Metódy strojového učenia na rozpoznávanie plazmidov v zostavených genómoch baktérií

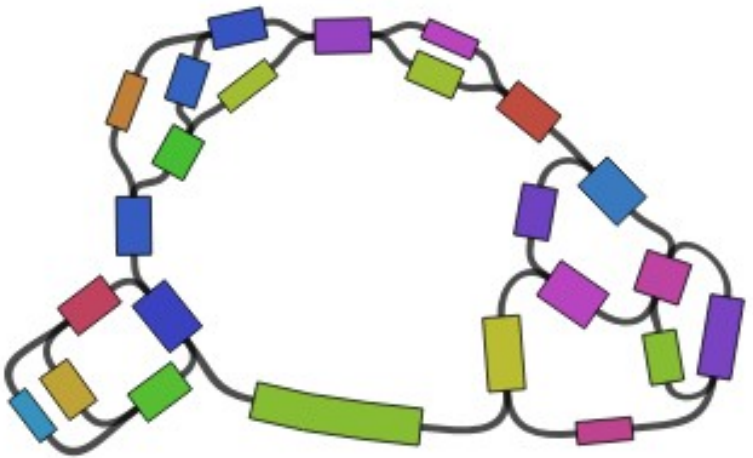
Juraj Vašut
Školiteľka: doc. Mgr. Bronislava Brejová, PhD.

Skladanie genómu pomocou krátkych čítaní

Výsledok skladania genómu

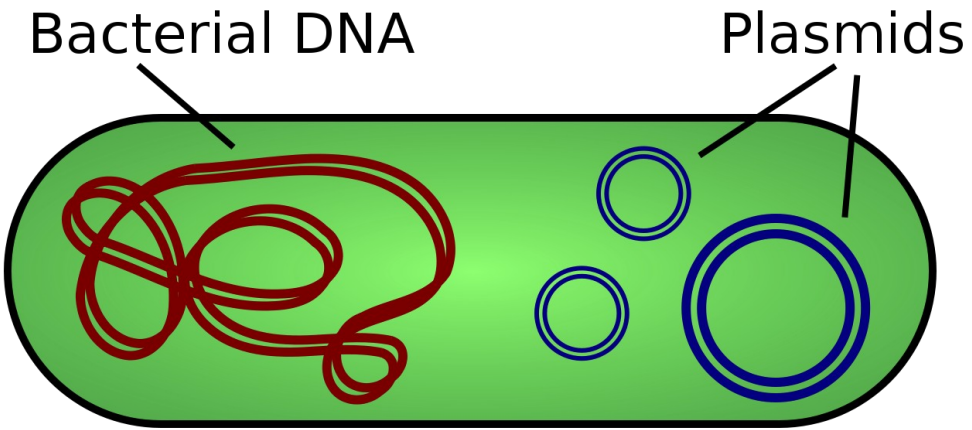
Graf

- Vrchol = kontig
- Hrana = prepojenie medzi kontigmi



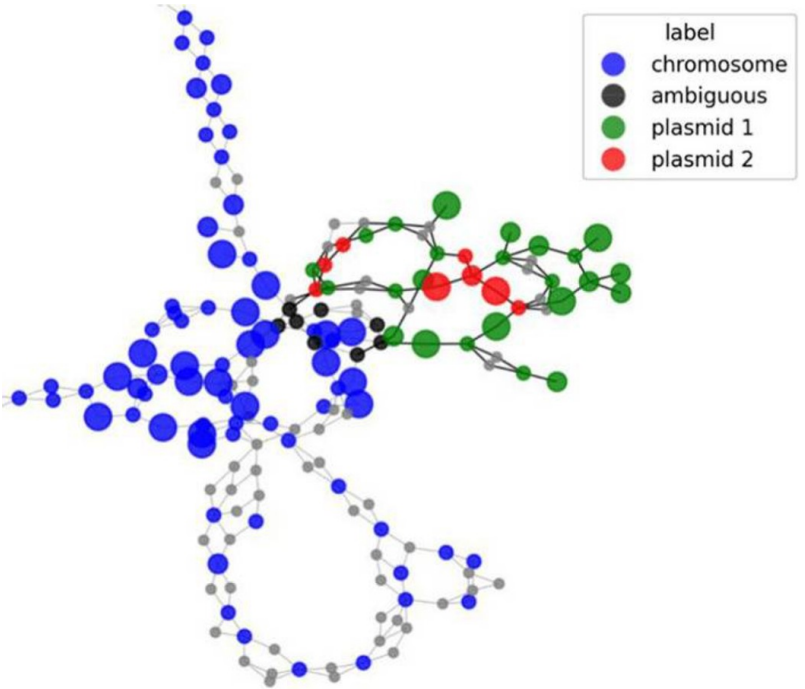
Plazmid

- Extrachromozomálna DNA molekula
- Obsahuje gény užitočné pre prežitie bunky (napr. rezistencie)
- Umožňuje horizontálny prenos génov

