Klasifikácia bakteriofágov na základe ich genomickej sekvencie

Andrej Baláž

Školiteľ: Jaroslav Budiš

Comenius University

21.06.2018

Čo sú to bakteriofágy

- vírusy napádajúce baktérie
- pomerne jednoduché organizmy
- najpočetnejšie organizmy na Zemi

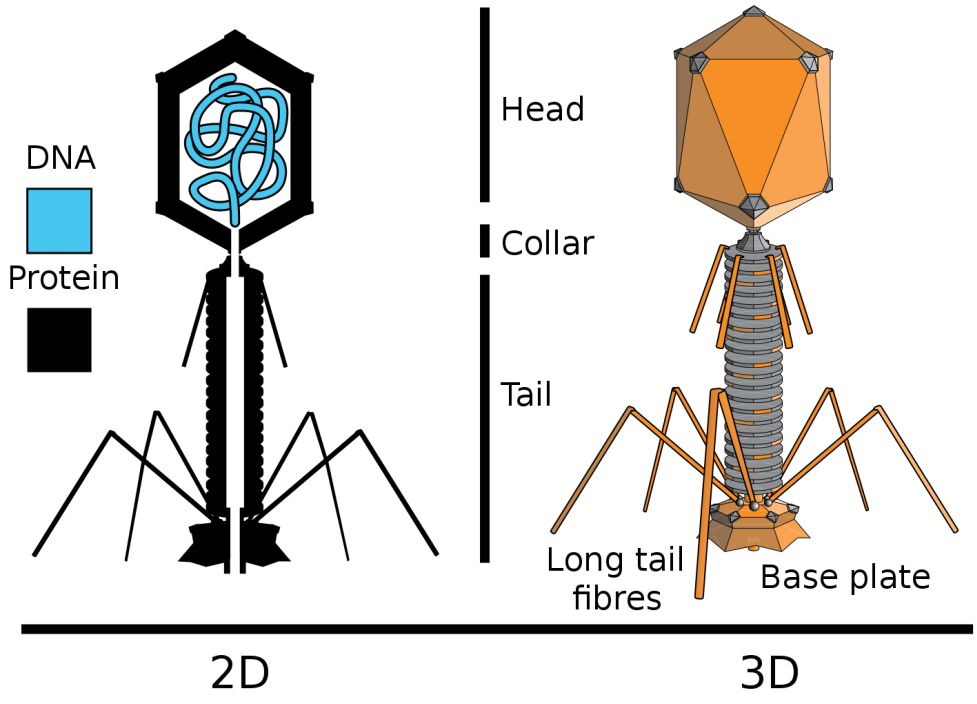


Figure 1: Štruktúra typického bakteriofága

Životný cyklus bakteriofágov

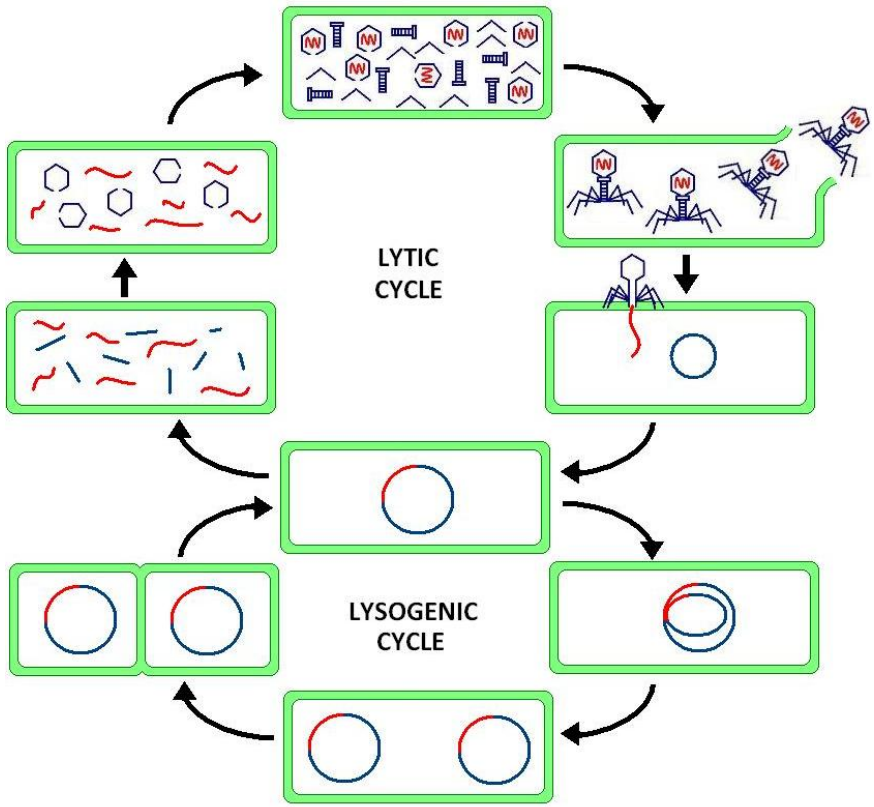


Figure 2: Životný cyklus bakteriofágov

Fágová terapia

- používa sa v Gruzínsku a Rusku
- oproti antibiotikám
  - vysoká špecificita
  - nízke riziko vedľajších účinkov
  - účinná aj proti baktériám v biofilmoch
  - možnosť použiť na multirezistentné kmene
- proces výroby nezaručuje presné zloženie liečiv ani neobjasňuje presný mechanizmus účinku

Ciele práce

- zozbierať dostupné informácie
- vytvoriť model na predikciu hostiteľa
