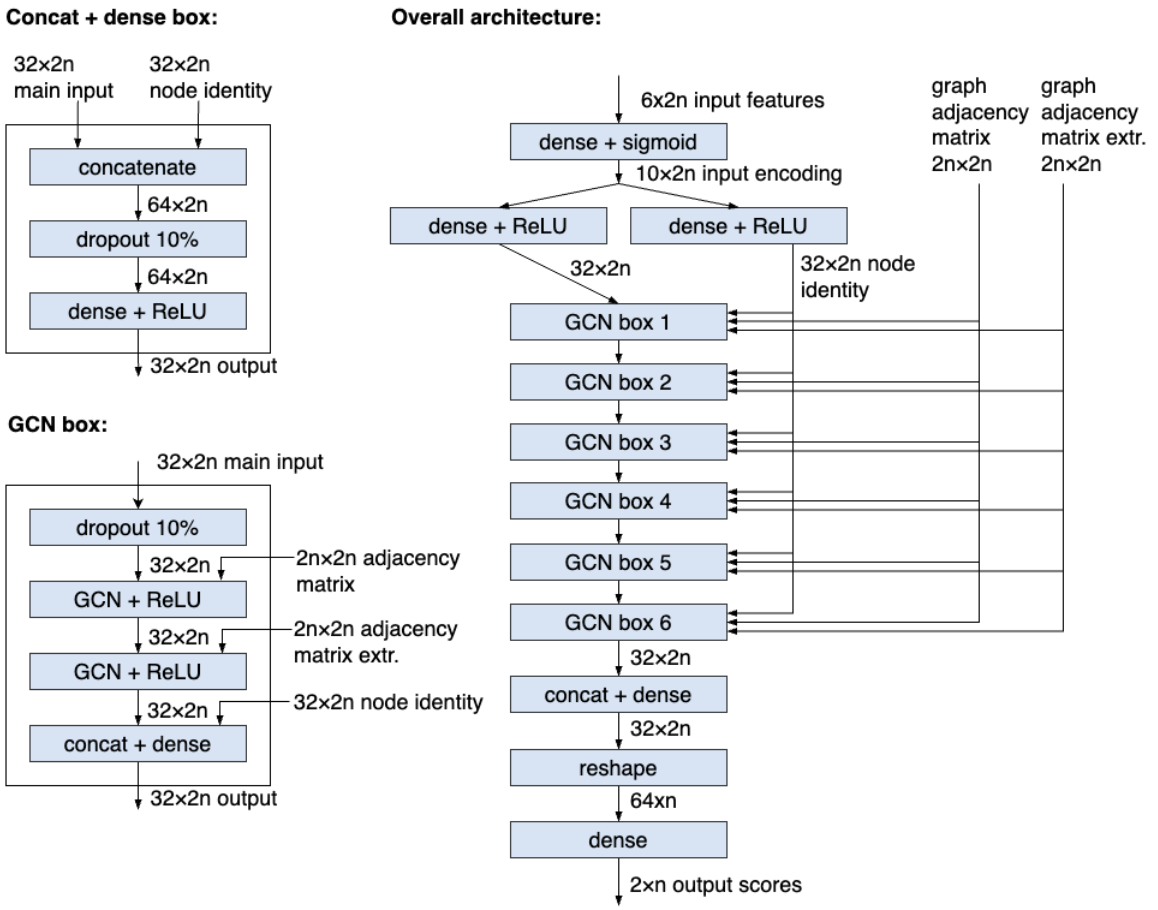
Spriemerovanie výsledných skóre sme neskúšali v ďalších kombináciách. V tabuľke môžeme vidieť, že má najvyššiu trénovaciu aj validačnú chybu z dôvodu dvojnásobného počtu vrcholov. Všetky verzie nadobudli lepšiu hodnotu AUC oproti pôvodnej verzii. Najvyššiu hodnotu $0,9426$ dosiahla pre plazmidy verzia v2.2 a pre chromozómy $0,9528$ verzia v1.2. F1 skóre bolo tiež veľmi podobné, väčšie rozdiely vznikli u plazmidov. Najväčšie rozdiely boli u metrík precíznosť a návratnosť – ak bola vyššia návratnosť, tak sa znížila precíznosť a naopak. Návratnosť u plazmidov bola vyššia u všetkých verzií oproti pôvodnej. Naopak precíznosť pôvodnej verzie u plazmidov bola jedna z najvyšších.

Okrem rôznej architektúry sme skúsili aj nastavenie pohyblivého prahu pre zlepšenie kompromisu medzi precíznosťou a návratnosťou. Podľa výsledkov na validačných dátach sa nastavila nová hranica skóre na klasifikácie molekuly do triedy plazmid a chromozóm miesto pôvodnej hranice 0,5 tak, aby sa maximalizovalo F1 skóre (tento postup bol použitý aj v programe plASgraph2). Výsledky sú v tabuľke 2.2 s označením "prah=áno".

Na základe výsledkov sme sa rozhodli pre architektúru s dvomi GCN vrstvami idúcimi za sebou (verzia v1) s nastavením prahu klasifikovania. Rozhodli sme sa pre túto verziu nakoľko má jednu z najvyšších hodnôt AUC a F1 skóre. Taktiež hodnoty precíznosti a návratnosti sú pomerne vyrovnané. Na obrázku 2.7 môžeme vidieť vizualizáciu celej architektúry.

Verzia	Prah	Mol.	AUC	Precíz.	Návrat.	F1	Tr. ch.	Val. ch.
v-orig	nie	plas	0,9256	0,8153	0,7501	0,7814	2191,78	537,65
v-orig	nie	chr	0,9385	0,9179	0,9607	0,9388	2191,78	537,65
v0	nie	plas	0,9379	0,7507	0,8686	0,8054	4524,56	1121,81
v0	nie	chr	0,9516	0,9438	0,9289	0,9363	4524,56	1121,81
v1	nie	plas	0,9391	0,7706	0,8534	0,8099	2169,06	526,41
v1	nie	chr	0,9507	0,9460	0,9191	0,9324	2169,06	526,41
v1.1	nie	plas	0,9394	0,7643	0,8624	0,8103	2013,43	497,48
v1.1	nie	chr	0,9529	0,9468	0,9284	0,9375	2013,43	497,48
v2	nie	plas	0,9295	0,7374	0,8704	0,7984	2349,55	610,32
v2	nie	chr	0,9457	0,9524	0,9005	0,9257	2349,55	610,32
v2.1	nie	plas	0,9385	0,7485	0,8806	0,8092	2074,40	535,16
v2.1	nie	chr	0,9500	0,9466	0,9212	0,9338	2074,40	535,16
v2.2	nie	plas	0,9426	0,8193	0,8031	0,8111	2084,10	512,44
v2.2	nie	chr	0,9499	0,9314	0,9478	0,9395	2084,10	512,44
v2.3	nie	plas	0,9404	0,8196	0,7911	0,8051	1962,14	465,86
v2.3	nie	chr	0,9503	0,9262	0,9541	0,9399	1962,14	465,86
v-orig	áno	plas	0,9256	0,7636	0,8333	0,7970	2191,78	537,65
v-orig	áno	chr	0,9385	0,9328	0,9439	0,9383	2191,78	537,65
v1	áno	plas	0,9391	0,7932	0,8181	0,8054	2169,06	526,41
v1	áno	chr	0,9507	0,9251	0,9525	0,9386	2169,06	526,41
v1.1	áno	plas	0,9394	0,7537	0,8788	0,8114	2013,43	497,48
v1.1	áno	chr	0,9529	0,9196	0,9628	0,9407	2013,43	497,48
v2	áno	plas	0,9295	0,7566	0,8417	0,7969	2349,55	610,32
v2	áno	chr	0,9457	0,9352	0,9318	0,9335	2349,55	610,32
v2.1	áno	plas	0,9385	0,7831	0,8426	0,8118	2074,40	535,16
v2.1	áno	chr	0,9500	0,9288	0,9543	0,9414	2074,40	535,16
v2.2	áno	plas	0,9426	0,7824	0,8552	0,8172	2084,10	512,44
v2.2	áno	chr	0,9499	0,9221	0,9598	0,9406	2084,10	512,44
v2.3	áno	plas	0,9404	0,8216	0,7896	0,8053	1962,14	465,86
v2.3	áno	chr	0,9503	0,9270	0,9529	0,9398	1962,14	465,86

