

Fylogenetické stromy

Porovnání metod

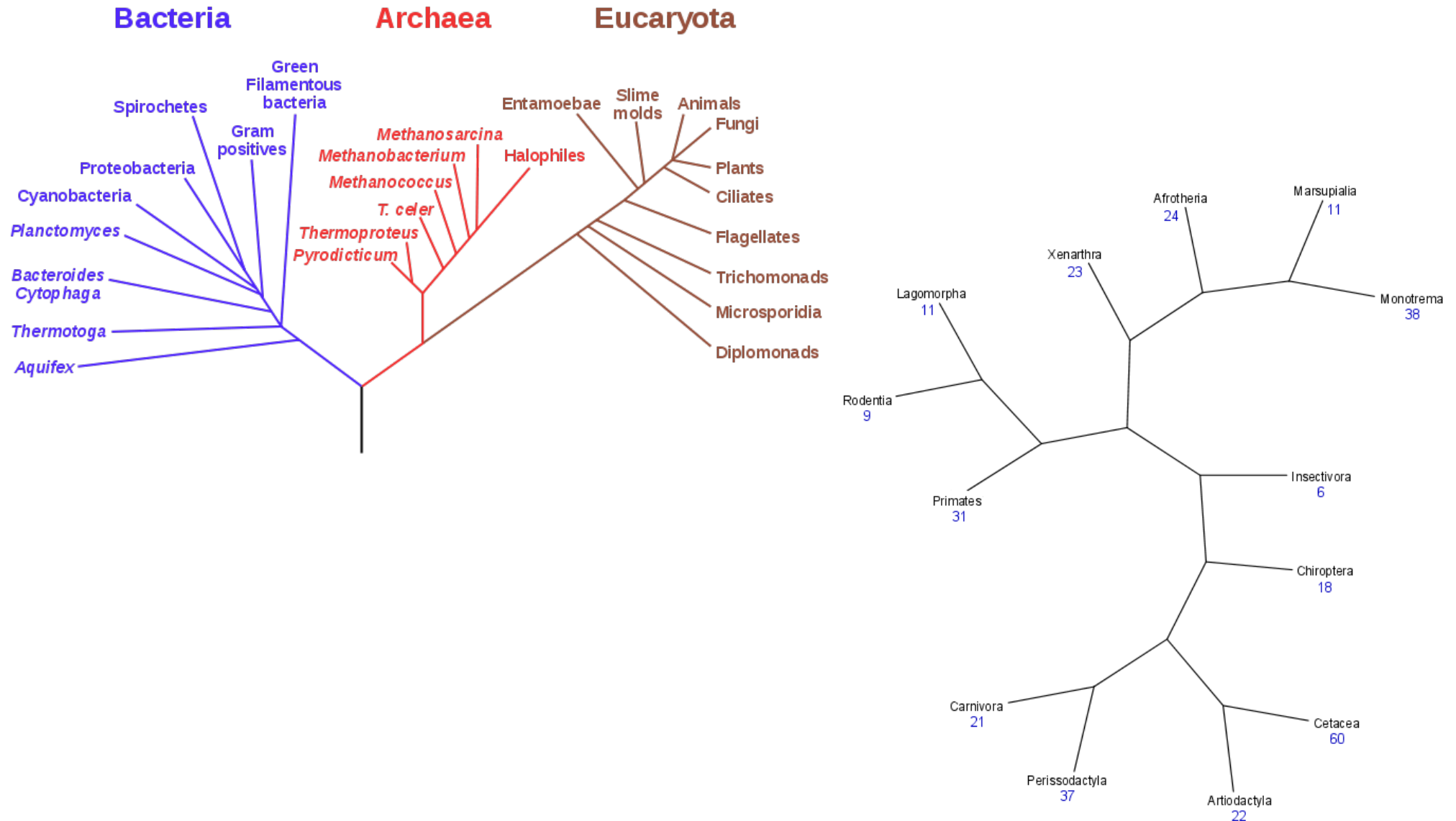
Michal Roder, Jiří Vejmola

Co to je?