- interpretovať model

Zbieranie dát

- FASTA formát

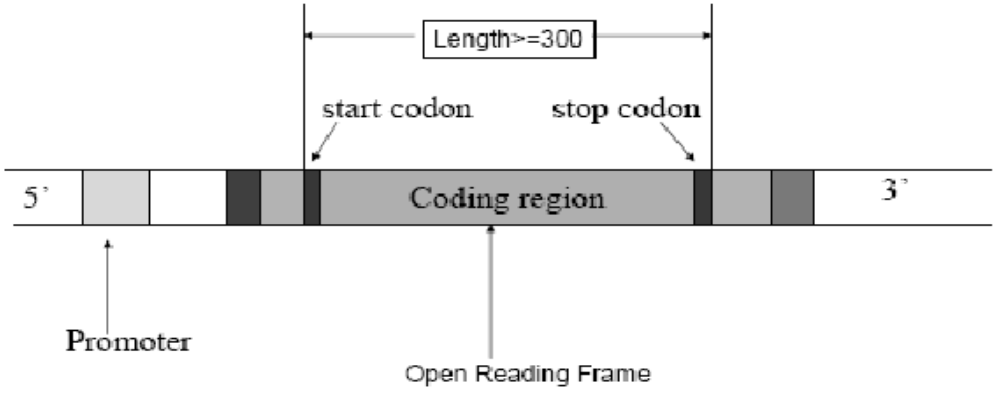
```
>NC_024124.2 Escherichia phage vB_EcoM_JS09, complete genome
ATGATTATTGAAACGGCTAAAGAAACGATTATTGGTTCAGGCGGGAAGAGCACAGCATTTACTATTCAAG
GCAATAGCAAAGTTTATAAGATTTTGTCTAATGACCTTTATACAAACAAAGAATTGGCTTGTGTACGTGA
ATTAATCACAAACTGTATCGATGGACAAATTCTTAATGGTTGCACTGATAAGTTTATTGTTCAGGCTCCA
GGTCGTTTAGACCCACGTTTTGTAGTTCGAGACTTTGGTCCTGGTATGAGTGATTTCACCATTCGAGGTA
ATGACGAAGAGCCAGGAATCTATAACTCATATTTTGCTTCAACTAAAACATCAAGTAACGATTTCATTGG
TGGATTCGGACTCGGCTCTAAAGCTCCGTTAGCATATACTGATACGTTTAACCTTACGTCTTATCACAAT
GGTGAAGTTCGTGGTTATGTAATTTATCAAGATGACAGTGGTCCACAGATTAAGCCAACCTTCGTAGATA
AGATGGGTCCTGATGACCGGACTGGCGTTGAAGTAGTTGTGCCTGTTAACCCAGAAGATTTTGAAAAGTT
CGCATCCGAAATTGCTTATGTTATGCGTCCTCTCGGCGATATTGCAGAAGTTCGTGGTGTTAAAGATATC
AAATACTTCCCGGAATTCGATGATGTTTATTTGGCCAAGGAAGCTCCATGGGGTGAACGTGGTAATATCA
```

