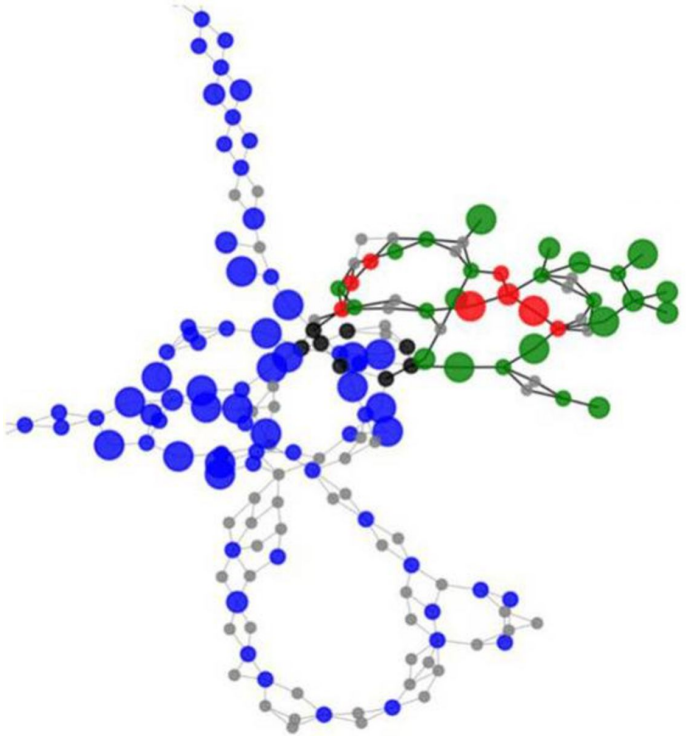
# Detekcia bakteriálnych plazmidov pomocou grafových neurónových sietí

Školiteľka: doc. Mgr. Bronislava Brejová, PhD.
Veronika Tordová

# Základné definície



Kontig

- dlhší súvislý úsek reprezentujúci časť DNA pozostávajúci z prekrývajúcich sa čítaní

Skladanie genómu (genome assembly)

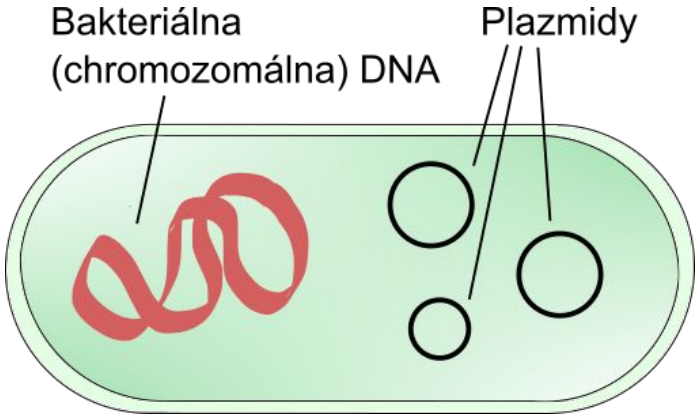
- skladanie krátkych prekrývajúcich sa čítaní do dlhších úsekov

Graf zostavenia genómu (assembly graph)

- vizualizácia poskladaného genómu
- vrcholy sú kontigy a hrany predstavujú možné prepojenia medzi nimi

# Plazmid

- malá molekula DNA uzavretá do kruhu
- oddelená od chromozomálnej DNA
- má charakteristické vlastnosti
- nesie rôzne gény, napr. gény rezistencie voči antibiotiku



# Cieľ práce

Cieľom práce je rozšíriť parametre existujúceho programu plASgraph2 extrakciou informácií zo vstupných dát.

# Klasifikácia plazmidových kontigov

Vstup

- kontigy pochádzajúce z bakteriálnych izolátov alebo metagenómu

Cieľ

- určenie pôvodu kontigov – či pochádzajú z plazmidov, baktérií alebo ďalších organizmov z metagenomickej vzorky

# Vstupné parametre a metódy

Vstupné parametre

- vlastnosti sekvencií – dĺžka sekvencie, obsah guanínu a cytozínu, podreťazce dĺžky k
- homológia – hľadanie podobností

Metódy

- metódy hlbokého učenia (Deeplasmid 2022), random forest (PlasForest 2021)
- väčšina nástrojov klasifikuje vrchol izolovane
- použitie informácie zo susedných vrcholov

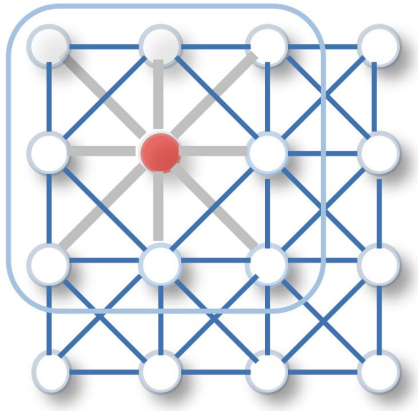
# plASgraph2: using GNNs to detect plasmid contigs from an assembly graph

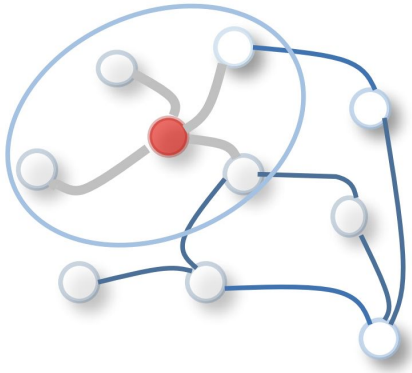
*2023, Janik Sielemann, Katharina Sielemann, Broňa Brejová, Tomáš Vinař, Cedric Chauve*