Tabuľka 2.2: Tabuľka s výsledkami rôznych návrhov architektúry



Obr. 2.7: Vizualizácia architektúry v1 s dvomi GCN vrstvami idúcimi za sebou. Obe GCN vrstvy zdieľajú spoločnú inštanciu.

ESKAPEE

Grafy zostavenia genómu finálnych trénovacích a testovacích dát pochádzajú z patogénov *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Enterobacter spp.* a *Escherichia coli* spolu označovaných ako ESKAPEE skupina. Dáta pochádzajú z niekoľkých štúdií [6, 28, 37, 14, 45, 41, 54, 2, 9] a boli použité ako trénovacia a testovacia množina v programe plAsgraph2. Trénovaciu množinu tvorí 140 grafov a testovaciu 224 grafov.

V tabuľke 2.3 môžeme vidieť výsledky na ESKAPEE dátach. Verzii s extremitami sa podarilo dosiahnuť lepšie výsledky na všetkých metrikách. Zlepšenie vidieť hlavne u plazmidov na precíznosti o takmer 0,04 a na F1 skóre o 0,02.

2.4.1 Štatistika extremít v trénovacích dátach

Hoci pridanie extremít do modelu prinieslo zlepšenie, nebolo až také výrazné. Pozreli sme sa teda na vlastnosti grafov v trénovacích dátach *Enterococcus faecium*. Vypočítali

Verzia	Mol.	AUC	Precíz.	Návrat.	F1	Tr. ch.	Val. ch.
orig	plaz.	0,9256	0,6645	0,7171	0,6898	4510,11	1038,85
orig	chrom	0,9359	0,9570	0,9620	0,9595	4510,11	1038,85
v1	plaz.	0,9362	0,7036	0,7288	0,7160	4589,54	1090,57
v1	chrom	0,9429	0,9620	0,9644	0,9632	4589,54	1090,57

Tabuľka 2.3: Výsledky experimentov na ESKAPEE dátach

Typ kontigu	Počet	Počet1	Počet2
Nejednoznačný	1658	748 (45,11 %)	645 (38,90 %)
Chromozóm	8260	468 (5,67 %)	339 (4,10 %)
Plazmid	2826	214 (7,57 %)	109 (3,86 %)
Neoznačený	930	144 (15,48 %)	120 (12,90 %)
Spolu	13674	1574 (11,51 %)	1213 (8,87 %)

Tabuľka 2.4: Štatistika na počtoch vrcholov s jedným typom suseda na jednom konci a viacerými typmi na druhom konci. Počet1 označuje počet kontigov s jedným typom suseda na jednom konci a viacerými typmi na druhom. Počet2 označuje počet kontigov s rovnakým typom suseda na jednom konci a viacerými typmi na druhom. V zátvorkách sú uvedené percentá z celkového počtu daného typu.

sme, koľko kontigov má susedov iba jedného typu na jednom konci a susedov viacerých typov na druhom konci. Práve pre takéto kontigy môže viesť zavedenie extremít najjednoduchšie k zlepšeniu presnosti.

V tabuľke 2.4 môžeme vidieť tri údaje pre každý typ kontigu. Počet označuje celkový počet kontigov s daným typom a dĺžkou nad 100 bp. Počet1 označuje kontigy s jedným typom susedov na jednom konci a s viacerými typmi na druhom konci. Počet2 označuje, koľko kontigov má na jednom konci iba susedov rovnakého typu ako je on a na druhom konci susedov viacerých typov. Typ kontigu a typ suseda označuje rozdelenie do štyroch tried – plazmid, chromozóm, nejednoznačný a neoznačený. Chromozómov, ktoré majú na jednom konci suseda chromozóm a na druhej viac typov je iba 4,10 %. Plazmidov, ktoré majú na jednom konci suseda plazmid a na druhej viac typov je 3,86 %, teda pre veľa plazmidov, ktoré majú na jednom konci iba susedov jedného typu, tento sused nie je plazmid. Na základe týchto metrík môžeme predpokladať, že vylepšenie pomocou extremít mohlo pomôcť iba malému percentu kontigov (3,27 %).

Kapitola 3

Homológia

V tejto kapitole popíšeme, ako sme program plASgraph2 rozšírili o použitie homológie. Štúdie a metódy používajúce homológiu sme bližšie popísali v časti 1.3.1. Budeme sa venovať rôznemu výberu proteínových domén a vyhodnotíme výsledky experimentov.

3.1 Log-odd skóre

Autori nástroja plASgraph2 už skúsili zahrnúť homológiu, ale tento pokus nie je súčasťou publikovanej štúdie. Použili na to kumulatívne skóre, ktoré získali pomocou logaritmu pomeru frekvencií domény v plazmidoch a domény v chromozómoch a plazmidoch tzv. *log-odd ratio*. Ďalej budeme toto skóre označovať ako log-odd skóre.

Predtým ako je možné vypočítať log-odd skóre, je potrebné nájsť podobnosti medzi vstupnými kontigmi a lokálnou databázou. Na to slúži nástroj BLAST [12], ktorý hľadá podobné oblasti medzi vstupnými sekvenciami a sekvenciami z lokálnej databázy. Výsledkom výpočtu okrem zarovnania sekvencií je aj percento identity $pIdent$ a štatistická významnosť e-hodnota. Na základe e-hodnoty môžeme odfiltrovať zarovnania s vysokou šancou náhodného pozorovania pre databázu danej veľkosti.

Lokálna databáza proteínových domén pre naše experimenty je vytvorená z proteínových domén z databázy Pfam [40] stiahnutých z databázy NCBI CDD [61]. Na hľadanie domén v sekvenciách použili nástroj rpstblastn [12] s nastavením odfiltrovania domén, ktoré dosiahnu e-hodnotu rovnakú alebo nižšiu ako $0,001$. Bol použitý aj parameter $max\_target\_seqs = 20\,000$ na ponechanie prvých 20 000 zarovnaní spĺňajúcich nastavenú e-hodnotu [53].

