

Grafové neurónové siete na identifikáciu plazmidov

Diplomová práca
Bc. Monika Buchalová
doc. Mgr. Tomáš Vinař, PhD.

Ciele na tento semester

- Implementácia attention mechanizmu
- Vytvorenie experimentov
- Evaluácia klasifikácie a jej vyhodnotenie

Plán práce

- Evaluácia pomocou scriptu z článku
- Porovnanie dvoch evaluačných scriptov, nájdenie chyby
- Úprava štruktúry grafu
- Vytvorenie a evaluácia experimentov pre vylepšenú verziu grafu
- Porovnanie výsledkov s existujúcimi nástrojmi

plASgraph2: using graph neural networks to detect plasmid contigs from an assembly graph

Cieľ: Detekovať **plazmidové kontigy** z assembly grafov bakteriálnych genómov, najmä v prípadoch, kde klasické sekvenčné metódy zlyhávajú.

plASgraph2 využíva **Graph Neural Network (GNN)** na klasifikáciu:

- Vstup: assembly graf + 6 číť pre každý kontig (napr. dĺžka, GC obsah, pokrytie, pentamer profil).
- Architektúra: 6 GCN vrstiev s identitným doplnením.
- Výstup: pravdepodobnosti pre *plasmid*, *chromozóm* alebo *ambivalentné* kontigy.

Tréning: Dáta z ESKAPEE skupiny patogénov.

- Hodnotené na nezávislých a neznámych dátach (non-ESKAPEE).
- Vyrovnaný výkon oproti (aj špecializovaným) SoTA nástrojom.

Prínos:

- Prvý GNN nástroj na túto úlohu.
- Nezávislý od databáz, vhodný pre nové alebo málo preskúmané druhy.
- Výrazne zlepšuje presnosť pre krátke alebo ťažko klasifikovateľné kontigy.

Graph Attention Networks

Problém: Tradičné GNN modely nedokážu efektívne rozlišovať dôležitosť susedných uzlov. Väčšina závisí na štruktúre grafu (napr. Laplaceova matica).

GAT – Graph Attention Networks využíva *self-attention* mechanizmus na susedné uzly:

- Dynamicky priradzuje váhy rôznym susedom.
- Nepotrebuje znalosť celej štruktúry vopred.
- Vhodné pre **transduktívne** aj **induktívne** úlohy.

Architektúra:

- Viacvrstvový model založený na **multi-head attention**.
- Efektívne paralelizovateľné.
- Vyhýba sa nákladným maticovým operáciám.

Accurate plasmid reconstruction from metagenomics data using assembly-alignment graphs and contrastive learning

Cieľ: Rekonštruovať plazmidy z metagenomických dát aj pri vysokej fragmentácii, nízkom pokrytí a mikrodiverzite.

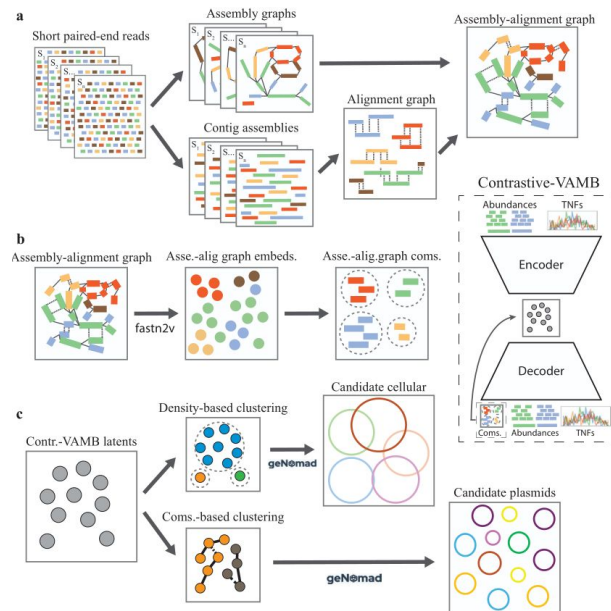
Hlavná myšlienka: PlasMAAG spája informácie z *assembly grafu* a *alignment grafu* do jedného **assembly-alignment graphu (AAG)** a binuje kontigy pomocou **contrastive learning**.