- "Strom života", evoluční strom
- znázornění evolučních vztahů mezi (různými) organismy (se společným předkem)
- uzly reprezentují nejbližší předky svých potomků
- délky hran mohou odpovídat časové vzdálenosti
- listy představují zkoumané organismy
- základní typy
 - zakořeněné (rooted)
 - předpokládají jednoho společného předka
 - nemusí být definovaný
 - bezkořenové (unrooted)
 - především příbuznost organismů

Ukázky stromů

Phylogenetic Tree of Life



Data

- zdroj: Influenza Virus Resource
 - <http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html>
- parametry:
 - typ - A
 - podtyp - H1N1
 - hostitel - člověk
 - protein - hemagglutinin
 - bez místa určení
 - do 31.12.1999
- sebrané vzorky:
 - 15 náhodně vybraných
 - různé státy
 - různá data

Použité programy

- ClustalW2
 - <http://www.ebi.ac.uk/Tools/msa/clustalw2/>
 - příprava dat
- PHYLIP
 - <http://evolution.genetics.washington.edu/phylip.html>
 - balík programů pro fylogenetiku
- BAli-Phy
 - <http://www.biomath.ucla.edu/msuchard/bali-phy/>
 - Bayesovský odhad zarovnání a fylogenetických stromů
- FigTree
 - <http://tree.bio.ed.ac.uk/software/figtree/>
 - prohlížeč dat

Distance-matrix: Fitch-Margoliash

- Distance-matrix
 - genetické vzdálenosti mezi každými dvěma páry sekvencí
 - neslouží k sestavení stromu
 - využito dalšími metodami, jako F-M, Neighbor-joining, ...
- Fitch-Margoliash
 - metoda vážených nejmenších čtverců pro shlukování na základě genetické vzdálenosti
 - nalezení optimálního stromu je NP-úplný problém
 - používání heuristik

Maximum parsimony

- hledání stromu s nejmenším počtem evolučních událostí pro popsání zkoumaných dat případně stromu s nejnižším ohodnocením
- jednotlivé události mohou mít různou váhu
- vhodné když různé události mají různé pravděpodobnosti vzniku
- NP-těžký problém
- využívání heuristik, většinou zahrnující gradientní metody

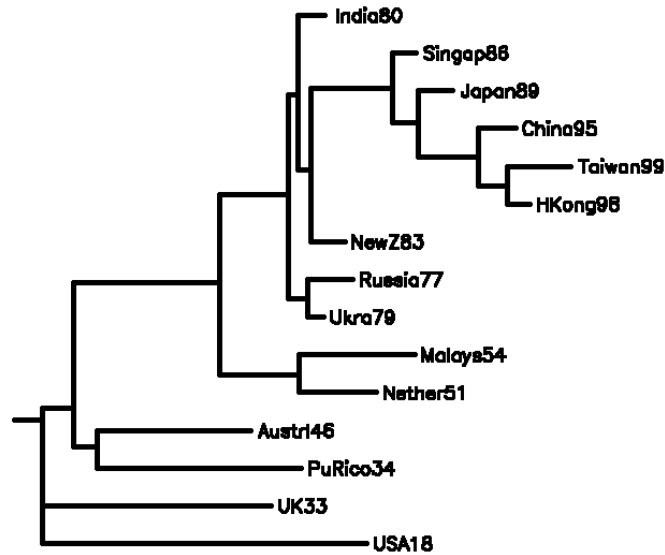
Maximum likelihood

- statistická metoda
- vyžaduje substituční model pro určení pravděpodobností jednotlivých mutací
- stromy vyžadující více mutací na vnitřních vrcholech mají nižší pravděpodobnost
- vhodné pro analýzu vzdáleně příbuzných sekvencí
- nebylo zatím dokázáno jako NP-úplné, avšak stále velmi náročné