Na základe výsledku z BLAST vieme určiť, ktoré domény sa zarovnali k vstupným kontigom. Vďaka tomu vieme spočítať frekvenciu každej domény d v plazmidových ($x_{P,d}$) a chromozómových $x_{C,d}$ kontigoch. Ak je kontig nejednoznačný, započíta sa do oboch početností. Každej početnosti sa pripočíta pseudopočet podľa príslušnej molekuly. Je rôzny pre chromozómy aj plazmidy, ale ich súčet je konštantný a v tomto

prípade rovný 10. Vypočíta sa na základe frekvencie danej molekuly.

Plazmidové skóre $s_{P,d}$ pre doménu d je určené vzťahom:

$$s_{P,d} = \log \frac{x_{P,d}}{\sum_{d'} x_{P,d'}} - \log \frac{\sum_{m \in \{C,P\}} x_{m,d}}{\sum_{m \in \{C,P\}} \sum_{d'} x_{m,d'}},$$

kde prvá časť reprezentuje logaritmus relatívnej frekvencie domény v doménach v plazmidoch. Druhá časť je relatívna frekvencia domény spomedzi všetkých domén v plazmidoch aj chromozómoch. Plazmidové skóre domény je log-odd skóre týchto dvoch relatívnych frekvencií. Skóre je pozitívne pre domény vyskytujúce sa viac v plazmidoch a negatívne pre vyskytujúce sa skôr v chromozómoch. Analogicky je vypočítané chromozómové skóre.

Pre každý kontig sa vypočíta plazmidové a chromozómové skóre ako súčet skóre domén, ktoré sa ku kontigu zarovnali. Tieto dve hodnoty sú pridané ako dva nové parametre pre každý kontig a vstupujú do neurónovej siete.

3.2 Naša implementácia

V našej implementácii sa chceme zamerať na viac homologických čŕt a zahrnúť informáciu o jednotlivých doménach, ktoré sa v kontigu nachádzajú. Vďaka tomu neurónová sieť bude mať informáciu o konkrétnych doménach nachádzajúcich sa v kontigoch, pričom počas trénovania môže zistiť, že tieto sekvencie sú špecifické pre konkrétne molekuly – chromozómy a plazmidy.

To docielime pridaním D parametrov, kde D je počet použitých domén dvomi prístupmi. V prvom prístupe bude platiť $x_d = 1$, ak sa doména d zarovnala ku kontigu. Ak sa nezarovnala tak, $x_d = 0$. V druhom prístupe bude $x_d \in [0, 1]$ reprezentovať percento identity, ktoré pochádza z hodnoty $pIdent$ z programu BLAST, teda $x_d = pIdent$. Ak sa doména zarovnala ku kontigu viackrát, vzali sme najvyššiu hodnotu $pIdent$ a ak sa doména ku kontigu nezarovnala, tak platí $x_d = 0$.

Ďalším krokom je výber domén, ktoré použijeme. Zvolili sme opäť dva prístupy: použitie log-odd skóre a Fisherov exaktný test pre každú doménu z nástroja BLAST. Nulová hypotéza je, že doména nie je špecifická pre danú molekulu (chromozóm alebo plazmid). Inak povedané, pravdepodobnosť pozorovania domény v danom plazmide je rovná pravdepodobnosti pozorovania v chromozóme a naopak. Alternatívna hypotéza je, že doméne je špecifická pre danú molekulu. Na výpočet testu pre plazmidové domény sme použili kontingenčnú tabuľku 3.1 a hladinu významnosti $\alpha = 0,05$. Vo výpočte početností domén v chromozómoch a plazmidoch sme nepoužili pseudopočty. Pre chromozómové domény sme použili analogickú tabuľku.

Na základe oboch prístupov vyberieme niekoľko najlepších domén pre plazmidy aj chromozómy. Podľa log-odd skóre použijeme n domén s najvyšším skóre a podľa

	Počet v plazmidoch	Počet v chromozómoch
Doména d	$x_{P,d}$	$x_{C,d}$
Ostatné domény	$\sum_{d'} x_{P,d'} - x_{P,d}$	$\sum_{d'} x_{C,d'} - x_{C,d}$

Tabuľka 3.1: Kontingenčná tabuľka použitá pre Fisherov exaktný test

Fisherovho testu n domén s najnižšou p-hodnotou.

V porovnaní s nástrojom Deeplasmid [4], kde tiež použili binárne vektory označujúce výskyt domény v kontigu, domény vybrali iba na základe ich frekvencie v chromozómoch a plazmidoch. V nástroji PlasForest [46] vypočítali na základe prekryvu sekvencií a domén nové parametre pre každý kontig, ale iba v podobe kumulatívnych hodnôt.

3.3 Výsledky

Enterococcus faecium

Na vývoj a trénovanie homológie sme použili grafy zostavenia genómu z patogénnej baktérie *E. faecium*, ktoré sme bližšie popísali v časti 2.4. Na experimenty sme použili základnú verziu programu plASgraph2 bez použitia extremít, popísanú v časti 1.5 s nastavením pre pohyblivý prah.

Log-odd skóre sme vypočítali na základe všetkých trénovacích dát a 2 116 domén, ktorých zarovnanie k sekvenciám *E. faecium* z nástroja BLAST malo e-hodnotu nižšiu ako $0,001$. Výsledky experimentov môžeme vidieť v tabuľke 3.2.

Na základe log-odd skóre sme vybrali 100 domén s najvyšším plazmidovým a 100 domén s najvyšším chromozómovým skóre, spolu 200 domén (experiment log-odd-100). Porovnali sme s výberom rovnakého počtu domén s najnižšou p hodnotou (experiment fisher-100). Všetky výsledky pre chrozmómy aj plazmidy boli lepšie v prospech výberu podľa Fisherovho testu.

Ďalej sme skúsili vybrať všetky domény, ktoré mali p-hodnotu nižšiu ako hladinu významnosti $\alpha = 0,05$ – 1 345 chromozómových domén a 333 plazmidových domén (experiment fisher-05). Rovnaké počty chromozómových aj plazmidových domén sme vybrali aj na základe log-odd skóre (experiment log-odd-fisher). Opäť mal lepšie výsledky experiment s použitím domén podľa Fishera.

Použitie hodnoty $pIDent$ miesto binárnych vektorov sme skúsili pre všetky predchádzajúce experimenty. V tabuľke ich môžeme vidieť ako experimenty s označením pId. Výsledky boli podobné, v niektorých prípadoch sa mierne zlepšili v porovnaní s experimentmi s binárnymi vektormi.

V porovnaní s pôvodnou verziou (experiment orig) sa zlepšili všetky metriky. Okrem

pôvodnej verzie bez homológie sme skúsili klasifikáciu s celkovým log-odd skóre vypočítaným pre každý kontig (experiment log-odd). V porovnaní s výberom podľa Fishera sa zlepšili metriky návratnosť, F1 skóre a presnosť pre chromozóm, ale pre plazmid boli všetky nižšie.

