Odhady taxonomické struktury bakterií

Josef Moudřík

Jan Vyhnánek

Cíl práce

- jak jednoduše zjistit hierarchickou příbuznost organismů?
  - Vzdálenost mezi genomy organismů odhadneme pomocí kompresních algoritmů a aproximace kolmogorovy komplexity (viz. přednáška)
    - Algoritmy GenCompress,Zlib7 a gzip2
  - Jiné metody pro zjištění vzdáleností – následující slajd
  - Ze vzdáleností mezi organismy následně vytvořit hierarchický strom
    - Algoritmy Neighbour joining a Upgma

- **Jak jsou jednotlivé metody spolehlivé?**

Vzdálenost genomů pomocí histogramu

- Kolikrát se sekvence délky *w* opakuje v genomu?
- Příklad histogramu sekvencí pro w=4

Organismus A	
Sekvence	Výskytů
AAAA	3
AAAG	5
AAGA	1
.	.
.	.
TTTT	2

Organismus B	
Sekvence	Výskytů
AAAA	2
AAAG	3
AAGA	6
.	.
.	.
TTTT	1

- Vzdálenost mezi organismy definujeme jako euklidovskou vzdálenost mezi normalizovanými vektory výskytů sekvencí

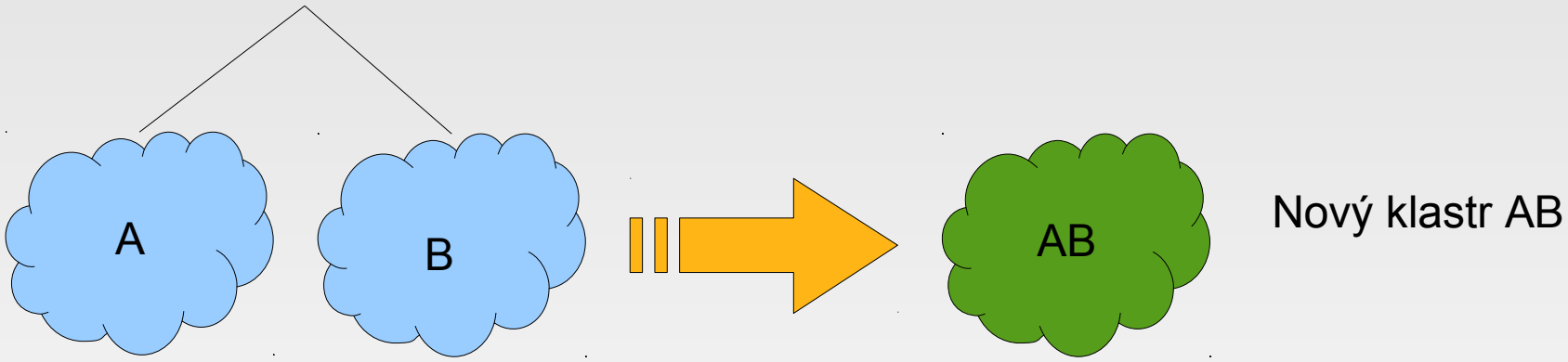
Vytváření taxonomického stromu

- Jak získat z distanční matice hierarchický strom?

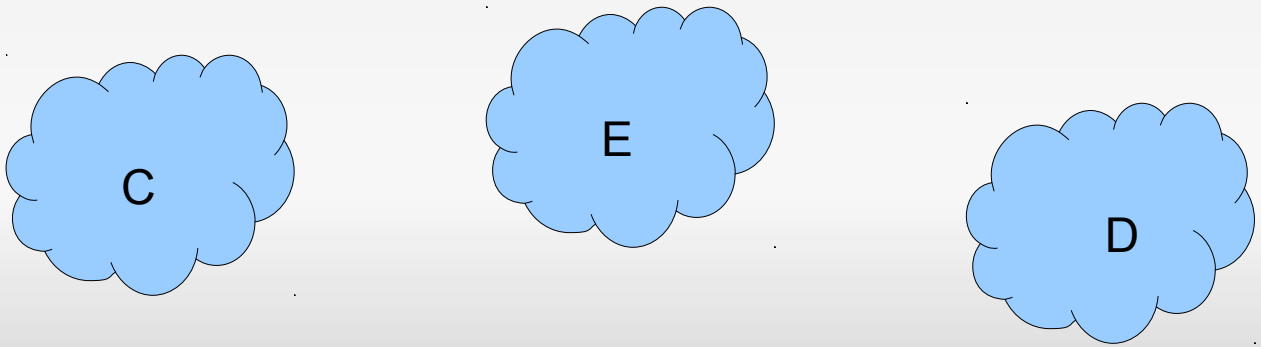
	A	B	C	D
A	0	0.67	0.93	0.87
B	0.67	0	0.78	0.91
C	0.93	0.78	0	0.82
D	0.87	0.91	0.82	0

