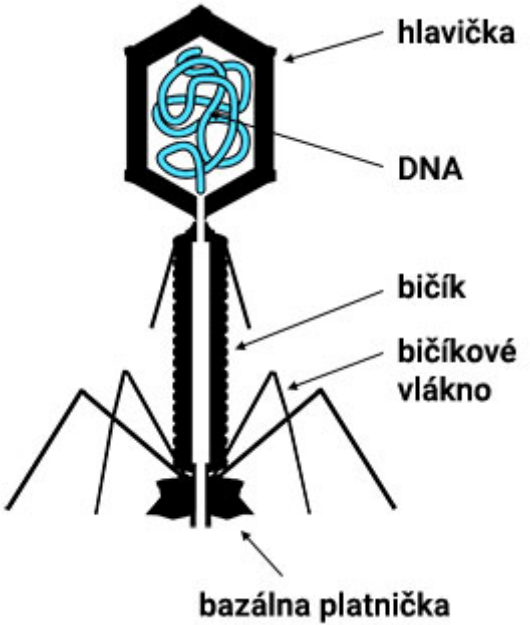
Program na detekciu endolyzínov z nespracovaných sekvenačných čítaní

Juraj Vašut
Školiteľ: Andrej Baláž, MSc.

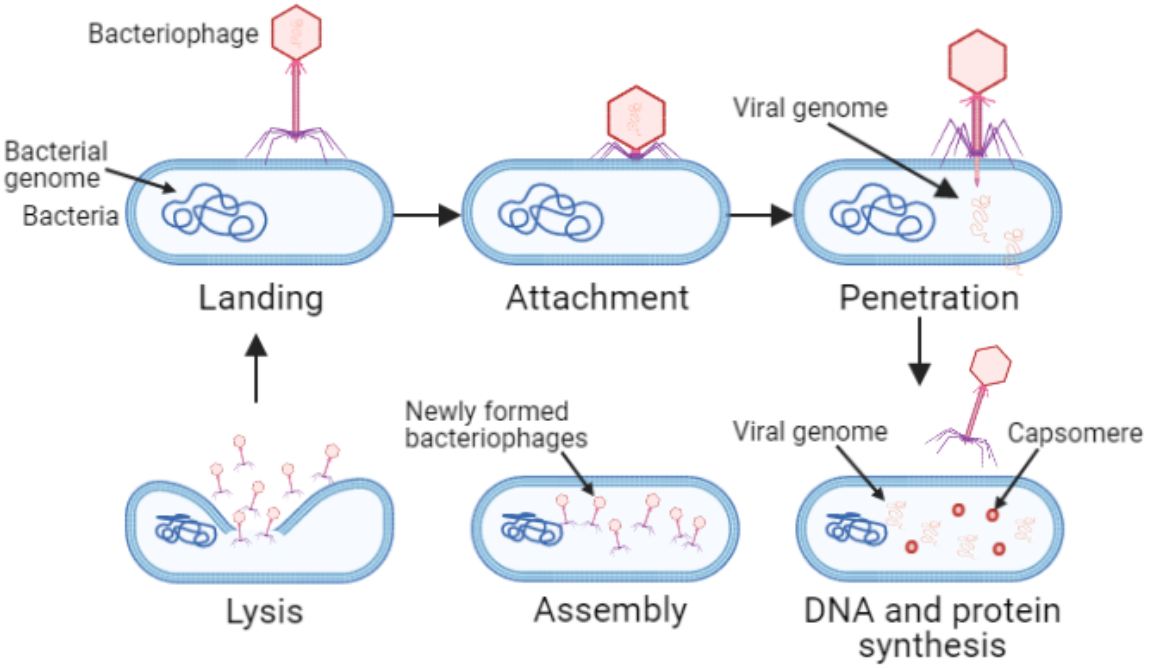
Baktériofág

- vírus
- infikuje konkrétne kmene baktérií
- 2 životné cykly: lytický a lyzogénny