Vstupy a použité dáta

- Assembly graph (kontigy + ich spojenia)
- Alignment graph (kontigy podobné naprieč vzorkami)
- K-mer profily (tetranukleotidy)
- Abundance signál
- Tréning na syntetických aj reálnych dátach (hospital sewage)

Výsledky

- Presnejší aj pri krátkych, fragmentovaných a nízko pokrytých kontigoch
- V reálnych dátach (hospital sewage) **+33 % validovaných plazmidov**



Implementácia attention mechanizmu

- from spektral.layers import GATConv
- use_attention: true

```
# create layers and store as instance variables (needed to save the model)
if self["use_attention"]:
    print("Using attention layer")
    setattr(self, gcn_name, GATConv(
        channels=self['n_channels'],
        attn_heads=1,
        concat_heads=False,
        dropout_rate=self['attention_dropout'],
        activation=self['gcn_activation'],
        kernel_regularizer=reg,
        use_bias=True
    ))
else:
    setattr(self, gcn_name, gcn_conv.GCNConv(
        channels=self['n_channels'],
        activation=self['gcn_activation'],
        kernel_regularizer=reg,
        use_bias=True
    ))
```

Experimenty

- Spolu 5 experimentov
- Každý experiment má samostatný adresár s výsledkami

```

  ▾ experiments
    ▾ exp0_added_attention
      > model
        ≡ eskapee_test_predictions.csv
        ≡ exp0_added_attention.yaml
        ≡ metrics_per_sample.csv
        ≡ metrics_summary.csv
        ≡ train.err
        ≡ train.log
      > exp1_crossentropy_do0.2
      > exp2_squaredhinge_do0.2
      > exp3_crossentropy_do0.3
      > exp4_squaredhinge_do0.3
      > original_plasgraph
        ≡ all_metrics_summary.csv
```

Výsledok experimentov

	experiment ∇	task ∇	precision ∇	recall ∇	f1 ∇	accuracy ∇	auroc ∇
1	exp2_squaredhinge_do0.2	plasmid	0.755	0.725705329153605	0.666666666666667	0.90467862820804	0.950098887598888
2	exp2_squaredhinge_do0.2	chromosome	0.96097468113459	1.0	0.956888544891641	0.926713459950454	0.961230471402765
3	exp0_added_attention	plasmid	0.756565656565657	0.908033826638478	0.733003300330033	0.941176470588235	0.990606766421984
4	exp0_added_attention	chromosome	0.963978992894656	1.0	0.967204301075269	0.943303571428571	0.989251835985312
5	original_plasgraph	plasmid	0.833333333333333	0.769230769230769	0.717142857142857	0.941391155002147	0.987921577780371
6	original_plasgraph	chromosome	0.968975468975469	1.0	0.963889894965592	0.940786910790807	0.98287728605333
7	exp4_squaredhinge_do0.3	plasmid	0.755	0.725705329153605	0.666666666666667	0.90467862820804	0.950098887598888
8	exp4_squaredhinge_do0.3	chromosome	0.96097468113459	1.0	0.956888544891641	0.926713459950454	0.961230471402765
9	exp1_crossentropy_do0.2	plasmid	0.756565656565657	0.908033826638478	0.733003300330033	0.941176470588235	0.990606766421984
10	exp1_crossentropy_do0.2	chromosome	0.963978992894656	1.0	0.967204301075269	0.943303571428571	0.989251835985312
11	exp3_crossentropy_do0.3	plasmid	0.756565656565657	0.908033826638478	0.733003300330033	0.941176470588235	0.990606766421984
12	exp3_crossentropy_do0.3	chromosome	0.963978992894656	1.0	0.967204301075269	0.943303571428571	0.989251835985312