- nástroj na identifikáciu plazmidových kontigov pomocou grafových neurónových sietí
- používa grafové konvolučné siete a šíri informáciu z vrchola do susedných vrcholov
- vstupné parametre založené na vlastnostiach sekvencií
  - stupeň vrchola v grafe zostavenia genómu
  - dĺžka kontigu vydelená dvomi miliónmi
  - logaritmus dĺžky kontigu
  - relatívne pokrytie
  - relatívny obsah GC
  - relatívny obsah podreťazcov dĺžky k = 5

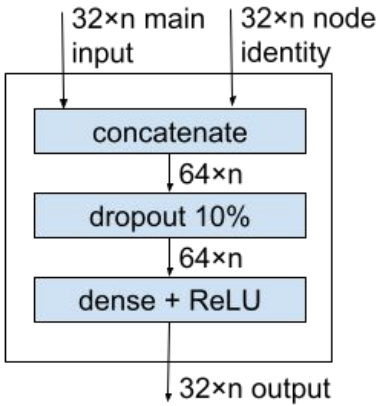
# Grafové konvolučné siete

- grafové neurónové siete
- vstup – zoznam vrcholov s parametrami a matica susednosti
- matica susednosti miesto mriežky
- L vrstiev → rozšírenie informácie do susedov vo vzdialenosti L

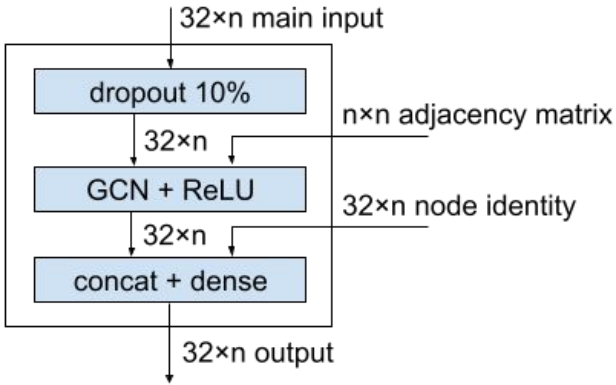




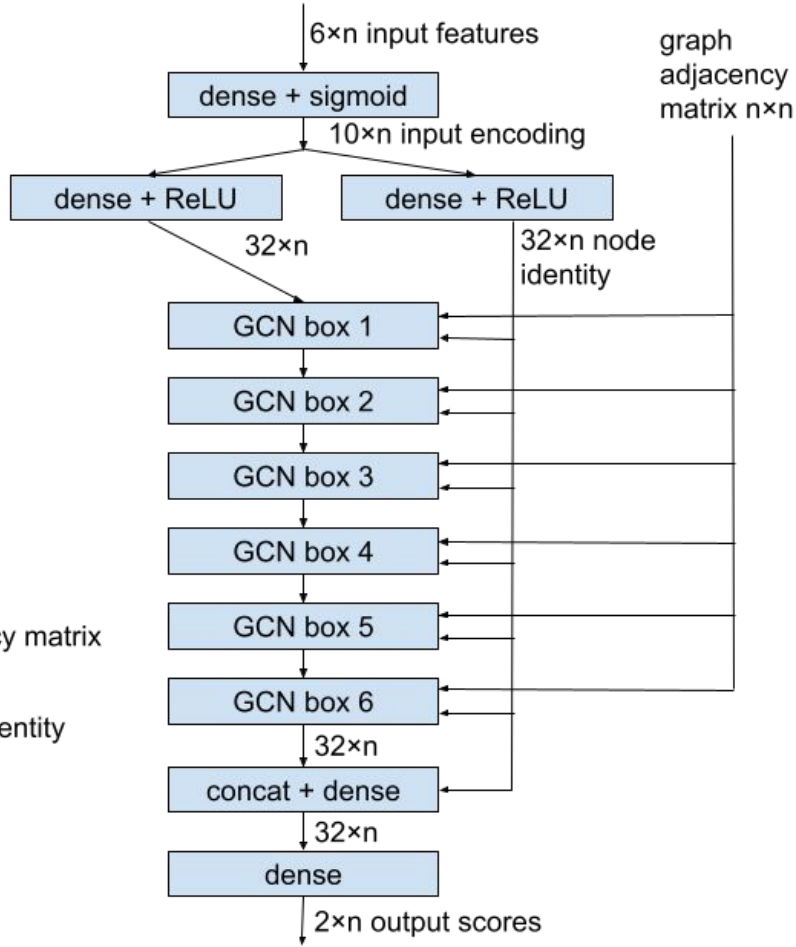
**Concat + dense box:**




**GCN box:**




**Overall architecture:**



# plASgraph2: using GNNs to detect plasmid contigs from an assembly graph

*Janik Sielemann, Katharina Sielemann, Broňa Brejová, Tomáš Vinař, Cedric Chauve*

- výsledkom je n dvojíc skóre – matica n x 2
- 4 triedy – plazmid, chromozóm, nejednoznačný, neoznačený
- klasifikačná úloha je rozdelená na dve – klasifikácia kontigu ako plazmidu a chromozómu

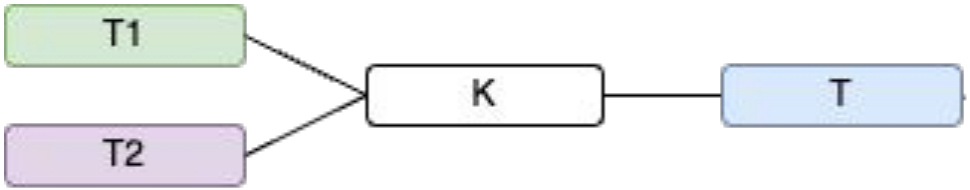
# 1. Rozšírenie – Extremity

Motivácia

- pri vytvorení grafu program nerozlišuje, či je susedný kontig pripojený z ľavej alebo z pravej strany kontigu, pracuje iba s celkovým počtom susedov