- hostiteľ

Zbieranie dát

- sťahovanie z verejných databáz:
  - GenBank - 6704 sekvencií
  - RefSeq - 2107 sekvencií
  - PhagesDB - 2491 sekvencií
- odstránenie duplikátov
  - 6277 finálnych sekvencií

Vyhľadávanie génov



- štart kodón = ATG, stop kodón = TAG, TAA, TGA
- Prokka
  - koordináty
  - preklad do proteínových sekvencií
  - priradenie funkcie
- nájdených 590246 génov

Delenie datasetu

- trénovací dataset (2787 záznamov)
- testovací dataset (699 záznamov)
- vyradené

rod	počet	rod	počet
Mycobacterium	1619	Streptococcus	354
Escherichia	323	Gordonia	293
Arthrobacter	240	Pseudomonas	236
Lactococcus	219	Staphylococcus	184

Table 1: Počty záznamov jednotlivých rodov

Zarovnávanie proteínov

- určenie podobnosti medzi proteínmi navzájom

  A C T C G C A A T A T G C T A G G C C A G C

  A C T _ _ _ _ T T A T G C T A T G C _ _ G C

  ▶
- získanie skóre zarovnaní
- vytvorenie matice podobnosti proteínov

Vyhľadávanie klastrov

- Markovov Klastrovací Algoritmus
  - deterministicky simuluje náhodné prechádzky po grafe
  - používa dve operácie - inflácia a expanzia
- 15017 klastrov
- anotácia klastrov

Tvorenie binárnej matice

- binárna matica *bakteriofag* $\times$ *klaster*
  - $a_{i,j} = 1$, ak bakteriofág i obsahuje gén z klastru j
  - $a_{i,j} = 0$, inak
- surová matica = 2787×15017
- redukcia matice pomocou Variance Threshold metódy (99%)
- redukovaná matica = 2787×1818

Analýza hlavných komponentov

- jedna z najbežnejších metód na vizualizáciu mnohorozmerných dát
- pre maticu $n \times m$ vytvorí $\leq n$ hlavných komponentov
- hlavné komponenty sú nekorelované a zachytávajú čo najviac variancie