Na ďalšie testovanie na ESKAPEE dátach a následnom spojení homológie s extremitami sme zvolili výber domén podľa Fishera s $p < 0,05$. Výber domén podľa Fishera s pIdent dosiahol pre chromozómy trochu lepšie výsledky oproti binárnym vektorom, ale pre plazmidy sa zlepšila iba návratnosť a presnosť sa zhoršila.

Experiment	Mol.	AUC	Precíz.	Návrat.	F1	Presn.	Tr. ch.	Val. ch.
orig	plazm.	0,9256	0,8153	0,7501	0,7814	0,8542	2191,78	537,65
orig	chrom	0,9385	0,9179	0,9607	0,9388	0,9023	2191,78	537,65
log-odd	plazm.	0,9660	0,8700	0,8830	0,8764	0,9147	1219,02	334,63
log-odd	chrom	0,9771	0,9563	0,9725	0,9643	0,9432	1219,02	334,63
log-odd-100	plazm.	0,9673	0,8639	0,8854	0,8745	0,9130	1018,56	302,12
log-odd-100	chrom	0,9717	0,9440	0,9721	0,9578	0,9326	1018,56	302,12
fisher-100	plazm.	0,9724	0,8703	0,9057	0,8877	0,9213	1280,90	318,14
fisher-100	chrom	0,9770	0,9485	0,9744	0,9613	0,9381	1280,90	318,14
log-odd-fisher	plazm.	0,9692	0,8759	0,8803	0,8781	0,9161	1418,12	364,10
log-odd-fisher	chrom	0,9759	0,9564	0,9648	0,9606	0,9377	1418,12	364,10
fisher-05	plazm.	0,9730	0,8799	0,8944	0,8871	0,9220	1140,30	364,98
fisher-05	chrom	0,9780	0,9594	0,9656	0,9625	0,9408	1140,30	364,98
log-odd-100-pId	plazm.	0,9663	0,8706	0,8740	0,8723	0,9123	1237,77	309,08
log-odd-100-pId	chrom	0,9718	0,9583	0,9652	0,9617	0,9394	1237,77	309,08
fisher-100-pId	plazm.	0,9709	0,8924	0,8734	0,8828	0,9204	992,58	326,98
fisher-100-pId	chrom	0,9749	0,9506	0,9797	0,9649	0,9439	992,58	326,98
log-odd-fisher-pId	plazm.	0,9658	0,8452	0,8902	0,8671	0,9064	935,03	352,56
log-odd-fisher-pId	chrom	0,9768	0,9635	0,9630	0,9632	0,9421	935,03	352,56
fisher-05-pId	plazm.	0,9717	0,8565	0,9039	0,8796	0,9154	1267,06	367,77
fisher-05-pId	chrom	0,9789	0,9620	0,9652	0,9636	0,9426	1267,06	367,77

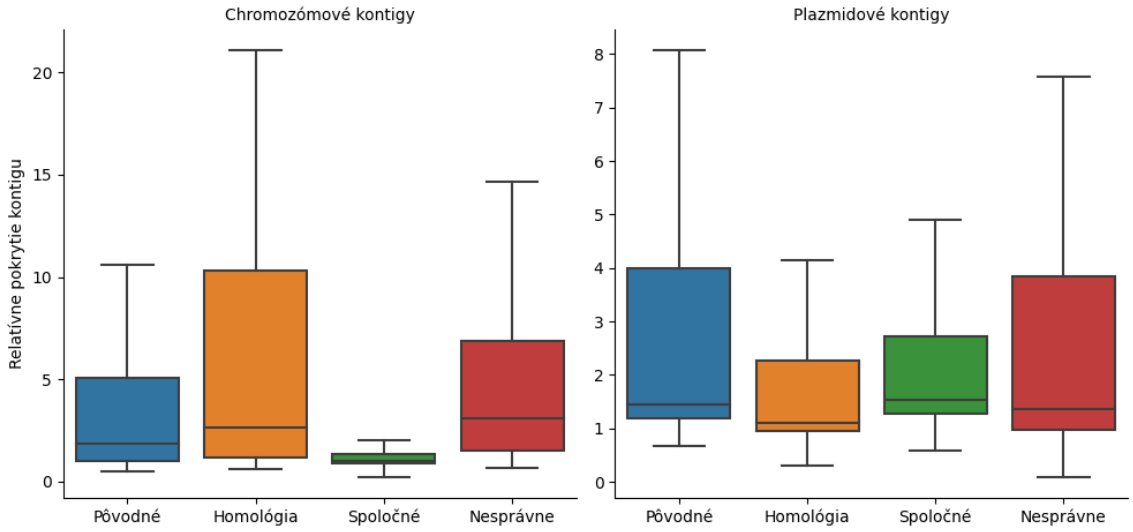
Tabuľka 3.2: Výsledky experimentov s použitím homológie

Pozreli sme sa bližšie na vlastnosti chromozómových a plazmidových kontigov, ktorým homológia pomohla. Rozdiely sme vizualizovali pomocou boxplotov bez odľahlých hodnôt. Kontigy sme rozdelili do 4 skupín – kontigy správne určené iba pôvodným programom (Pôvodné), iba s pomocou homológie (Homológia), množinu spoločných (Spoločné) a nesprávne určené kontigy (Nesprávne).

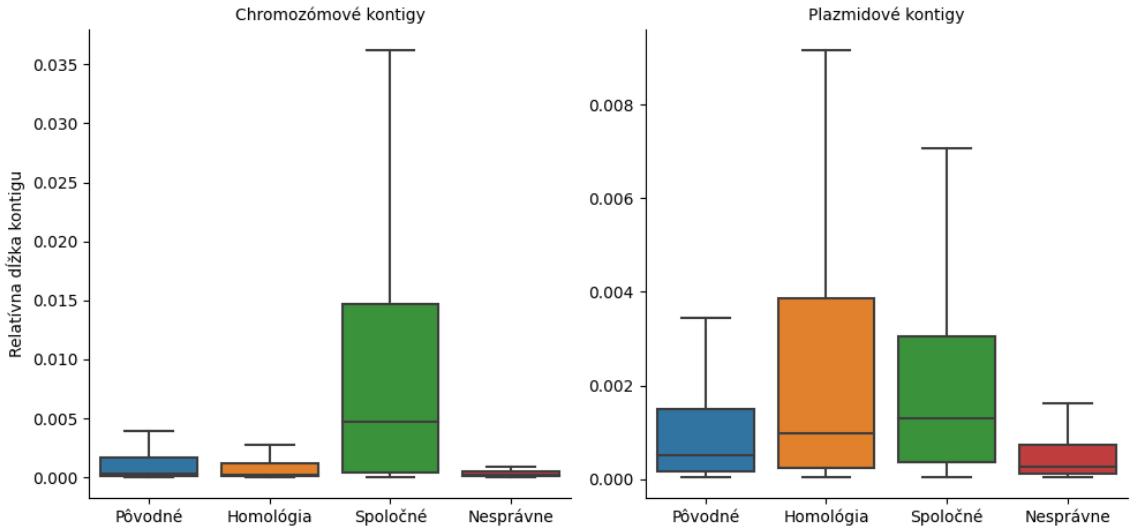
Na obrázku 3.1 môžeme vidieť porovnanie pre relatívne pokrytie, ktoré sme bližšie vysvetlili v časti 1.5. Použitie homológie zvýšilo správne určenie chromozómových

kontigov s vyšším pokrytím a plazmidových kontigov s nižším pokrytím. Vzhľadom na použitú normalizáciu majú typické chromozomálne kontigy pokrytie blízko jedna a teda tie s vyšším pokrytím sú pre neurónovú sieť bez homológie ťažšie klasifikovateľné.