Myšlienka

- ak má kontig K na jednom konci jedného suseda typu T a na druhom konci má susedov dvoch a viac typov T1, T2, ..., tak je pravdepodobné, že daný kontig K bude toho istého typu ako je jeho sused T



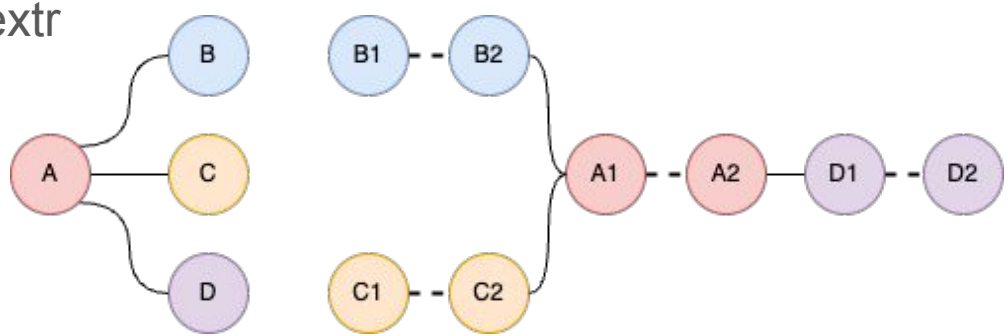
# Návrh architektúry

Pôvodná architektúra

- každý kontig je 1 vrchol → n vrcholov

Nová architektúra – graf susednosti intervalov (ang. Interval Adjacency Graph)

- každý kontig sú 2 vrcholy – extremity → 2n vrcholov
- každá extremita má rovnaké vl. ako pôvodný kontig, no iný počet susedov
- 2 matice susedností A a A_extr

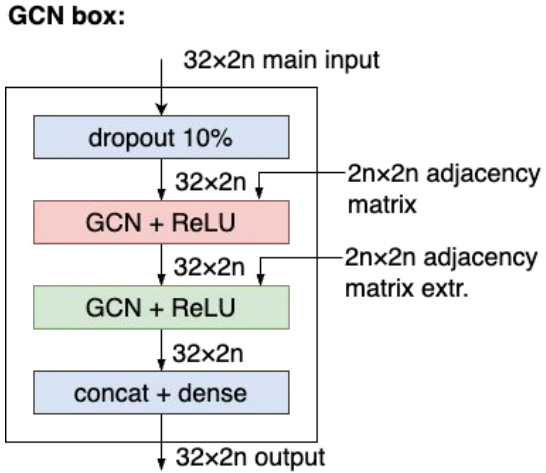


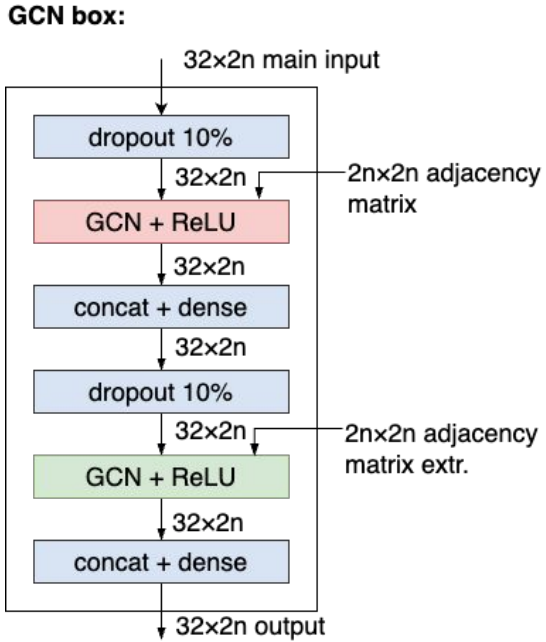
# Ako dostaneme n výsledných dvojíc skóre?

- spriemerovanie výsledných skóre
  - výsledkom architektúry je 2n dvojíc skóre
  - pred testovaním sa spriemerujú
- zmena rozmeru matice (reshape)
  - v trénovaní zmeníme maticu z 2n x k na n x 2k pred vstupom do poslednej plne prepojenej vrstvy

# Ako dve matice susedností použijeme?

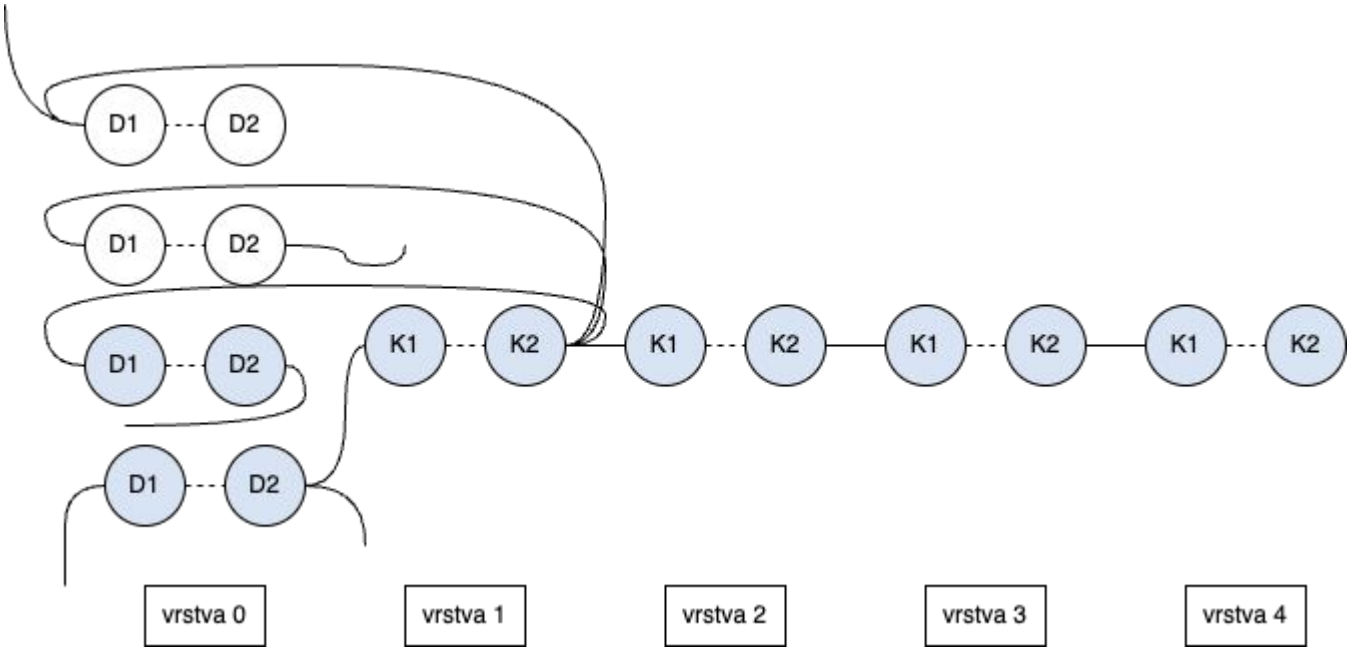
- použitie dvoch grafových konvolučných vrstiev
- rovnaké/rôzne parametre