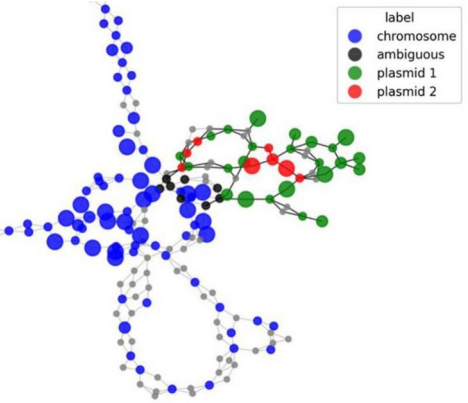


Plazmid binning

- Cieľom je detekcia skupín kontigov, ktoré pochádzajú z rovnakého plazmidu
- Druhy binningu
  - Riadené referenciou
  - de-novo



Metódy binningu plazmidov



- Recycler
  - Odstraňuje cykly z grafu
  - Predpokladá rovnomerné pokrytie sekvenovaných plazmidov
- PlasmidSPAdes
  - Odhadne pokrytie chromozómu a odstráni ho z grafu
  - Z komponentov grafu vytvorí určí biny plazmidov
- Gplas
  - Začne na kontigoch, ktoré sú plazmidové a vytvára z nich prechádzku v grafe
  - Rozširuje na základe podobnosti medzi pokrytím kontigu a priemerným pokrytím aktuálnej prechádzky
- HyAsP
  - Iteratívne odstraňuje prechádzky z grafu.
  - Rozširuje pomocou pokrytia, obsahu báz G a C a vysokej hustoty plazmidových génov v prechádzke

Použitá metóda

- Klasifikácia dvojíc kontigov podľa príslušnosti k molekule
- Zhlukovanie kontigov na základe získanej klasifikácie

Vstup

- Informácie o referenčnom genóme (hybridné zostavenie)
  - Klasifikácia kontigov
  - Úplnosť kontigov

- Trénovacie dáta (zostavenie krátkych čítaní)
  - Kontigy v grafe
  - Mapovanie kontigov na referenčný genóm

Klasifikácia vstupu

- True
  - existuje spoločný kontig v referenčnom genóme
- False
  - niektorý kontig je mapovaný len na úplné kontigy z referenčného genómu
- Unknown
  - ostatné prípady
  - odfiltrované

Použité znaky

- Znaky individuálnych kontigov
  - Dĺžka kontigu
  - Relatívne pokrytie kontigu čítaniami
  - Relatívny obsah báz G a C
  - Stupeň vrchola v grafe
  - Relatívny obsah k-merov

- Znaky párov
  - Vzdialenosť kontigov v grafe
  - Rozdiel:
    - Dĺžok kontigov
    - Relatívne pokrytie kontigu čítaniami
    - Relatívny obsah báz G a C
    - Stupeň vrchola v grafe
    - Relatívny obsah k-merov

Výsledky tréningu

Klasifikátor	Znaky	Veľkosť (kB)	Presnosť	True Positive	False Negative	False Positive	True Negative
Logistická regresia	**Individuálne**	1	0,63205	119726	20649	58412	16079
	Párové	1	0,63786	121015	19360	58452	16039
	Kombinácia	1	0,63416	120455	19920	58687	15804
Naivný Bayesov	**Individuálne**	1	0,66524	135432	4943	66985	7506
	Párové	1	0,66396	135322	5053	67150	7341
	Kombinácia	2	0,66548	**135584**	**4791**	67085	7406
Gradient Boosting	**Individuálne**	133	0,78039	111409	28966	18221	56270
	Párové	133	0,66626	92210	48165	23545	50946
	Kombinácia	133	0,78227	111741	28634	18149	56342
K najbližších susedov	**Individuálne**	150024	0,85289	119173	21202	10407	64084
	Párové	92323	0,64175	91257	49188	27858	46633
	Kombinácia	178777	0,84345	117917	22458	11180	63311
Náhodný les	**Individuálne**	1284314	**0,96234**	134242	6133	**1958**	**72533**
	Párové	4227336	0,69236	98355	42020	24082	50409
	Kombinácia	1263241	0,95441	133181	7194	2602	71889