Vytváření stromu - UPGMA

- V každém kroku jsou propojeny dva nejbližší klastry
- Vzdálenost mezi dvěma klastry je definována jako průměr vzájemných vzdáleností jejich prvků

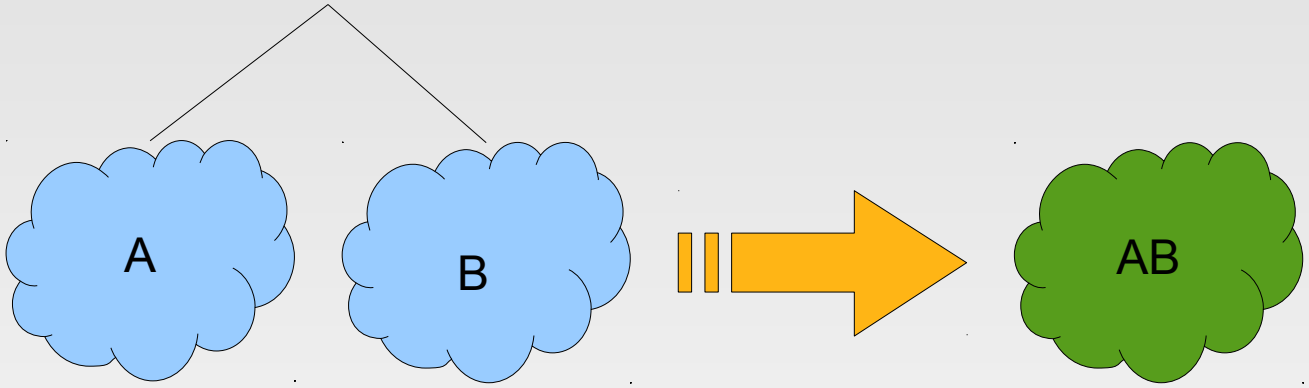


d(A,B) je minimální



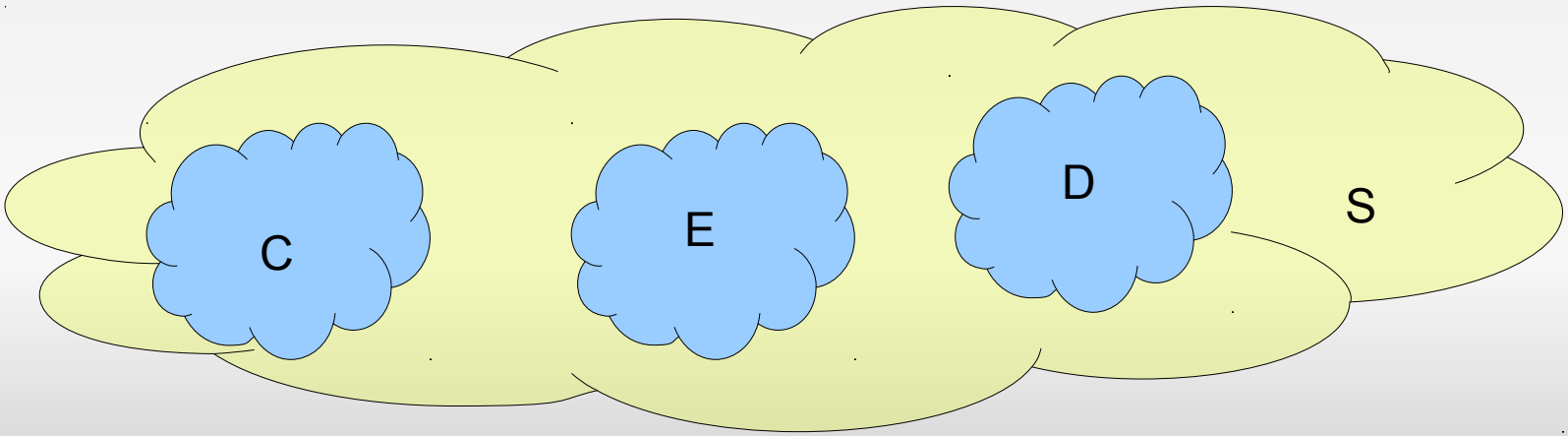
Vytváření stromu - Neighbour joining

- V každém kroku spojí dva klastry, které jsou blízko sobě a zároveň daleko od ostatních klastrů