# Testovanie architektúry

- vygenerovali sme syntetické dáta
- dlhé a krátke vrcholy typu plazmid a chromozóm s rozdielnym obsahom GC



# Testovanie architektúry

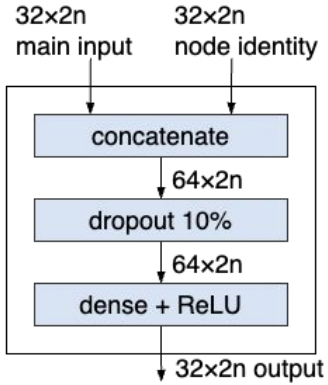
- 10 umelých grafov
- porovnali sme výsledky pôvodného programu s extremitovým
- pôvodný program nevedel správne určiť skóre kontigov – priemerné skóre krátkych kontigov sa pohybovalo okolo 0,5
- extremitový program správne určil všetky kontigy

# Výsledky *E. faecium*

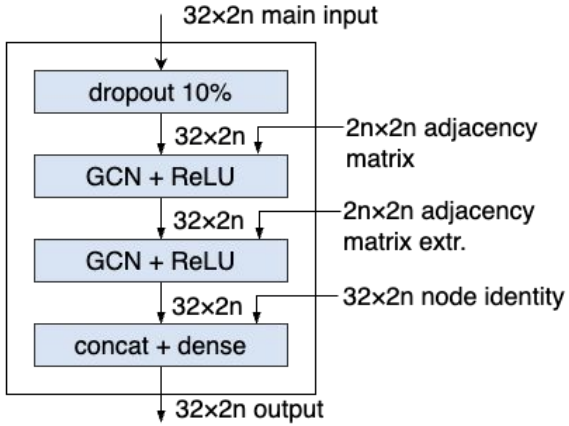
- 46 grafov, 13674 kontigov dlhších ako 100 bp,
- 1 658 nejednoznačných vrcholov, 8 260 chromozómov, 2 826 plazmidov a 930 neoznačených vrcholov
- testovacie dáta obsahujú 60 grafov

| Experiment | Molekula | AUC | Precision | Recall | F1 | Trén. chyba | Val. chyba |
|---|---|---|---|---|---|---|---|
| orig | plazm. | 0,9256 | 0,7636 | 0,8333 | 0,7970 | 2191,78 | 537,65 |
| orig | chrom. | 0,9385 | 0,9328 | 0,9439 | 0,9383 | 2191,78 | 537,65 |
| extr | plazm. | **0,9391** | **0,7932** | 0,8181 | **0,8054** | **2169,06** | **526,41** |
| extr | chrom. | **0,9507** | 0,9251 | **0,9525** | **0,9386** | **2169,06** | **526,41** |