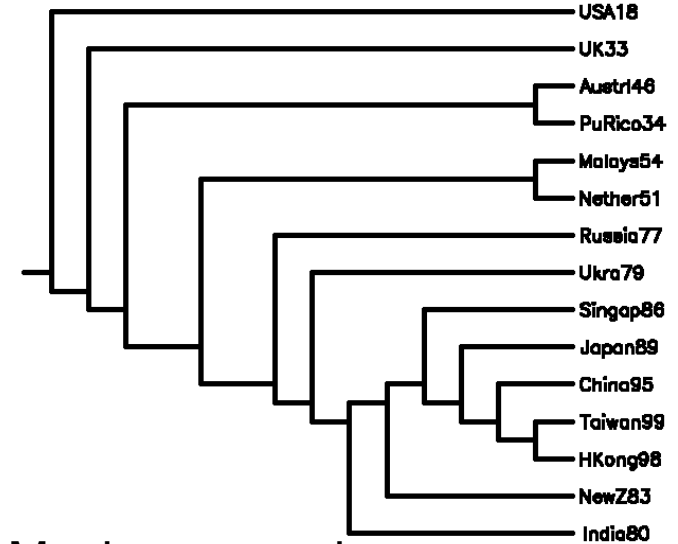
Markov chain Monte Carlo

- třída statistických algoritmů
- generování vzorků distribuce pravděpodobnosti, která je založena na Markovových řetězcích
- Markovovy řetězce slouží k popisu substitucí v sekvencích
- BAli-Phy používá variantu Gibbsova vzorkování k prohledávání prostoru všech zarovnání a stromů
- výpočetně náročné, výsledky se zpřesňují s časem