Taktiež homológia pomohla skôr kratším chromozómovým kontigom v porovnaní iba s pôvodným programom, ale v prípade plazmidov lepšie určila dlhšie kontigy ako môžeme vidieť na obrázku 3.2.



Obr. 3.1: Relatívne pokrytie kontigu v testovacích dátach *E. faecium*



Obr. 3.2: Relatívna dĺžka kontigu v testovacích dátach *E. faecium*. Predstavuje dĺžku kontigu vydelenú dvomi miliónmi

Vybrané domény

Pozreli sme sa bližšie na domény, ktoré jednotlivé metódy vybrali na dátach *E. faecium*. Prvých 100 plazmidových domén podľa log-odd skóre a podľa Fisherovho testu sa líši o 34 domén a majú spoločných 66 domén. Prvých 100 chromozómových domén podľa log-odd skóre a podľa Fisherovho testu sa líši o 44 domén a majú spoločných 56 domén.

Prvých 1 345 chromozómových domén podľa log-odd skóre a podľa Fisherovho testu sa líši o 39 domén. Prvých 333 plazmidových domén podľa log-odd skóre a podľa Fisherovho testu sa líši o 12 domén. Domény, ktoré sú odlišné vo výbere najlepších 100 sa nachádzajú vo väčšom výbere a žiadna z nich nie je iba v jednej skupine.

Ďalej v tabuľke 3.3 uvádzame zoznam najlepších domén pre plazmidy a chromozómy vybrané podľa Fisherovho testu. Okrem p-hodnoty môžeme vidieť aj log-odd skóre.

Plazmidové domény			Chromozómové domény		
Doména	Log-odd	p-hodnota	Doména	Log-odd	p-hodnota
pfam00239	2,3175	$7{,}4811\times10^{-315}$	pfam00005	0,0397	$4{,}0144\times10^{-31}$
pfam00872	1,9242	$1{,}5677\times10^{-266}$	pfam13520	0,0705	$8{,}6264\times10^{-25}$
pfam13610	2,3757	$1{,}5330\times10^{-235}$	pfam00324	0,0715	$2{,}1804\times10^{-24}$
pfam00665	1,9455	$8{,}8547\times10^{-230}$	pfam00664	0,0703	$7{,}9170\times10^{-24}$
pfam02796	2,5209	$2{,}7123\times10^{-203}$	pfam07992	0,0651	$7{,}4927\times10^{-23}$

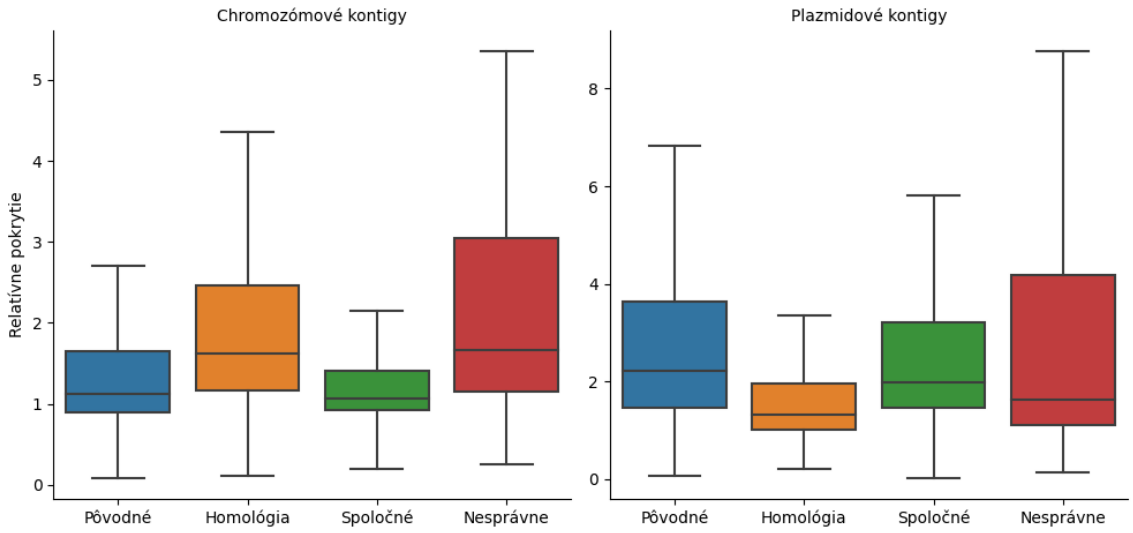
Tabuľka 3.3: Zoznam domén s najnižšou p-hodnotou

ESKAPEE

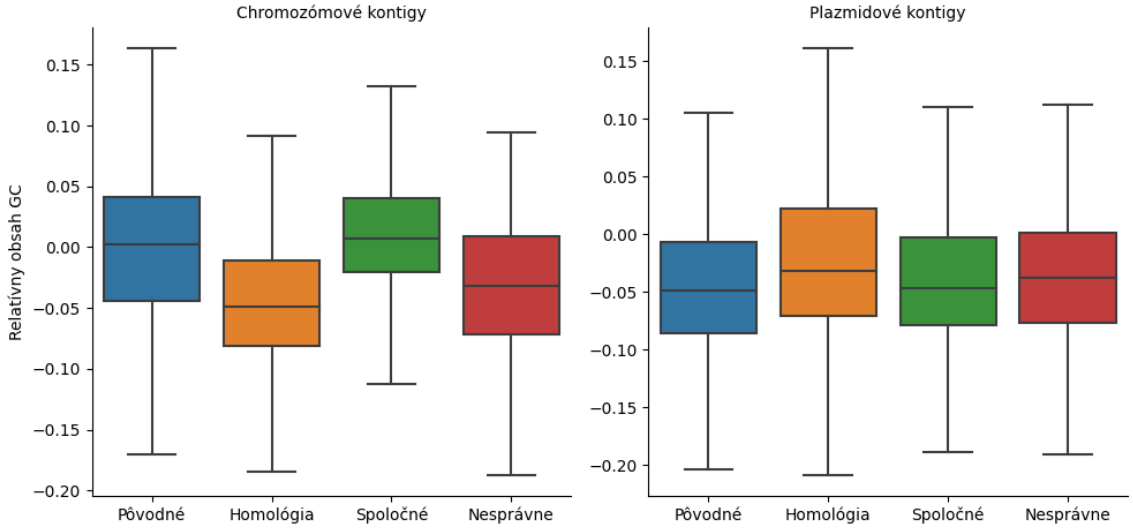
Prístup s homológiou sme vyhodnotili aj na dátach ESKAPEE, ktoré sme bližšie popísali v časti 2.4. Na základe Fisherovho testu sme určili 1 564 chromozómových domén a 730 plazmidových domén. Výsledky môžeme vidieť v tabuľke 3.4. V porovnaní s pôvodnou verziou bez homológie (experiment orig) sa zlepšili všetky metriky. Najvýraznejší posun je v precíznosti a návratnosti pre klasifikáciu plazmidov a zlepšila sa aj presnosť pre chromozómy aj plazmidy.