Typy výstupov

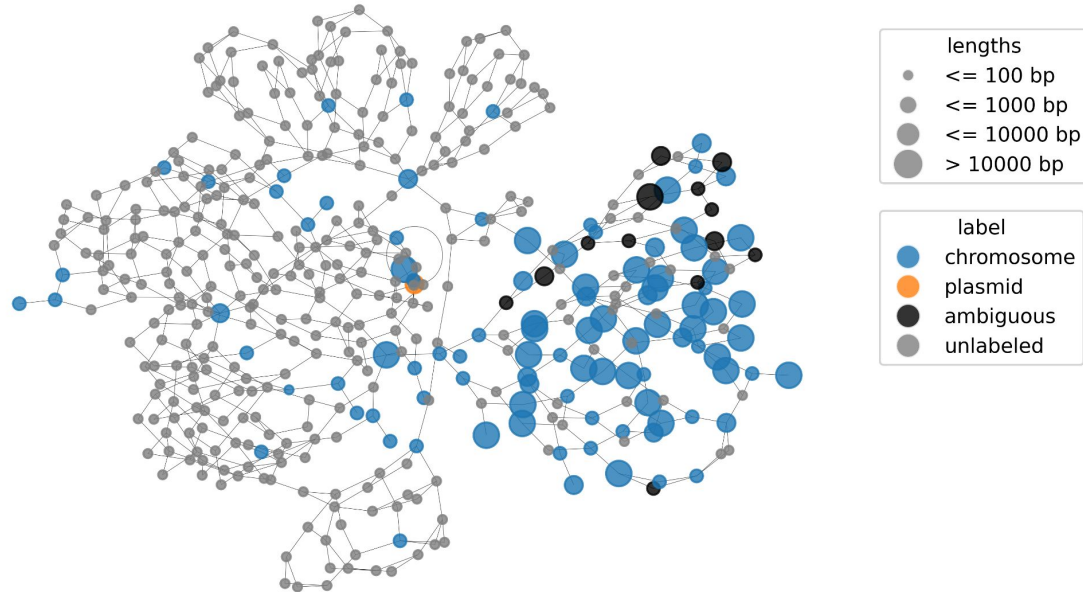
- Klasifikačná tabuľka
- Vizualizácia
 - predikovaných hodnôt
 - skór plazmidov
 - predikovaných hodnôt a skór plazmidov

Klasifikačná tabuľka

	sample ∇	÷	contig ∇	÷	length ∇	÷	plasmid_score ∇	÷	chrom_score ∇	÷	label ∇
1	robertson-benchmark_abau-SAMN10163228-s		Contig_10_1640.24		821		0.41805542		0.9521531		chromosome
2	robertson-benchmark_abau-SAMN10163228-s		Contig_11_767.706		51152		0.019140191		0.9914889		chromosome
3	robertson-benchmark_abau-SAMN10163228-s		Contig_12_918.955		4640		0.029671378		0.99260145		chromosome
4	robertson-benchmark_abau-SAMN10163228-s		Contig_13_1264.15		1391		0.034935888		0.9948362		chromosome
5	robertson-benchmark_abau-SAMN10163228-s		Contig_14_811.2		206383		0.0063471906		0.9974102		chromosome
6	robertson-benchmark_abau-SAMN10163228-s		Contig_15_2414.86		561		0.77814317		0.6525892		ambiguous
7	robertson-benchmark_abau-SAMN10163228-s		Contig_16_841.944		35336		0.016371338		0.99232996		chromosome
8	robertson-benchmark_abau-SAMN10163228-s		Contig_17_1076.09		3641		0.21813945		0.9061179		chromosome
9	robertson-benchmark_abau-SAMN10163228-s		Contig_18_4926.17		1309		0.24568398		0.9878564		chromosome
10	robertson-benchmark_abau-SAMN10163228-s		Contig_19_1699.71		666		0.6907502		0.7744854		ambiguous
11	robertson-benchmark_abau-SAMN10163228-s		Contig_1_5268.88_Circ		6078		0.9993891		0.006230306		plasmid
12	robertson-benchmark_abau-SAMN10163228-s		Contig_20_875.325		68758		0.014300094		0.99505913		chromosome
13	robertson-benchmark_abau-SAMN10163228-s		Contig_21_711.676		98081		0.009913929		0.9957453		chromosome
14	robertson-benchmark_abau-SAMN10163228-s		Contig_22_785.743		26572		0.015932921		0.9930146		chromosome
15	robertson-benchmark_abau-SAMN10163228-s		Contig_23_802.926		297386		0.007147357		0.9970932		chromosome
16	robertson-benchmark_abau-SAMN10163228-s		Contig_24_733.887		25562		0.04498018		0.9784079		chromosome
17	robertson-benchmark_abau-SAMN10163228-s		Contig_25_819.655		19826		0.02057231		0.9907411		chromosome

