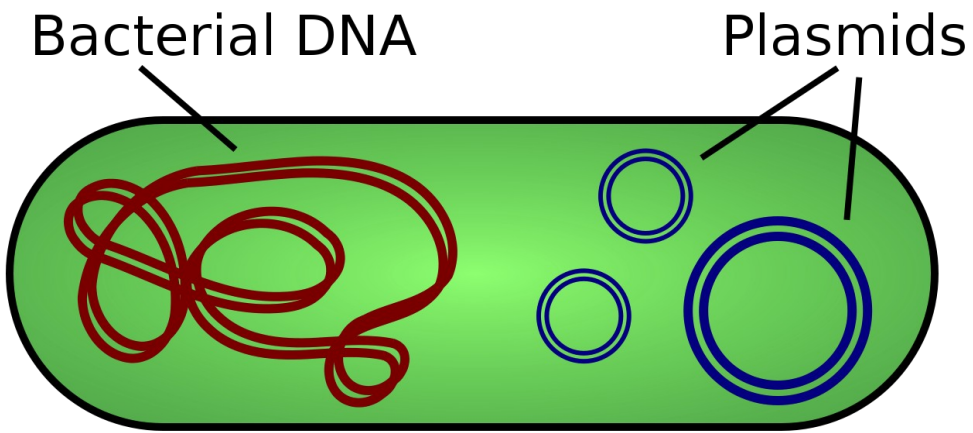
Plasmid binning

Juraj Vašut
doc. Mgr. Broňa Brejová, PhD.

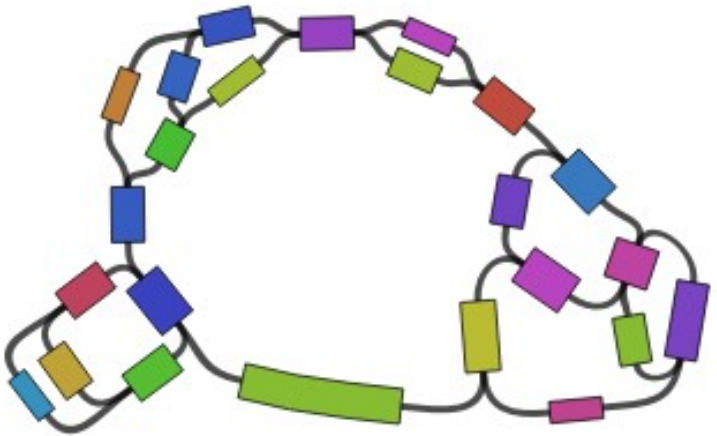
Plazmid

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



Výsledok skladania genómu

- Graf
  - Vrchol = kontig
  - Hrana = prepojenie medzi kontigmi



Binning

- Zoskupenie kontigov podľa príslušnosti k molekule DNA
- 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

Cieľ práce

- Pre dvojice kontigov pomocou informácie v grafe určiť, či patria do rovnakej molekuly DNA
- Rozdeliť kontigy pomocou určenej príslušnosti do skupín reprezentujúcich molekuly DNA

Vstup

- Informácie o referenčnom genóme
  - Klasifikácia kontigov
  - Úplnosť kontigov
- Trénovacie dáta
  - Kontigy v grafe
  - Mapovanie kontigov na referenčný genóm

Použité znaky

- Dĺžka kontigu
- Relatívne pokrytie kontigu čítaniami
- Relatívny obsah báz G a C
- Stupeň vrchola v grafe

Klasifikácia vstupu

- True
  - existuje spoločný kontig v referenčnom genóme
- False
  - niektorý kontig je mapovaný len na úplné kontigy z referenčného genómu
  - Množiny neúplných kontigov sú disjunktné
- Unknown
  - ostatné prípady

Trénovacia množina

- Veľkosť = 530444
  - True = 400882
  - False = 129562
- Vyvážená množina = 259124
  - True = 129562
  - False = 129562

Výsledky klasifikácie (pôvodný vstup)

	Logická regresia	
	0	1
0	9656	22668
1	1844	98444

	Random forest	
	0	1
0	31792	532
1	631	99657

	Naive Bayes	
	0	1
0	10549	21775
1	4515	95773

	K nearest neighbors	
	0	1
0	27087	5237
1	3565	96723

	Gradient Boosting	
	0	1
0	20061	12263
1	1174	99114

Výsledky klasifikácie (vyvážená trénovacia množina)

	Logická regresia	
	0	1
0	17701	14623
1	27120	73168

	Random forest	
	0	1
0	32144	180
1	1441	98847

	Naive Bayes	
	0	1
0	21369	10955
1	25374	74914

	K nearest neighbors	
	0	1
0	29746	2578
1	10917	89371

	Gradient Boosting	
	0	1
0	27658	4666
1	11621	88667

Budúca práca

- Pridanie znakov pre klasifikáciu
  - Relatívny obsah kmerov
  - Vzdialenosť kontigov v grafe
- Zhlukovanie pomocou informácií získaných z klasifikácie

Ďakujem za pozornosť