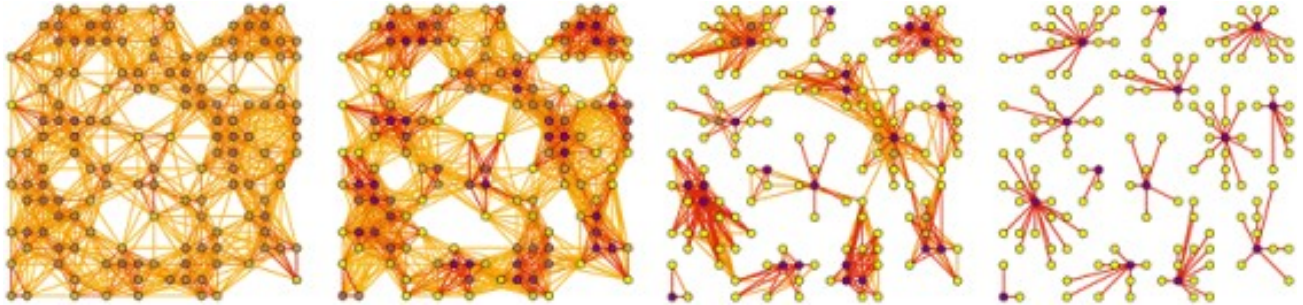
Vstup pre zhlukovanie

- Matica pravdepodobností M
  - $M_{i,j}$ je pravdepodobnosť, že kontigy i a j sú z rovnakej molekuly
  - Výsledok klasifikácie

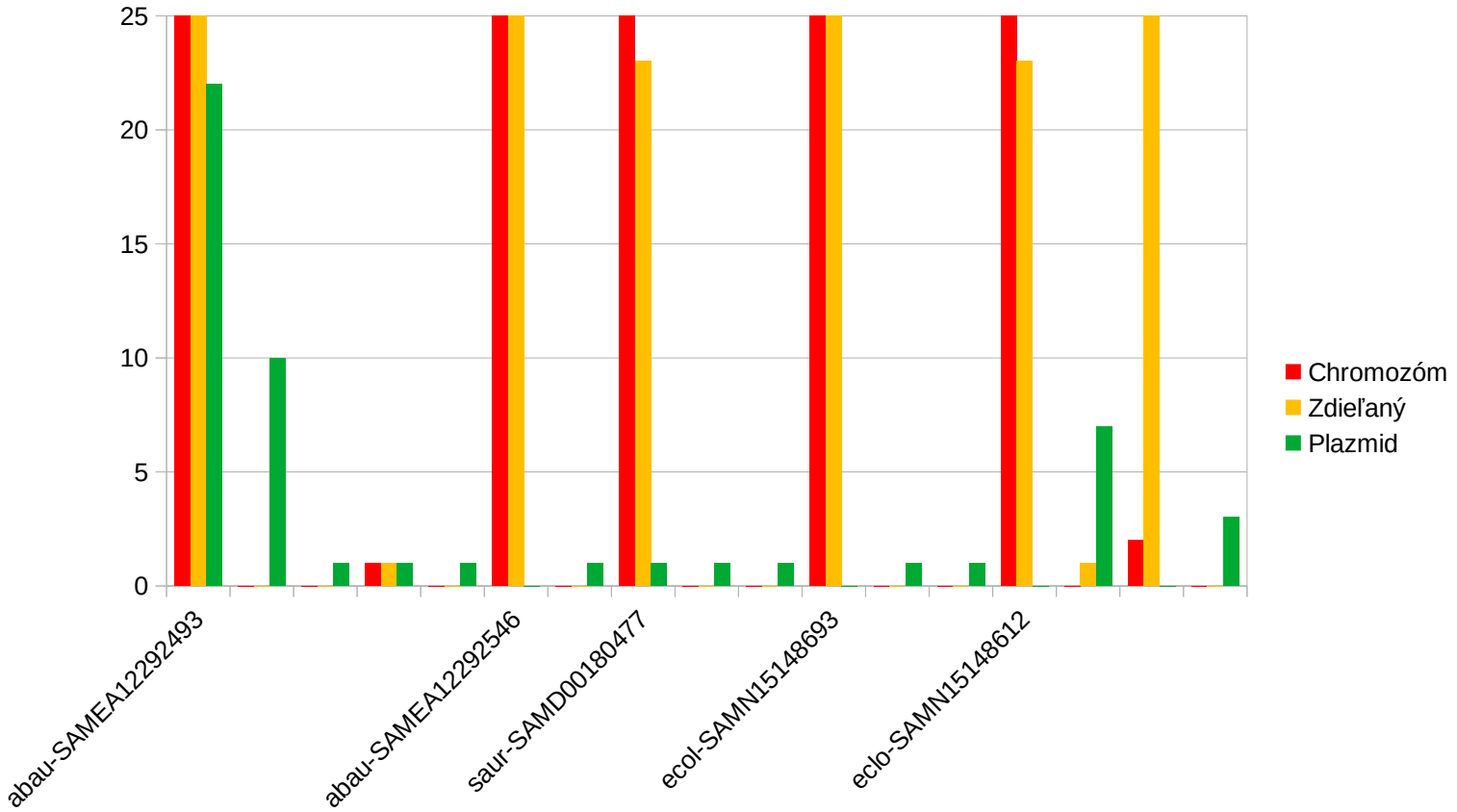
0	0,25	0,33	0,33	0	0	0
0,33	0	0,33	0,33	0,33	0	0
0,33	0,25	0	0,33	0	0	0
0,33	0,25	0,33	0	0	0	0
0	0,25	0	0	0	0,5	0,5
0	0	0	0	0,33	0	0,5
0	0	0	0	0,33	0,5	0