Analýza hlavných komponentov

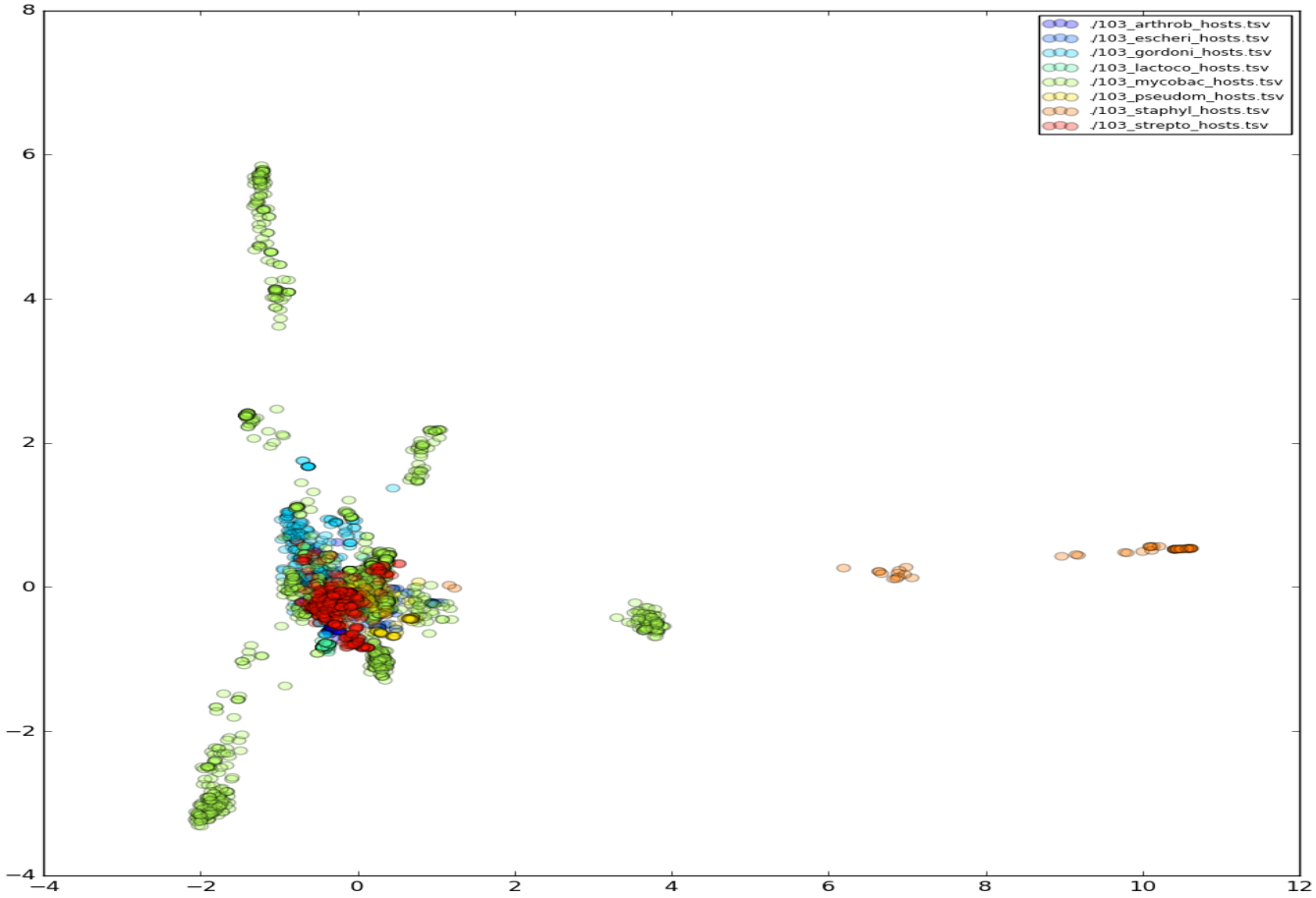


Figure 3: Principal component 0 (11,57%) and principal component 1 (9,19%)

Stromový klasifikátor

- pythonová knižnica scikit learn - decision tree
- 8 rozhodovacích stromov (pre každý rod jeden)
- každý strom nám odpovedá na otázku či daný fág napáda baktériu z konkrétneho rodu
- jednotlivé stromy je možné interpretovať