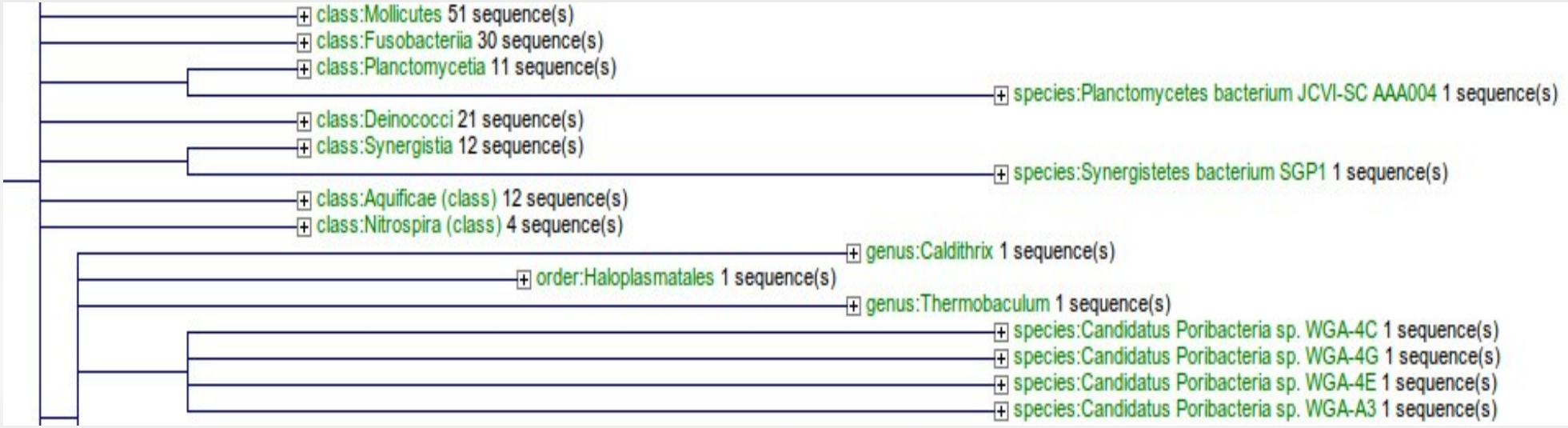
Nový klastr AB

$$|S|\cdot d(A,B)-\sum_{X\in S} d(A,X)-\sum_{X\in S} d(B,X)$$ je minimální

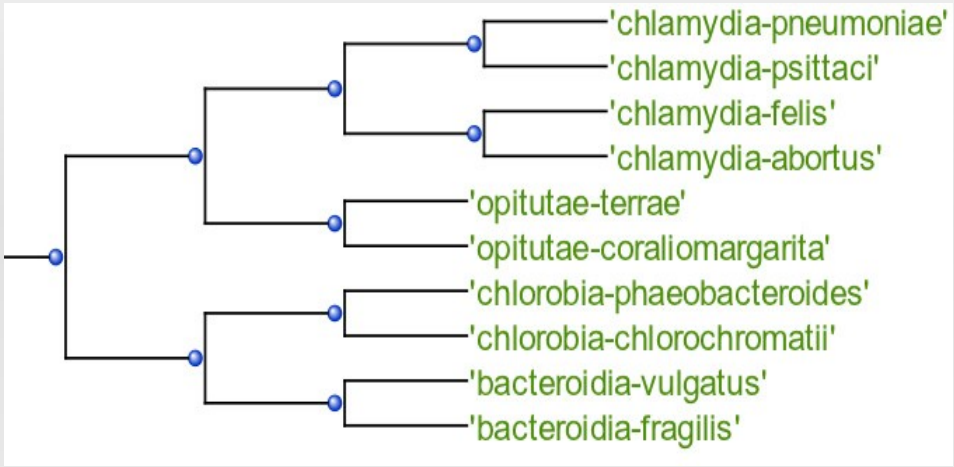


Výběr testovacích dat

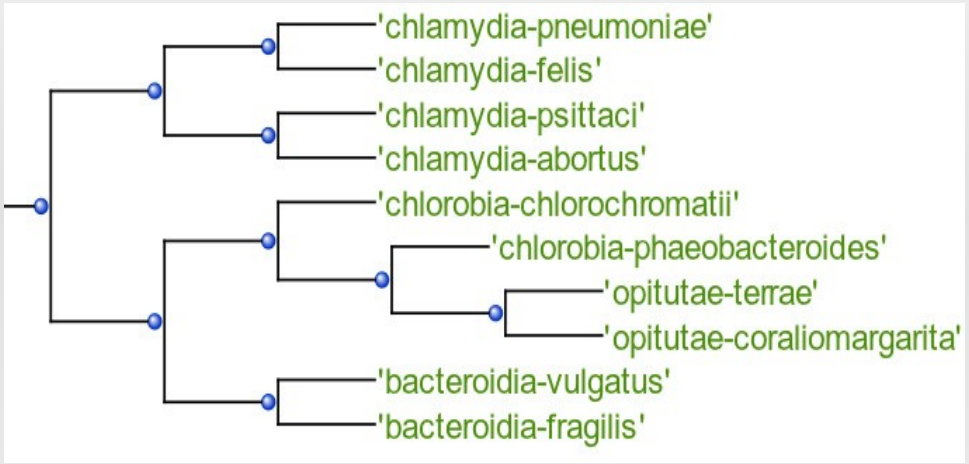
- Porovnáváme celé genomy organismů
  - K použití více specifických dat nám chybí lepší biologické znalosti
  - Nevýhoda: příliš dlouhé sekvence
- Pro vyhodnocení výsledků je třeba expertní taxonomie
  - U bakterií dobře vypracováno:



Výsledky

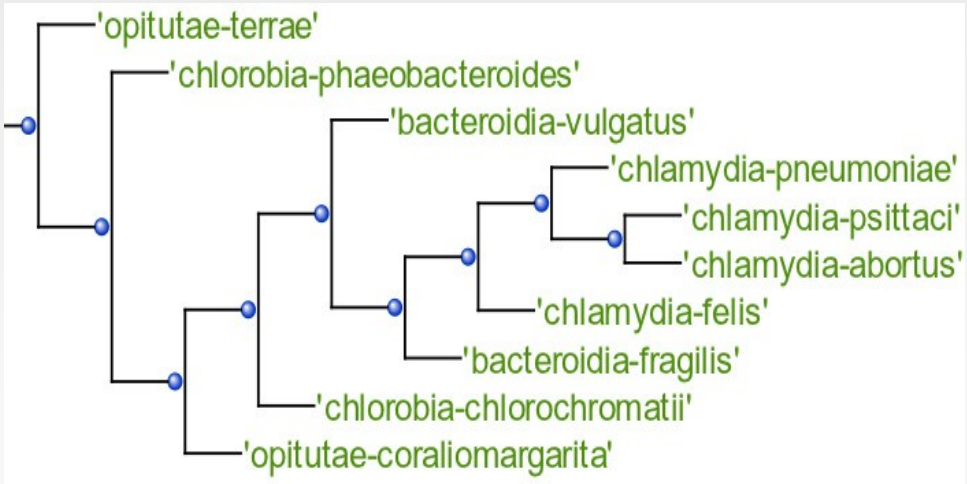


Skutečná taxonomie



Histogram (w=10) + Neighbour joining

Nejlepší, čeho jsme dosáhli



Bzip2+UPGMA – jak moc je to špatný výsledek?

Diskuze

- Nejspolehlivější výsledky:
  - Histogram + Neighbour joining
- Algoritmus GenCompress se nám na dlouhých sekvencích nepodařilo použít
- Problém: jak hodnotit kvalitu výsledků?
  - My je hodnotili ”podle oka”
- Nebyly dobré výsledky jenom náhoda?
  - další testování nezbytné, na větším vzorku organismů
- Co by šlo zlepšit?
  - vybrat pro porovnávání specifickou část genomu s důležitou informací – nutná rada experta