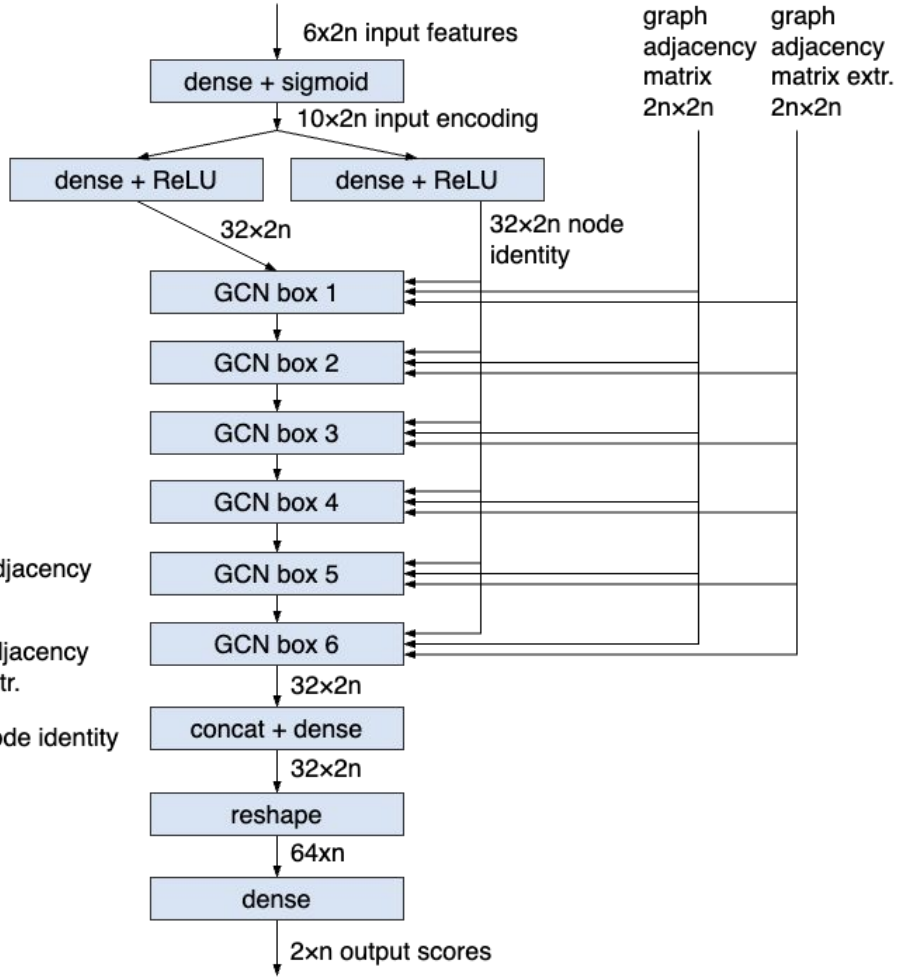
**Concat + dense box:**




**GCN box:**




**Overall architecture:**



# Výsledky ESKAPEE

- trénovaciu množinu tvorí 140 grafov a testovaciu 224 grafov

| Experiment | Molekula | AUC | Precision | Recall | F1 | Trén. chyba | Val. chyba |
|---|---|---|---|---|---|---|---|
| orig | plazm. | 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 | plazm. | **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 |

# 2. Rozšírenie – Homológia

- princíp 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
- Deeplasmid (2022), Platon (2020)

# Homológia – ako použijeme informácie?

- chceme použiť viac homologických čŕt a zahrnúť informáciu o jednotlivých doménach, ktoré sa v kontigu nachádzajú
- pridanie D parametrov
- binárne – doména sa nachádza alebo nenachádza v kontigu
- reálne v intervale [0,1] – percento identity
- napr. program Deeplasmid (2022)

# Homológia – ako domény vyberieme?

Log odd skóre

- čím vyššie skóre tým špecifickejšie pre molekulu

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

Fisherov exaktný test

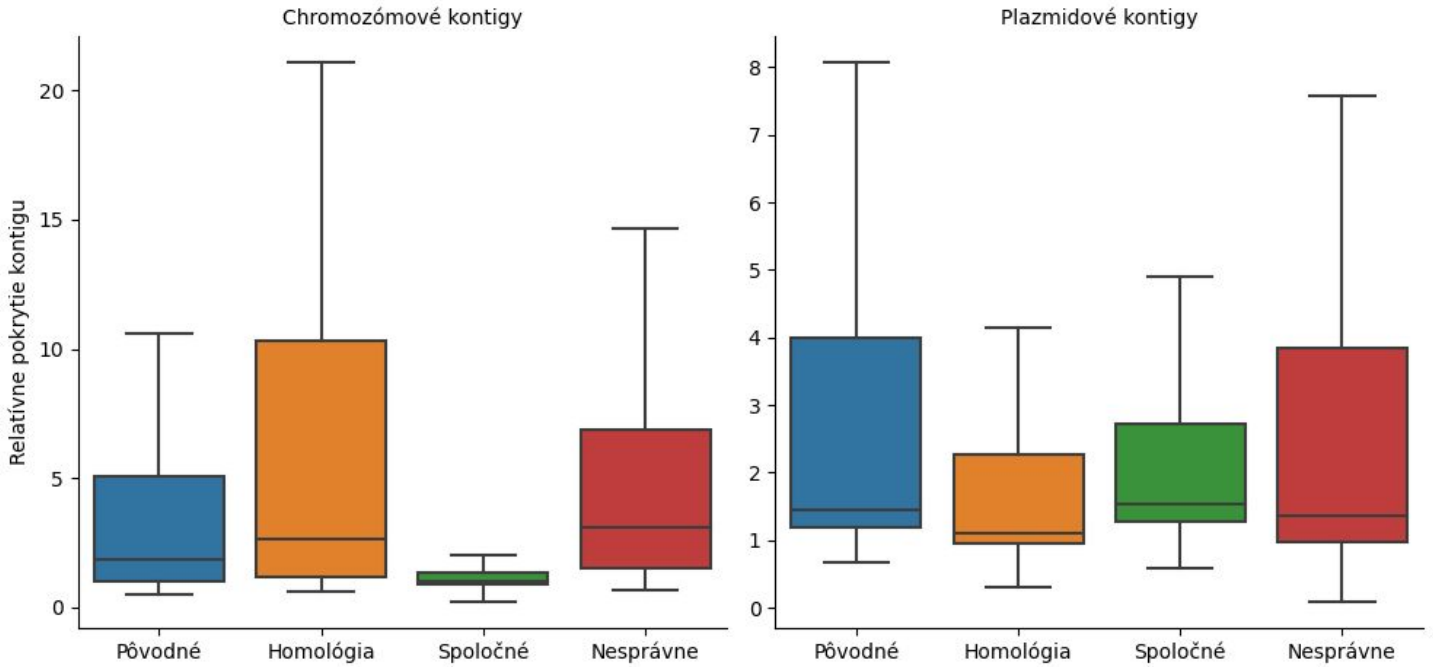
- H0: doména nie je špecifická pre danú molekulu – pravdepodobnosť pozorovania domény v danom plazmide je rovná pravdepodobnosti pozorovania v chromozóme a naopak
- 1564 chromozómových a 730 plazmidových domén