Lytický cyklus

- Infekcia baktérie
- Replikácia baktériofágu
- Rozklad bunkovej membrány



Rozklad bunkovej membrány

- Enzým endolyzín
- Spôsobuje smrť bunky

Súčasný stav

Nástroje
- Anotácia genómu
- Súkromné nástroje

RASTtk
- Bakteriálny genóm

PhATE
- Virálny genóm

Naše riešenie

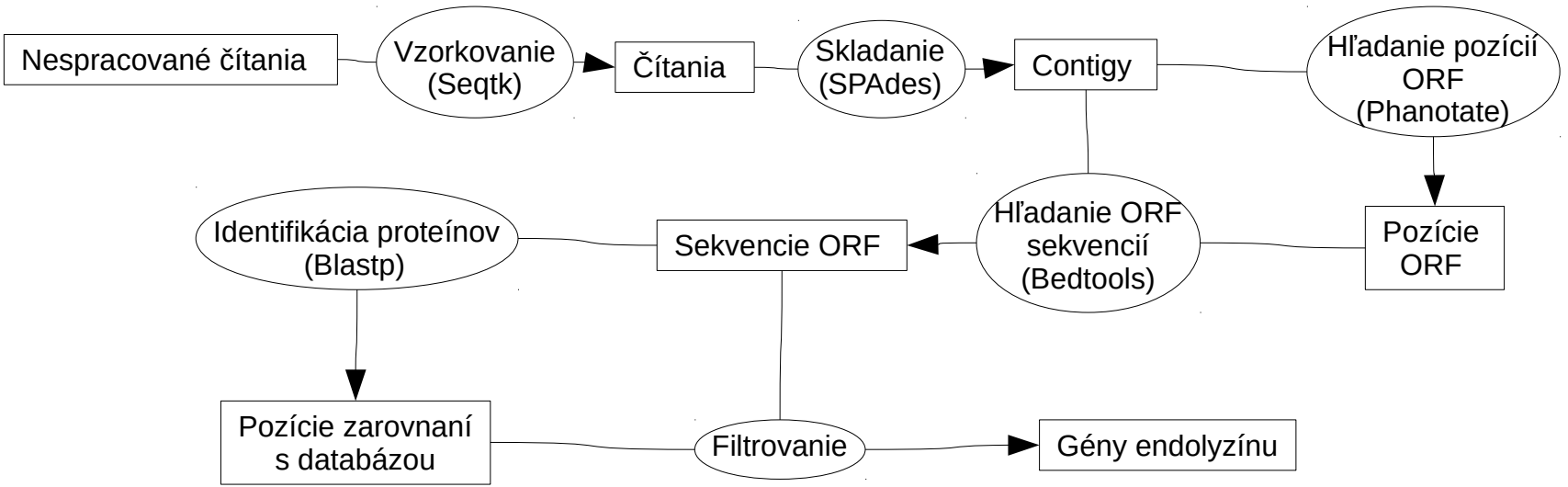
Inštalácia:

```
- conda install phendol
```

Spustenie:

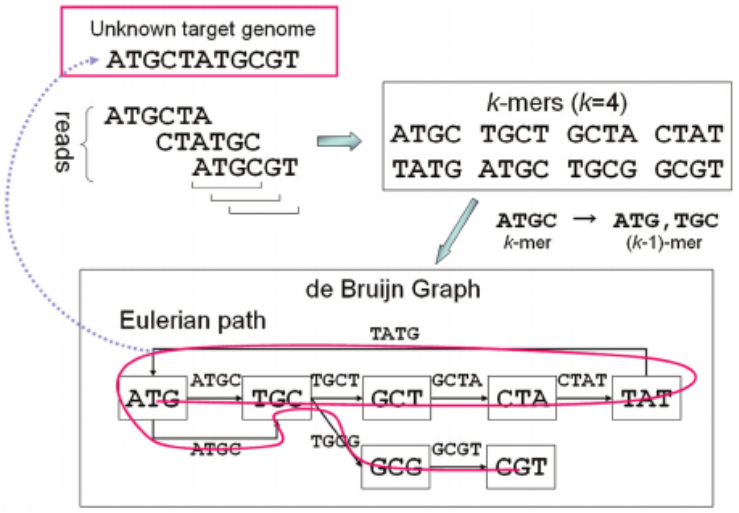
```
- phendol [options] -r1 <reads1> -r2 <reads2>
```

Chod programu



SPAdes

- zostavenie krátkych čítaní bakteriálneho genómu
- využíva k-mery na zostrojenie de Bruijnovho grafu