Zhlukovanie

- Cieľ: Rozdelenie kontigov do zhlukov
- Metóda Markovovského zhlukovania
  - využíva maticu pravdepodobností
  - Inflation
    - Umocnenie prvkov matice a normalizácia
  - Expansion
    - Umocnenie matice
  - Pomocou expansion a inflation odstraňuje slabé prepojenia



Výsledky zhlukovania



Záver

- Výsledky
  - Implementácia metódy na plasmid binning využívajúcej klasifikáciu a zhlukovanie
  - Vyhodnotenie vhodnosti rôznych klasifikačných metód na rozlišovanie príslušnosti dvojíc kontigov k molekulám
- Ďalší výskum
  - Použitie odlišných znakov na klasifikáciu
  - Spoľahlivejšie odstránenie chromozomálnych kontigov

Ďakujem za pozornosť.

1. Aké sú možnosti úpravy vášho riešenia pre prípad veľkých grafov, kde vzdialenosti medzi vrcholmi sú podobné?

- Veľké grafy sú často nežiadúce
- Možná úprava:
  - Namiesto vzdialenosti vrcholov v grafe použiť vzdialenosti v bp

2. Aká bola časová náročnosť výpočtu celého procesu riešenia pre vami zvolenú dátovú množinu?

- Tréning: 6 min 38.142973 s
- Klasifikácia: 1.899325 s
- Zhlukovanie: 1.281990 s

3. Poznáte softvér PlasmidHunter? Bol by použiteľný ako jedna možnosť na porovnanie Vášho riešenia?

- PlasmidHunter rieši problém chromozóm vs. plazmid

Slabinou je využitie predikcií programu plASgraph2 namiesto správnej klasifikácie kontigov pri zostavovaní trénovacích a testovacích dát. Nakoľko tieto predikcie môžu obsahovať veľa chýb, nie je šťastné sa nimi riadiť pri vývoji modelu.

- Predikcie pomohli lepšej klasifikácií párov kontigov, hlavne pri rozhodnutí o zdieľaných kontigoch

Z obsahovej stránky by som uvítala, keby sa výsledné zhluky porovnávali aj so správnymi zhlukmi z hybridných zostavení. Autor sa snaží odhadovať skutočný počet zhlukov len na základe výsledkov svojich metód, hoci má aspoň čiastočné informácie aj o správnom počte.

- Zarovnanie kontigov na hybridné kontigy obsahuje aj situácie, kde sa jeden kontig zarovná na viaceré hybridné
- Hybridné kontigy sú v niektorých vzorkách neúplné