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

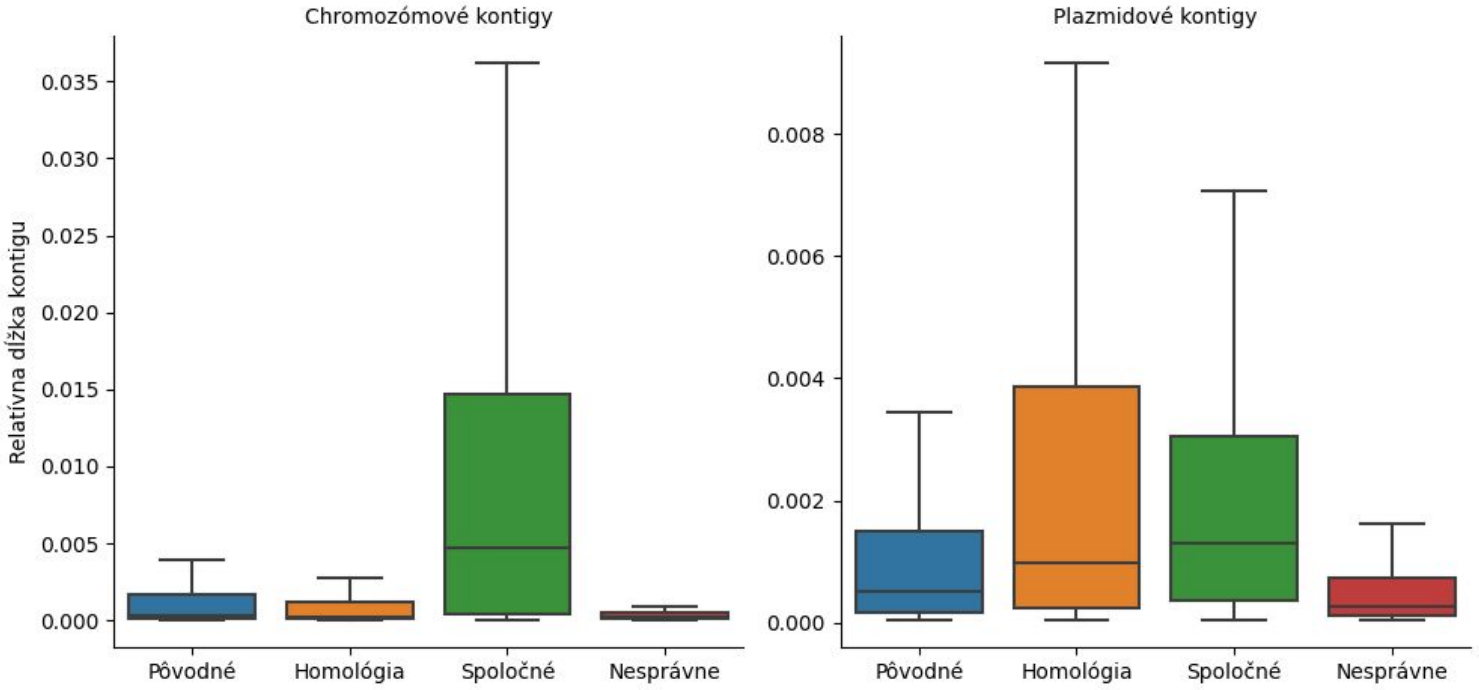
# Výsledky E. faecium

| Experiment | Mol. | AUC | Precision | Recall | F1 | Accuracy | # hmg. par. |
|---|---|---|---|---|---|---|---|
| orig | plazm. | 0,9256 | 0,7636 | 0,8333 | 0,7970 | 0,8542 | 0 |
| orig | chrom | 0,9385 | 0,9328 | 0,9439 | 0,9383 | 0,9023 | 0 |
| log-odd-score | plazm. | 0,9660 | 0,8700 | 0,8830 | 0,8764 | 0,9147 | 2 |
| log-odd-score | chrom | 0,9771 | 0,9563 | **0,9725** | **0,9643** | **0,9432** | 2 |
| log-odd | plazm. | 0,9692 | 0,8759 | 0,8803 | 0,8781 | 0,9161 | 333 |
| log-odd | chrom | 0,9759 | 0,9564 | 0,9648 | 0,9606 | 0,9377 | 1345 |
| fisher | plazm. | **0,9730** | **0,8799** | **0,8944** | **0,8871** | **0,9220** | 333 |
| fisher | chrom | **0,9780** | **0,9594** | 0,9656 | 0,9625 | 0,9408 | 1345 |