Phanotate

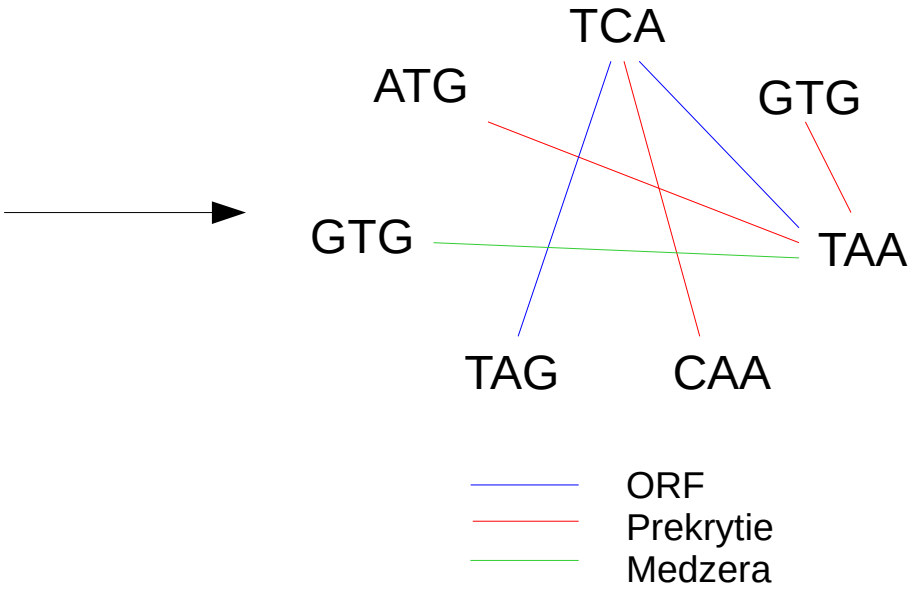
- Hľadanie otvorených čítacích rámcov (ORF) v genóme baktériofágu
- Genóm vníma ako vážený graf štart a stop kodónov
- Hľadá najkratšiu cestu pomocou Bellman-Ford algoritmu

Čítací rámec 1: ATGTCAGTGTAACAATAGTG

Čítací rámec 2: ATGTCAGTGTAACAATAGTG

Čítací rámec 3: ATGTCAGTGTAACAATAGTG

Zelená = štart kodón
Modrá = stop kodón
Svetlá = rovnaké vlákno
Tmavá = komplementárne vlákno



Blastp

- hľadanie podobností medzi sekvenciami
- rýchlejší ako Smith-Waterman algoritmus
- vytvára zo sekvencie slová rovnakej dĺžky
- zo slov vytvorí vyhľadávací strom v ktorom potom hľadá zhody z databázou

Filtrovanie

Parametre:

- minimálna dĺžka endolyzínu
- identita (I)
- pokrytie (C)

Identita (I)

Počet zhôd: 43
Počet stĺpcov: 50
Identita: 43/50 = 86%

Pokrytie (C)

EFDYMIRIMFEQGRKSLHCGEWYSHICDHRMRHTCRMDAVPVTWVMQKNA Z databázy

Počet AK v zarovnaní: 7
Počet AK sekvencie v databáze: 50
Pokrytie: 7/50 = 14%

Príprava testovacích vstupov

Nástroj: InSilicoSeq
Vstup: náhodné kompletné genómy baktériofágov z databázy vírusov dostupnej na NCBI
Výstup: [počet genómov][sekvenátor][počet čítaní]_R1(2).fastq.gz
Počet genómov: 10, 50
Sekvenátor: MiSeq, HiSeq, NovaSeq
Počet čítaní: 1 milión, 5 miliónov