Experiment	Mol.	AUC	Precíz.	Návrat.	F1	Presn.	Tr. ch.	Val. ch.
orig	plazm.	0,9256	0,6645	0,7171	0,6898	0,9096	4510,11	1038,85
orig	chrom	0,9359	0,9570	0,9620	0,9595	0,9272	4510,11	1038,85
hmg	plazm.	0,9656	0,7975	0,8308	0,8138	0,9467	1597,77	651,22
hmg	chrom	0,9678	0,9680	0,9775	0,9727	0,9509	1597,77	651,22

Tabuľka 3.4: Výsledky experimentov s použitím homológie na dátach ESKAPEE



Obr. 3.3: Relatívne pokrytie kontigu v testovacích dátach ESKAPEE



Obr. 3.4: Relatívny obsah guanínu a cytozínu kontigov v testovacích dátach ESKAPEE

Pozreli sme sa tiež na vlastnosti kontigov, ktorým pomohlo použitie homológie. Kontigy sme rozdelili do 4 skupín podobne ako v časti výsledkov *E. faecium*.

Na obrázku 3.3 môžeme vidieť porovnanie pre relatívne pokrytie, ktoré sme bližšie vysvetlili v časti 1.5. Použitie homológie zvýšilo správne určenie chromozómových kontigov s vyšším pokrytím a plazmidových kontigov s nižším pokrytím podobne ako aj v dátach *E. faecium*.

Nakoľko ESKAPEE tvorí viac organizmov, ktoré môžu mať odlišný obsah guanínu a cytozínu, pozreli sme sa aj na relatívny obsah GC (bližšie vysvetlený v časti 1.5) v testovacích dátach. Homológia pomohla chromozómovým kontigom s nižším obsahom GC a plazmidom s vyšším obsahom GC ako môžeme vidieť na obrázku 3.4.

3.4 Spojenie extremít a homológie

Na záver sme spojili extremity s homológiou. Výsledky uvádzame v tabuľkách 3.5 pre dáta *E. faecium* a 3.6 pre dáta ESKAPEE. V experimente extr sme použili verziu v1 extremít s nastavením pohyblivého prahu. V oboch prípadoch najlepšie výsledky dosiahla homológia (experiment hmg). Kombinácia homológie a extremít (experiment hmg-extr) zlepšila výsledky v porovnaní iba s použitím extremít na oboch dátových množinách.

Experiment	Mol.	AUC	Precíz.	Návrat.	F1	Presn.	Tr. ch.	Val. ch.
orig	plazm.	0,9256	0,8153	0,7501	0,7814	0,8542	2191,78	537,65
orig	chrom	0,9385	0,9179	0,9607	0,9388	0,9023	2191,78	537,65
extr	plazm.	0,9391	0,7706	0,8534	0,8099	0,8642	2169,06	526,41
extr	chrom	0,9507	0,9460	0,9191	0,9324	0,9019	2169,06	526,41
hmg	plazm.	0,9730	0,8799	0,8944	0,8871	0,9220	1140,30	364,98
hmg	chrom	0,9780	0,9594	0,9656	0,9625	0,9408	1140,30	364,98
hmg-extr	plazm.	0,9635	0,8298	0,9013	0,8640	0,9026	1786,73	464,26
hmg-extr	chrom	0,9707	0,9429	0,9654	0,9540	0,9267	1786,73	464,26

Tabuľka 3.5: Výsledky experimentov s použitím homológie a extremít na dátach *E. faecium*

Experiment	Mol.	AUC	Precíz.	Návrat.	F1	Presn.	Tr. ch.	Val. ch.
orig	plazm.	0,9256	0,6645	0,7171	0,6898	0,9096	4510,11	1038,85
orig	chrom	0,9359	0,9570	0,9620	0,9595	0,9272	4510,11	1038,85
extr	plazm.	0,9362	0,7036	0,7288	0,7160	0,9189	4589,54	1090,57
extr	chrom	0,9429	0,9620	0,9644	0,9632	0,9339	4589,54	1090,57
hmg	plazm.	0,9656	0,7975	0,8308	0,8138	0,9467	1597,77	651,22
hmg	chrom	0,9678	0,9680	0,9775	0,9727	0,9509	1597,77	651,22
hmg-extr	plazm.	0,9290	0,7969	0,7579	0,7769	0,9390	2864,36	914,43
hmg-extr	chrom	0,9327	0,9640	0,9791	0,9715	0,9484	2864,36	914,43

Tabuľka 3.6: Výsledky experimentov s použitím homológie a extremít na dátach ESKAPEE

Záver

Detekcia plazmidov zo sekvenačných dát je dôležitou úlohou vzhľadom na rozšírenie rezistencie voči antibiotikám v baktériách. Naším cieľom bolo rozšírenie programu plASgraph2 slúžiaceho na identifikáciu plazmidov v grafoch zostavenia genómu pochádzajúcich zo sekvenačných dát bakteriálnych izolátov o využitie ďalších informácií.

Prvou časťou bola príprava a implementácia architektúry programu pre zapojenie informácie o tom, ktoré konce susedných kontigov v grafoch sú prepojené. Modifikovali sme štruktúru grafu, kde sme každý vrchol rozdelili na dva, pomocou čoho sme reprezentovali rôzne konce kontigu. Graf obsahoval dve matice susedností, prvá reprezentovala susednosti medzi kontigmi a druhá susednosti extremít v kontigoch. Nová architektúra šíri informáciu najprv medzi kontigmi a potom si túto informáciu odovzdajú extremity medzi sebou. Architektúru sme otestovali na syntetických grafoch, na ktorých sme ukázali, ako sa šíri informácia od jednej extremity k druhej. Extremity mierne zlepšili výsledné metriky v dátach z patogénu *Enterococcus faecium* aj z patogénov ESKAPEE.

Druhou časťou bolo pridanie informácie o homológii (podobnosti vstupných sekvencií so známymi) do parametrov popisujúcich jednotlivé klasifikované sekvencie. Vlastnosti sme pridali v podobe binárnych hodnôt a vyjadrovali prítomnosť proteínových domén v kontigoch. Skúsili sme použiť aj percento identity medzi doménami a vstupnými kontigmi z nástroja BLAST [12]. Zo všetkých nájdených proteínových domén sme pre použitie v našej sieti vybrali také, ktoré majú najlepší potenciál odlišovať chromozómy od plazmidov, pričom sme použili dve metódy: log-odd skóre a Fisherov exaktný test. Homológia očakávane zlepšila skóre na oboch dátových množinách. Spojenie oboch prístupov však nadobudlo horšie výsledky ako samotné použitie homológie, ale lepšie ako použitie iba extremitového prístupu.