# Relatívne pokrytie kontigu v testovacích dátach *E. faecium*



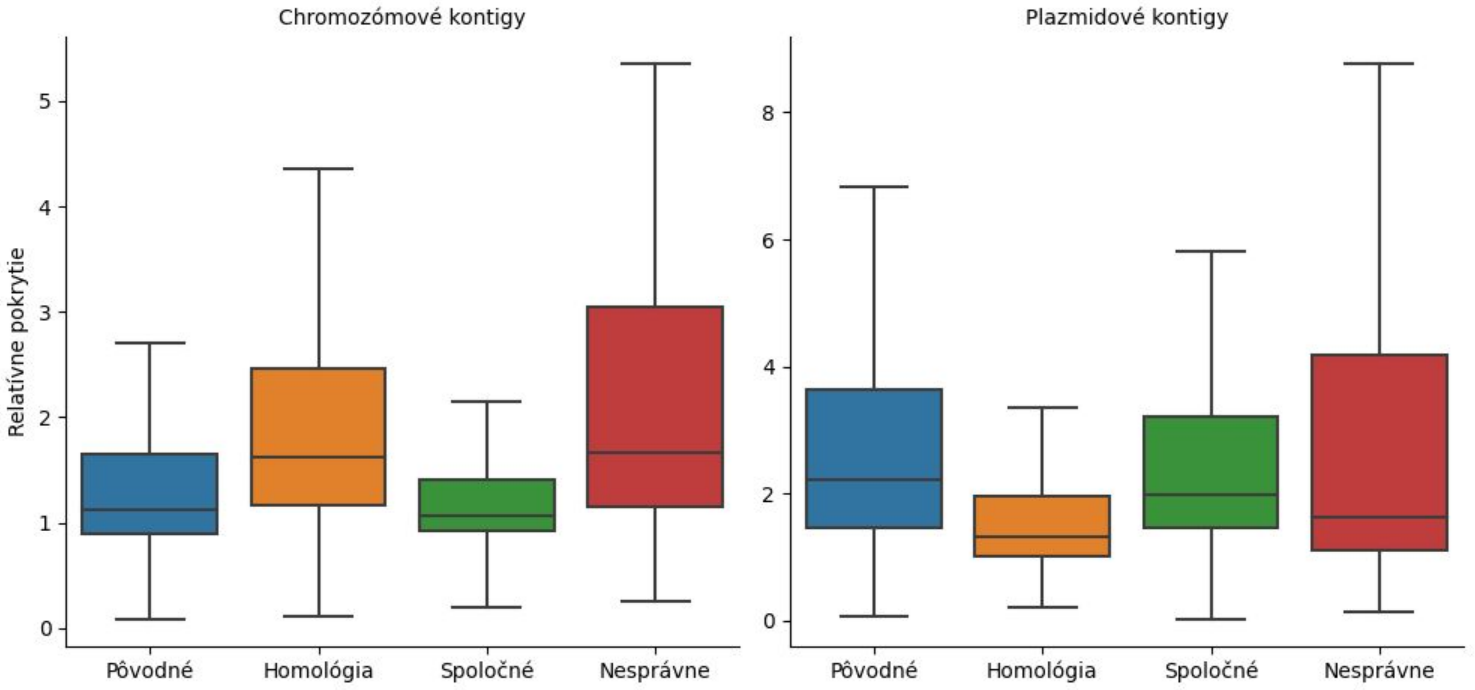
# Relatívna dĺžka kontigu v testovacích dátach *E. faecium*



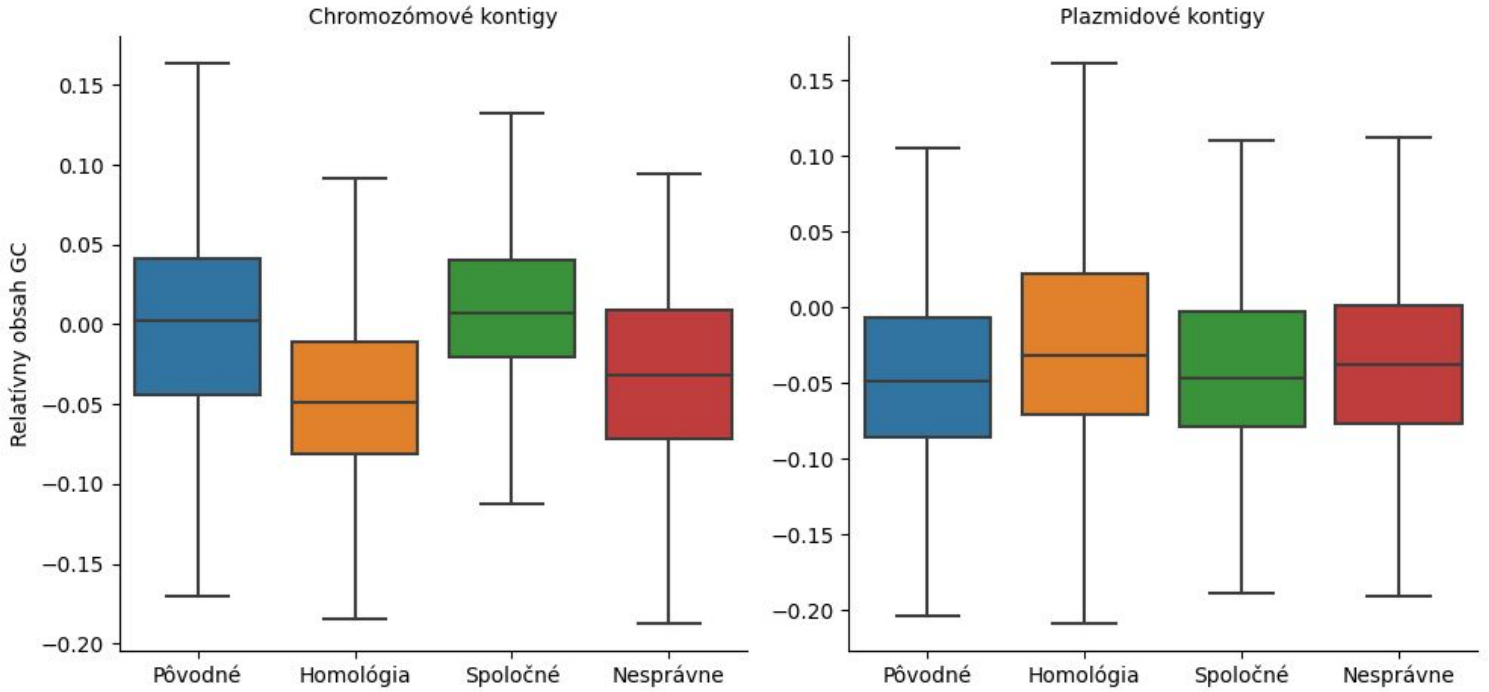
# Výsledky ESKAPEE

| Experiment | Mol. | AUC | Precision | Recall | F1 | Accuracy | # hmg. par. |
|---|---|---|---|---|---|---|---|
| orig | plazm. | 0,9256 | 0,6645 | 0,7171 | 0,6898 | 0,9096 | 0 |
| orig | chrom | 0,9359 | 0,9570 | 0,9620 | 0,9595 | 0,9272 | 0 |
| hmg | plazm. | **0,9656** | **0,7975** | **0,8308** | **0,8138** | **0,9467** | 730 |
| hmg | chrom | **0,9678** | **0,9680** | **0,9775** | **0,9727** | **0,9509** | 1564 |