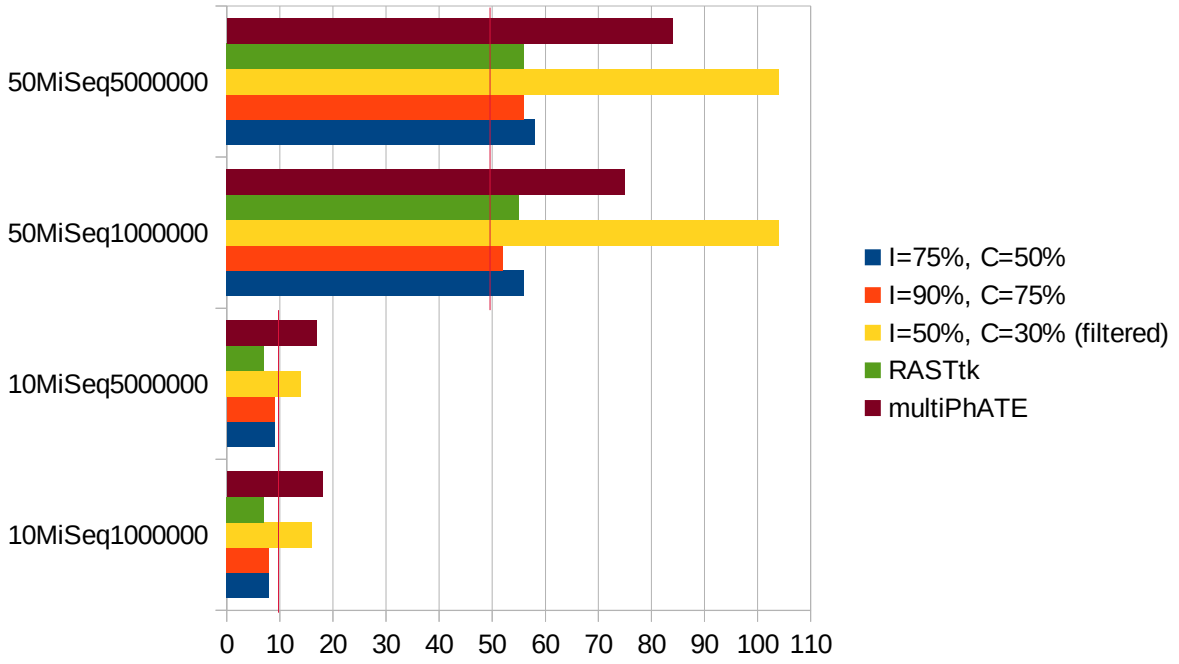
Testovacie nastavenia

Vzorkovanie: 2 milióny
Minimálna dĺžka endolyzínu: 50 aminokyselín
Identita: 50%, 75%, 90%
Pokrytie: 30%, 50%, 75%

Výsledky

	Phendol					
	I=75%, C=50%	I=90%, C=75%	I=50%, C=30%	I=50%, C=30% (filtered)	RASTtk	multiPhATE
10MiSeq1000000	8	8	64	16	7	18
10MiSeq5000000	9	9	58	14	7	17
10HiSeq1000000	9	9	65	13	5	--
10HiSeq5000000	12	12	110	20	9	--
10NovaSeq1000000	16	16	162	34	5	--
10NovaSeq5000000	13	12	73	20	12	--
50MiSeq1000000	56	52	441	104	55	75
50MiSeq5000000	58	56	510	104	56	84
50HiSeq1000000	50	49	413	104	--	--
50HiSeq5000000	55	53	442	98	50	--
50NovaSeq1000000	53	50	441	101	73	--
50NovaSeq5000000	53	45	463	105	45	--

Výsledky


50MiSeq5000000
50MiSeq1000000
10MiSeq5000000
10MiSeq1000000
0
10
20
30
40
50
60
70
80
90
100
110
I=75%, C=50%
I=90%, C=75%
I=50%, C=30% (filtered)
RASTtk
multiPhATE

Ďakujem za pozornosť.

Vyhodnotenie výsledkov

- Každý baktériofág má 1 endolyzín
- Nehľadá na kontigoch, ale na ORF

Identity

- Identitu možné znížiť na 50%
- Predpoklad, že aj keď genóm je výrazne odlišný, štruktúra jednotlivých podjednotiek endolyzínu je podobná

Falošné pozitíva

- Prijateľné
- Testovateľné

Problémy

- Zachovanie informácie o vlákne obsahujúcom ORF
- InterProScan nahradený Blastp
- Namiesto filtrovania contigov filtrovanie výsledných endolyzínov
- Spúšťanie porovnávaných nástrojov

Organizácia textu

Snaha o minimalizáciu popisov so zachovaním čitateľnosti