Existuje niekoľko možností, ako by sa dal program ďalej rozšíriť. Prvou z nich je použiť pokročilejšie konvolučné alebo iné grafové neurónové siete, napríklad z knižnice Spektral [22]. Ďalšou možnosťou je pridať viac informácií na základe podobností sekvencií. Lokálnu databázu je možné rozšíriť o ďalšie špecifické sekvencie, ktoré by sa pridali ako binárny vektor alebo iným spôsobom. Výber domén môžeme sprísniť pomocou korekcie niekoľkonásobného testovania. Namiesto výskytu jednotlivých domén v kontigoch sa dá pozrieť na výskyt aspoň jednej domény zo skupiny. Skupiny sa dajú

vytvoriť na základe známych proteínových rodín alebo ich charakteristík. Poslednou možnosťou, ktorú spomenieme, je vytvorenie skóre, ktorého hodnoty by boli špecifické pre chromozómy a plazmidy, ako už je napríklad používané log-odd skóre alebo skóre z nástroja Platon [52]. Skóre môže byť založené na podobnosti sekvencií alebo aj iných charakteristikách, ktorými sa plazmidy a chromozómy odlišujú.

Literatúra

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Je□rey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geo□rey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Mislav Acman, Ruobing Wang, Lucy van Dorp, Liam P Shaw, Qi Wang, Nina Luhmann, Yuyao Yin, Shijun Sun, Hongbin Chen, Hui Wang, et al. Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene bla_{NDM}. *Nature communications*, 13(1):1131, 2022.

[3] Sergey Aganezov, Ilya Zban, Vitaly Aksenov, Nikita Alexeev, and Michael C Schatz. Recovering rearranged cancer chromosomes from karyotype graphs. *BMC bioinformatics*, 20:1–11, 2019.

[4] William B Andreopoulos, Alexander M Geller, Miriam Lucke, Jan Balewski, Alicia Clum, Natalia N Ivanova, and Asaf Levy. Deeplasmid: deep learning accurately separates plasmids from bacterial chromosomes. *Nucleic acids research*, 50(3):e17–e17, 2022.

[5] Dmitry Antipov, Mikhail Raiko, Alla Lapidus, and Pavel A Pevzner. Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics*, 36(14):4126–4129, 2020.

[6] Sergio Arredondo-Alonso, Malbert RC Rogers, Johanna C Braat, Tess D Verschuuren, Janetta Top, Jukka Corander, Rob JL Willems, and Anita C Schürch. mlplasmids: A user-friendly tool to predict plasmid-and chromosome-derived sequences for single species. *Microbial genomics*, 4(11), 2018.

[7] Malay Kumar Basu, Eugenia Poliakov, and Igor B Rogozin. Domain mobility in proteins: functional and evolutionary implications. *Briefings in bioinformatics*, 10(3):205–216, 2009.

[8] John E Bennett, Raphael Dolin, and Martin J Blaser. *Mandell, douglas, and bennett's principles and practice of infectious diseases E-book*. Elsevier Health Sciences, 2019.

[9] Ian Boostrom, Edward AR Portal, Owen B Spiller, Timothy R Walsh, and Kirsty Sands. Comparing long-read assemblers to explore the potential of a sustainable low-cost, low-infrastructure approach to sequence antimicrobial resistant bacteria with oxford nanopore sequencing. *Frontiers in Microbiology*, 13:796465, 2022.

[10] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1):59–60, 2015.

[11] Marija Buljan and Alex Bateman. The evolution of protein domain families. *Biochemical Society Transactions*, 37(4):751–755, 2009.

[12] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC bioinformatics*, 10:1–9, 2009.

[13] Alessandra Carattoli. Plasmids and the spread of resistance. *International journal of medical microbiology*, 303(6-7):298–304, 2013.

[14] Agnes P Chan, Yongwook Choi, Thomas H Clarke, Lauren M Brinkac, Richard C White, Michael R Jacobs, Robert A Bonomo, Mark D Adams, and Derrick E Fouts. AbGRI4, a novel antibiotic resistance island in multiply antibiotic-resistant Acinetobacter baumannii clinical isolates. *Journal of Antimicrobial Chemotherapy*, 75(10):2760–2768, 2020.

[15] I-Min A Chen, Ken Chu, Krishnaveni Palaniappan, Anna Ratner, Jinghua Huang, Marcel Huntemann, Patrick Hajek, Stephan J Ritter, Cody Webb, Dongying Wu, Neha J Varghese, T B K Reddy, Supratim Mukherjee, Galina Ovchinnikova, Matt Nolan, Rekha Seshadri, Simon Roux, Axel Visel, Tanja Woyke, Emiley A Eloe-Fadrosh, Nikos C Kyrpides, and Natalia N Ivanova. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Research*, 51(D1):D723–D732, 11 2022.

[16] David P. Clark, Nanette J. Pazdernik, and Michelle R. McGehee. Chapter 23 - Plasmids. In David P. Clark, Nanette J. Pazdernik, and Michelle R. McGehee,

editors, *Molecular Biology (Third Edition)*, pages 712–748. Academic Cell, third edition edition, 2019.

[17] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

[18] Sara El-Metwally, Taher Hamza, Magdi Zakaria, and Mohamed Helmy. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS computational biology*, 9(12):e1003345, 2013.

[19] Zhencheng Fang, Jie Tan, Shufang Wu, Mo Li, Congmin Xu, Zhongjie Xie, and Huaiqiu Zhu. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, 8(6):giz066, 2019.

[20] Laura S Frost, Raphael Leplae, Anne O Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005.

[21] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics*, 17(6):333–351, 2016.

[22] Daniele Grattarola and Cesare Alippi. Graph neural networks in tensorflow and keras with spektral [application notes]. *IEEE Computational Intelligence Magazine*, 16(1):99–106, 2021.

[23] Nature Publishing Group. plasmid / plasmids. `https://www.nature.com/scitable/definition/plasmid-plasmids-28/`, 2014. Retrieved Nov 11, 2022.

[24] The GFA Format Specification Working Group. Graphical fragment assembly (GFA) format specification. `https://gfa-spec.github.io/GFA-spec/GFA1.html`, Jun 2022. Retrieved May 09, 2024.

[25] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[26] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[27] Haruka Hasegawa, Erika Suzuki, and Sumio Maeda. Horizontal plasmid transfer by transformation in Escherichia coli: environmental factors and possible mechanisms. *Frontiers in microbiology*, 9:2365, 2018.

[28] Miyako Hikichi, Miki Nagao, Kazunori Murase, Chihiro Aikawa, Takashi Nozawa, Akemi Yoshida, Taisei Kikuchi, and Ichiro Nakagawa. Complete genome sequences of eight methicillin-resistant staphylococcus aureus strains isolated from patients in japan. *Microbiology Resource Announcements*, 8(47):10–1128, 2019.