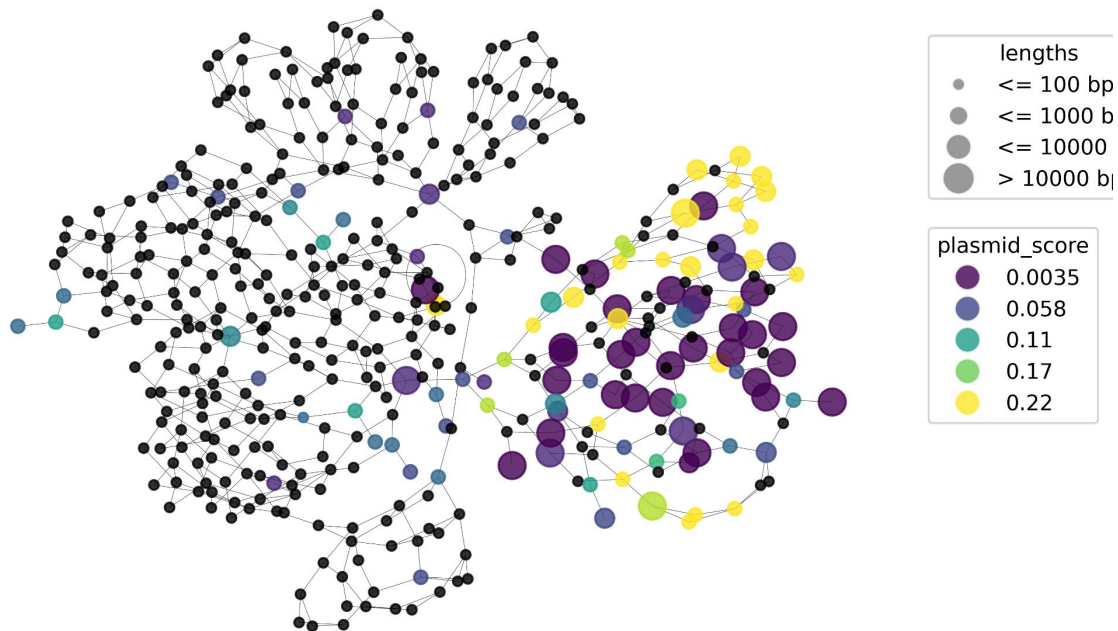
Vizualizácia predikovaných hodnôt

Sample robertson-benchmark_abau-SAMN10163228-s, label



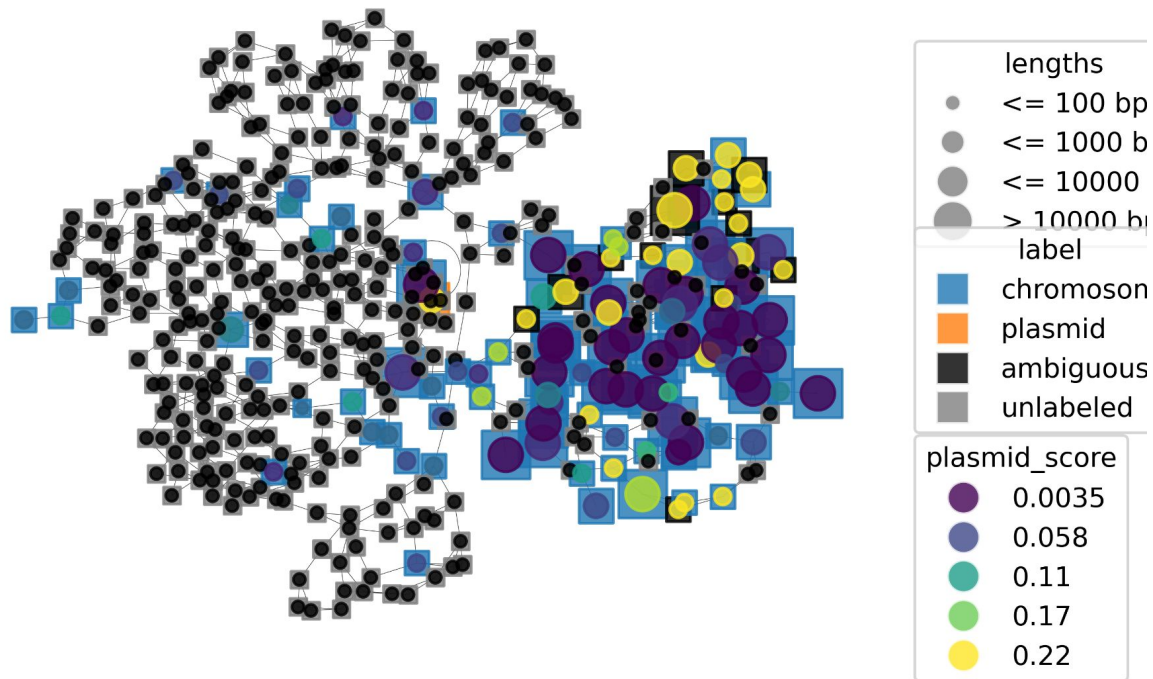
Vizualizácia skór plazmidov

Sample robertson-benchmark_abau-SAMN10163228-s, plasmid_score



Vizualizácia predikovaných hodnôt a skór plazmidov

Sample robertson-benchmark_abau-SAMN10163228-s, label:plasmid_score



Ďakujem za pozornosť