Stromový klasifikátor

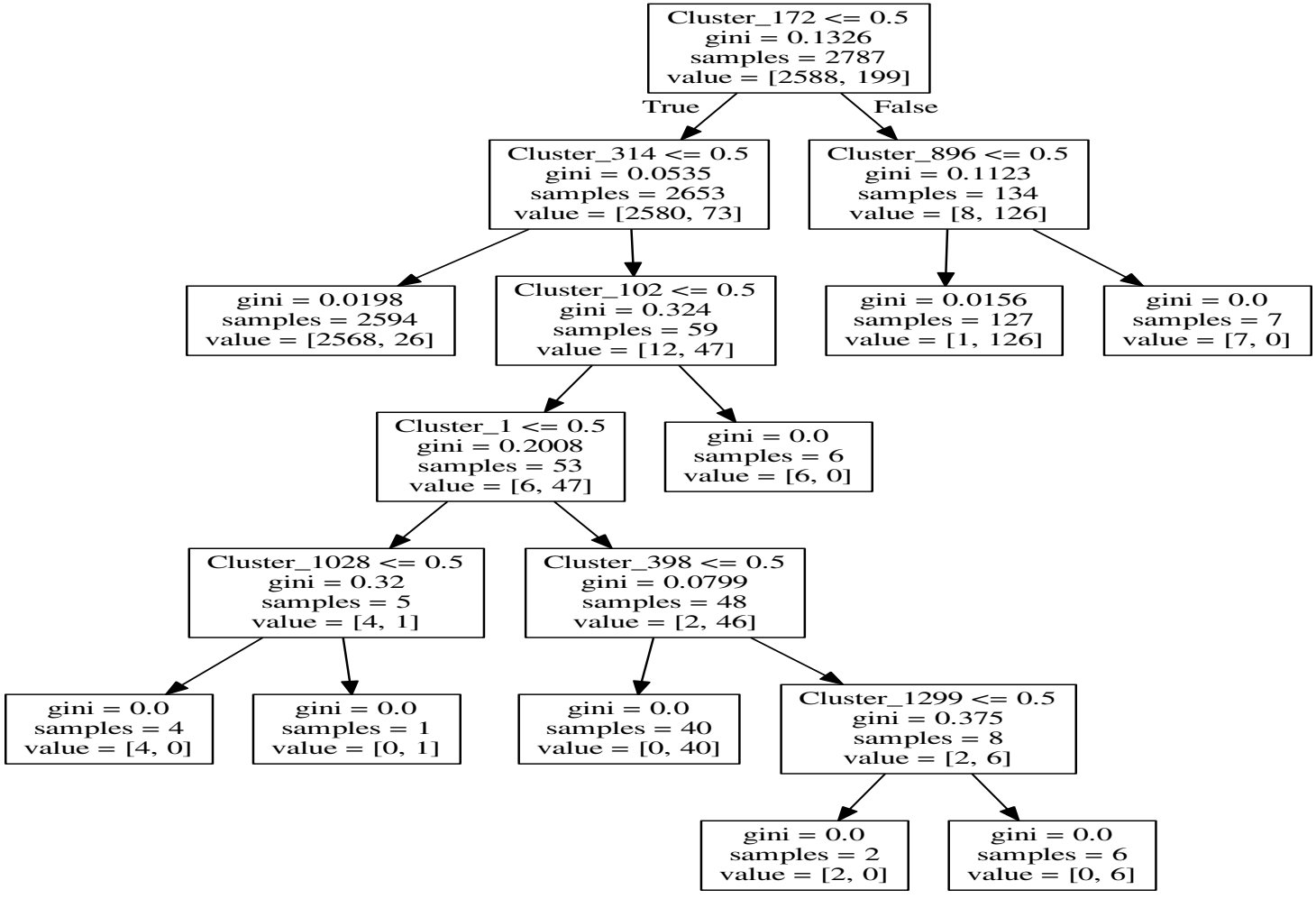


Figure 4: Arthrobacter model

Evaluácia modelov

model	correct	fpos	fneg	accuracy	fposp	fnegp
arthrob	683	11	5	97.71%	1.57%	0.71%
escheri	679	17	3	97.13%	2.43%	0.42%
gordoni	647	39	13	92.56%	5.57%	1.85%
lactoco	680	3	16	97.28%	0.42%	2.28%
mycobac	686	11	2	98.14%	1.57%	0.28%
pseudom	686	6	7	98.14%	0.85%	1.00%
staphyl	685	0	14	97.99%	0.00%	2.00%
strepto	686	2	11	98.14%	0.28%	1.57%

Následná práca

- Získanie ďalších dát (množstvo/kvalita)
- Rozšírenie predikcie na druhy/kmene
- Predikovanie z neúplných sekvencií

Ďakujem za pozornosť