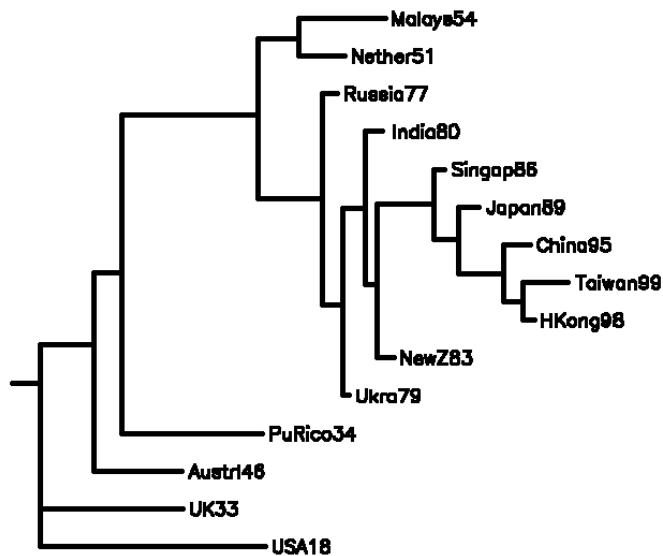
Srovnání - kořenové stromy



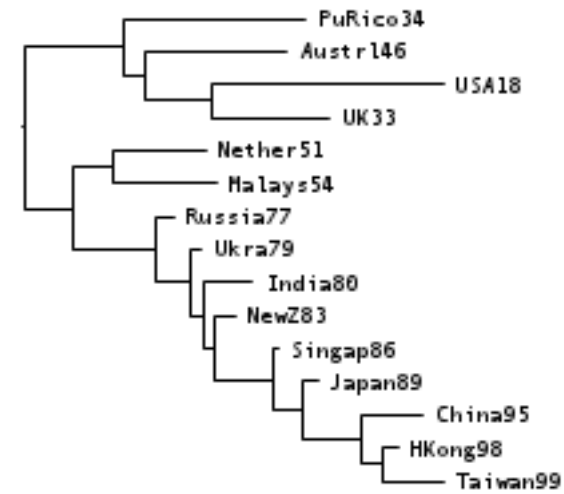
Distance-matrix



Maximum parsimony

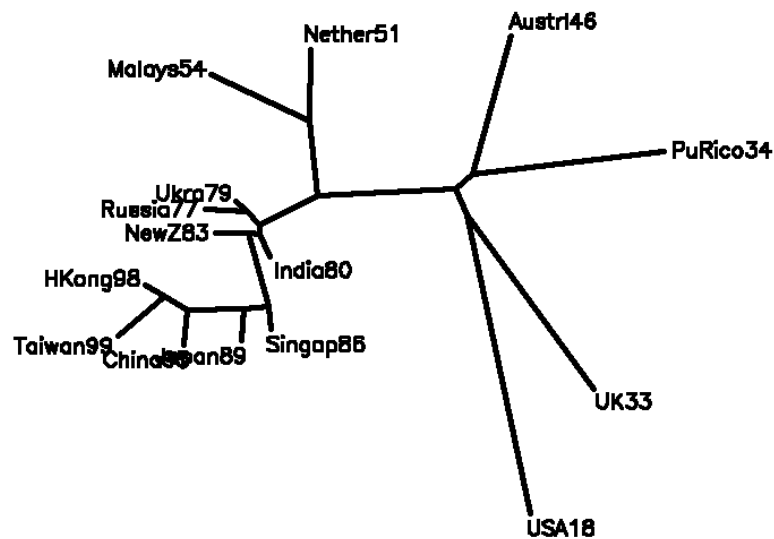


Maximum likelihood

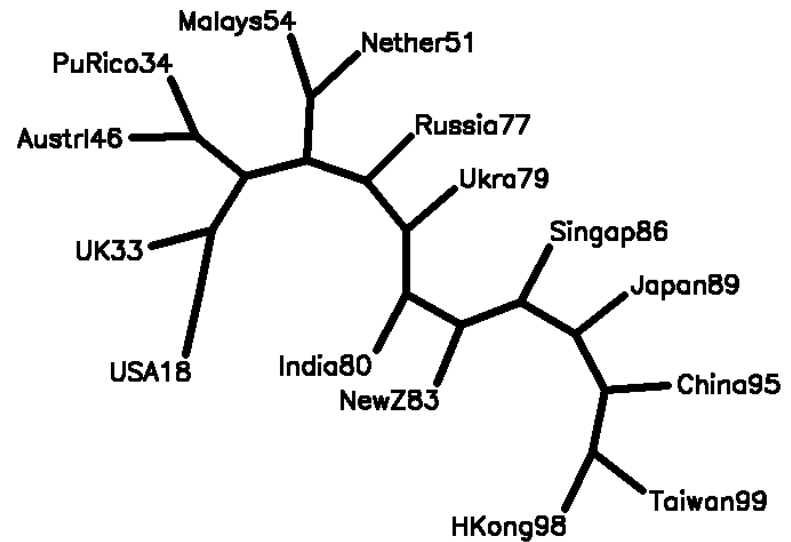


Markov chain Monte Carlo

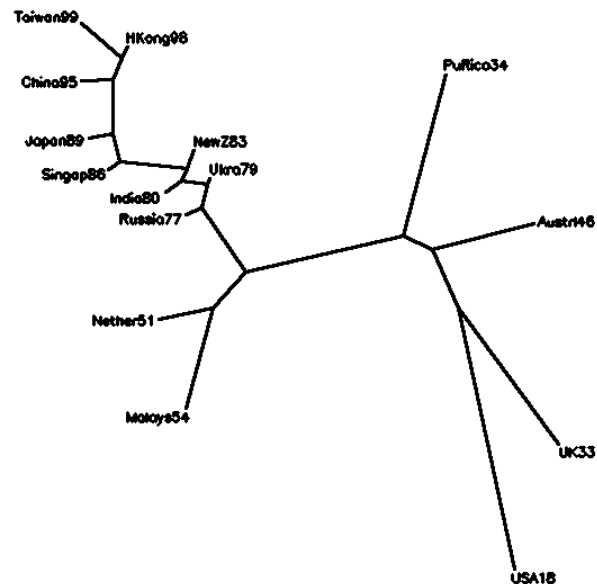
Srovnání - bezkořenové stromy



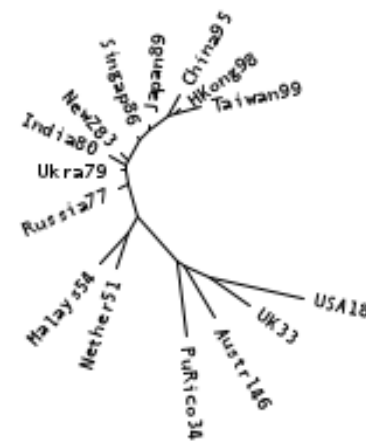
Distance-matrix



Maximum parsimony



Maximum likelihood



Markov Chain Monte Carlo

Porovnání metod

- mezi výslednými stromy pouze minimální rozdíly, základní struktura stejná
- uspořádání především podle stáří
- data vybrána pro prezentaci programů a metod
 - nevhodný výběr dat?

Konec

Děkujeme za pozornost.