[29] National Human Genome Research Institute Home. Chromosome. https://www.genome.gov/genetics-glossary/Chromosome. Retrieved Jan 10, 2023.

[30] National Human Genome Research Institute Home. Genome. https://www.genome.gov/genetics-glossary/Genome. Retrieved Jan 10, 2023.

[31] National Human Genome Research Institute Home. Plasmid. https://www.genome.gov/genetics-glossary/Plasmid. Retrieved Jan 10, 2023.

[32] Laura M Kasman and La Donna Porter. Bacteriophages. In *StatPearls [Internet]*. StatPearls Publishing, 2022.

[33] Thomas Kipf. How powerful are graph convolutional networks?, Sep 2016.

[34] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[35] Pawel S Krawczyk, Leszek Lipinski, and Andrzej Dziembowski. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic acids research*, 46(6):e35–e35, 2018.

[36] Miriam Land, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, Ole Lund, Guruprased Kora, Trudy Wassenaar, et al. Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, 15:141–161, 2015.

[37] Bárbara Magalhães, Laurence Senn, and Dominique S Blanc. High-quality complete genome sequences of three pseudomonas aeruginosa isolates retrieved from patients hospitalized in intensive care units. *Microbiology resource announcements*, 8(9):10–1128, 2019.

[38] ROBERT V Masterson, PAUL R Russell, and ALAN G Atherly. Nitrogen fixation (nif) genes and large plasmids of Rhizobium japonicum. *Journal of Bacteriology*, 152(2):928–931, 1982.

[39] Jason R Miller, Sergey Koren, and Granger Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[40] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. Pfam: The protein families database in 2021. *Nucleic acids research*, 49(D1):D412–D419, 2021.

[41] Hisaya K Ono, Yasunori Suzuki, Hiroaki Kubota, Krisana Asano, Shinji Takai, Akio Nakane, and Dong-Liang Hu. Complete genome sequence of Staphylococcus aureus strain 834, isolated from a septic patient in Japan. *Microbiology Resource Announcements*, 10(9):10–1128, 2021.

[42] Alex Orlek, Nicole Stoesser, Muna F Anjum, Michel Doumith, Matthew J Ellington, Tim Peto, Derrick Crook, Neil Woodford, A Sarah Walker, Hang Phan, et al. Plasmid classification in an era of whole-genome sequencing: application in studies of antibiotic resistance epidemiology. *Frontiers in microbiology*, 8:182, 2017.

[43] Chandan Pal, Johan Bengtsson-Palme, Erik Kristiansson, and DG Larsson. Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC genomics*, 16(1):1–14, 2015.

[44] David Pellow, Itzik Mizrahi, and Ron Shamir. PlasClass improves plasmid sequence classification. *PLoS computational biology*, 16(4):e1007781, 2020.

[45] Silke Peter, Mattia Bosio, Caspar Gross, Daniela Bezdan, Javier Gutierrez, Philipp Oberhettinger, Jan Liese, Wichard Vogel, Daniela Dörfel, Lennard Berger, et al. Tracking of antibiotic resistance transfer and rapid plasmid evolution in a hospital setting by Nanopore sequencing. *Msphere*, 5(4):e00525–20, 2020.

[46] Léa Pradier, Tazzio Tissot, Anna-Sophie Fiston-Lavier, and Stéphanie Bedhomme. PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets. *BMC bioinformatics*, 22(1):1–17, 2021.

[47] Lianrong Pu and Ron Shamir. 3CAC: improving the classification of phages and plasmids in metagenomic assemblies using assembly graphs. *Bioinformatics*, 38(Supplement_2):ii56–ii61, 2022.

[48] Lianrong Pu and Ron Shamir. 4CAC: 4-class classification of metagenome assemblies using machine learning and assembly graphs. *bioRxiv*, pages 2023–01, 2023.

[49] Maria S Ramirez, German M Traglia, David L Lin, Tung Tran, and Marcelo E Tolmasky. Plasmid-mediated antibiotic resistance and virulence in gram-negatives: the Klebsiella pneumoniae paradigm. *Microbiology spectrum*, 2(5):2–5, 2014.

[50] Daniel J Rankin, Eduardo PC Rocha, and Sam P Brown. What traits are carried on mobile genetic elements, and why? *Heredity*, 106(1):1–10, 2011.

[51] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[52] Oliver Schwengers, Patrick Barth, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty, and Alexander Goesmann. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microbial genomics*, 6(10), 2020.

[53] Nidhi Shah, Michael G Nute, Tandy Warnow, and Mihai Pop. Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows. *Bioinformatics*, 35(9):1613–1614, 2019.

[54] Liam P Shaw, Kevin K Chau, James Kavanagh, Manal AbuOun, Emma Stubberfield, H Soon Gweon, Leanne Barker, Gillian Rodger, Mike J Bowes, Alasdair TM Hubbard, et al. Niche and local geography shape the pangenome of wastewater- and livestock-associated Enterobacteriaceae. *Science advances*, 7(15):eabe3868, 2021.

[55] Masaki Shintani, Zoe K Sanchez, and Kazuhide Kimbara. Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy. *Frontiers in microbiology*, 6:242, 2015.

[56] Janik Sielemann, Katharina Sielemann, Broňa Brejová, Tomáš Vinař, and Cedric Chauve. plASgraph2: using graph neural networks to detect plasmid contigs from an assembly graph. *Frontiers in Microbiology*, 14:1267695, 2023.

[57] Masahiro Sota. Horizontal gene transfer mediated by plasmids. *Plasmids: current research and future trends*, pages 111–182, 2008.

[58] Alexandre Souvorov, Richa Agarwala, and David J Lipman. SKESA: strategic k-mer extension for scrupulous assemblies. *Genome biology*, 19(1):153, 2018.

[59] Alessandro Sperduti and Antonina Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8(3):714–735, 1997.

[60] Linda van der Graaf-Van Bloois, Jaap A Wagenaar, and Aldert L Zomer. RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning. *Microbial genomics*, 7(11), 2021.

[61] Jiyao Wang, Farideh Chitsaz, Myra K Derbyshire, Noreen R Gonzales, Marc Gwadz, Shennan Lu, Gabriele H Marchler, James S Song, Narmada Thanki, Roxanne A Yamashita, et al. The conserved domain database in 2023. *Nucleic Acids Research*, 51(D1):D384–D388, 2023.

[62] Yan Wang, Hang Zhang, Haolin Zhong, and Zhidong Xue. Protein domain identification methods and online resources. *Computational and structural biotechnology journal*, 19:1145–1153, 2021.

[63] Ryan R Wick, Louise M Judd, Claire L Gorrie, and Kathryn E Holt. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS computational biology*, 13(6):e1005595, 2017.

[64] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[65] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[66] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for tra□c forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[67] Ziwei Zhang, Peng Cui, and Wenwu Zhu. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):249–270, 2020.

[68] Fengfeng Zhou and Ying Xu. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*, 26(16):2051–2052, 2010.

[69] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

Príloha

Elektronické prílohy so zdrojovými kódmi a príkladom vstupných dát sa nachádzajú na priloženom pamäťovom médiu.