# Relatívne pokrytie kontigu v testovacích dátach ESKAPEE



# Rel. obsah GC kontigov v testovacích dátach ESKAPEE



# Spojenie homológie a extremít

| Experiment | Mol. | AUC | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|---|
| orig | plazm. | 0,9256 | 0,6645 | 0,7171 | 0,6898 | 0,9096 |
| orig | chrom | 0,9359 | 0,9570 | 0,9620 | 0,9595 | 0,9272 |
| extr | plazm. | 0,9362 | 0,7036 | 0,7288 | 0,7160 | 0,9189 |
| extr | chrom | 0,9429 | 0,9620 | 0,9644 | 0,9632 | 0,9339 |
| hmg | plazm. | **0,9656** | **0,7975** | **0,8308** | **0,8138** | **0,9467** |
| hmg | chrom | **0,9678** | **0,9680** | 0,9775 | **0,9727** | **0,9509** |
| hmg-extr | plazm. | 0,9290 | 0,7969 | 0,7579 | 0,7769 | 0,9390 |
| hmg-extr | chrom | 0,9327 | 0,9640 | **0,9791** | 0,9715 | 0,9484 |

# Zhrnutie a návrh do budúcna

Čo sa nám podarilo

- príprava niekoľkých možností rozšírenia vstupných parametrov
- niekoľko verzií pre použitie extremít aj homológie

Návrhy do budúcna

- preskúmanie prečo spojenie oboch prístupov prinieslo zhoršenie
- použitie iných grafových neurónových sietí
- kumulatívne skóre z proteínových domén
- lokálna databáza pozostávajúca z iných sekvencií

1. Aké by boli výsledky, keby sme použili iba homológiu a nepoužili grafové spojenia? (Príp. aké vysledky majú predchádzajúce nástroje, ktoré používajú iba homológiu?)

| Mol. | AUC | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|
| plazm. | 0,9049 | 0,7535 | 0,5881 | 0,6606 | 0,9153 |
| chrom | 0,8947 | 0,9529 | 0,9779 | 0,9652 | 0,9368 |

| Method | SS | DB | AUROC | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|---|---|
| **A: Plasmid classification, contigs >100 bp, *n*=38,110** | | | | | | | |
| plASgraph2 | – | – | **0.991** | 0.906 | 0.908 | **0.808** | **0.935** |
| mlplasmids | X | – | 0.896 | 0.273 | **0.957** | 0.480 | 0.641 |
| PlasClass | – | – | 0.892 | 0.381 | 0.939 | 0.617 | 0.794 |
| PlasForest | – | X | n/a | 0.486 | 0.939 | 0.711 | 0.852 |
| Platon | – | X | n/a | **1** | 0.5 | 0.667 | 0.924 |
| Deeplasmid | – | X | n/a | n/a | n/a | n/a | n/a |
| RFPlasmid | X | X | 0.973 | 0.854 | 0.789 | 0.667 | 0.885 |
| **B: Chromosome classification, contigs >100 bp, *n*=38,110** | | | | | | | |
| plASgraph2 | – | – | **0.991** | 0.975 | **1** | 0.968 | 0.943 |
| mlplasmids | X | – | 0.908 | **1** | 0.540 | 0.697 | 0.609 |
| PlasClass | – | – | 0.878 | **1** | 0.738 | 0.840 | 0.766 |
| PlasForest | – | X | n/a | 0.992 | 0.771 | 0.855 | 0.795 |
| Platon | – | X | n/a | 0.957 | **1** | **0.973** | **0.952** |
| Deeplasmid | – | X | n/a | n/a | n/a | n/a | n/a |
| RFPlasmid | X | X | 0.959 | 0.982 | 0.936 | 0.933 | 0.893 |

2. Prečo spojenie extremít a homológie nefunguje? Na prvý pohľad vyzerá, že ide o problém v trénovaní, keď pridanie extremít do originálneho algoritmu trénovaniu chybu zníži, ale pri pridaní do homológie ju zvýši.

# Drobnosti

- V práci chýba akýkoľvek popis trénovania. T.j. koľko ako dlho sa model trénoval, akým optimalizátorom, aký bol learning rate, …
- Podľa kódu vyzerá, že learning rate sa počas trénovania neznižuje. Toto by som už v budúcnosti nerobil.
- V kóde sa používa “early stopping” (t.j. zastavím trénovanie keď validačná chyba ďalej neklesá a potom sa vrátim k najlepšej epoche doteraz). Toto je asi “najlepší” spôsob ako ľahko overfitnúť validačný dataset. Tiež by som to v budúcnosti už nerobil. Je viacero spôsobov ako zistiť vhodnú dĺžku trénovania. Osobne odporúčam https://arxiv.org/abs/1608.03983 alebo https://openaccess.thecvf.com/content/CVPR2022/html/Zhai_Scaling_Vision_Transformers_CVPR_2022_paper.
- Zvážiť písanie po anglicky, lebo pojmy ako “dlhá krátkodobá pamäť, precíznosť, návratnosť, presnosť” znejú hrozne.
- Obr. 2.4/5. by si zaslúžili farebne (alebo inak výrazne) odlíšiť dva druhy matíc susednosti, ktoré tam vstupujú.
- Tabulka 2.1 je neprehľadná a neviem z nej rýchlo vyčítať nič.
- Tabuľky s výsledkami 2.3, 3.2 je tiež neprehľadná, treba vyznačiť či menšia alebo väčšia hodnota je lepšia a tiež boldom vyznačiť najlepšie hodnoty.