Vytváření fylogenetických stromů na základě alignmentů

Tomáš Novotný

Jaroslav Knotek

Alignmenty - opakování

Existují dvě základní varianty alignmentů: globální a lokální

- Globální: Hledáme nejlepší zarovnání dvou celých sekvencí o délkách m a n.
- Lokální: Hledáme takové souvislé podřetězce v obou sekvencích, že jejich globální alignment má maximální možné skóre
- Máme několik možností spočtení skóre – lze vybrat různá ohodnocení pro match, mismatch a gap, použít různé penalizace vzhledem k velikosti díry, apod.
- Alignment lze spočítat dynamickým programováním v čase $O(mn)$ s využitím paměti $O(n)$.

Alignmenty v našem programu

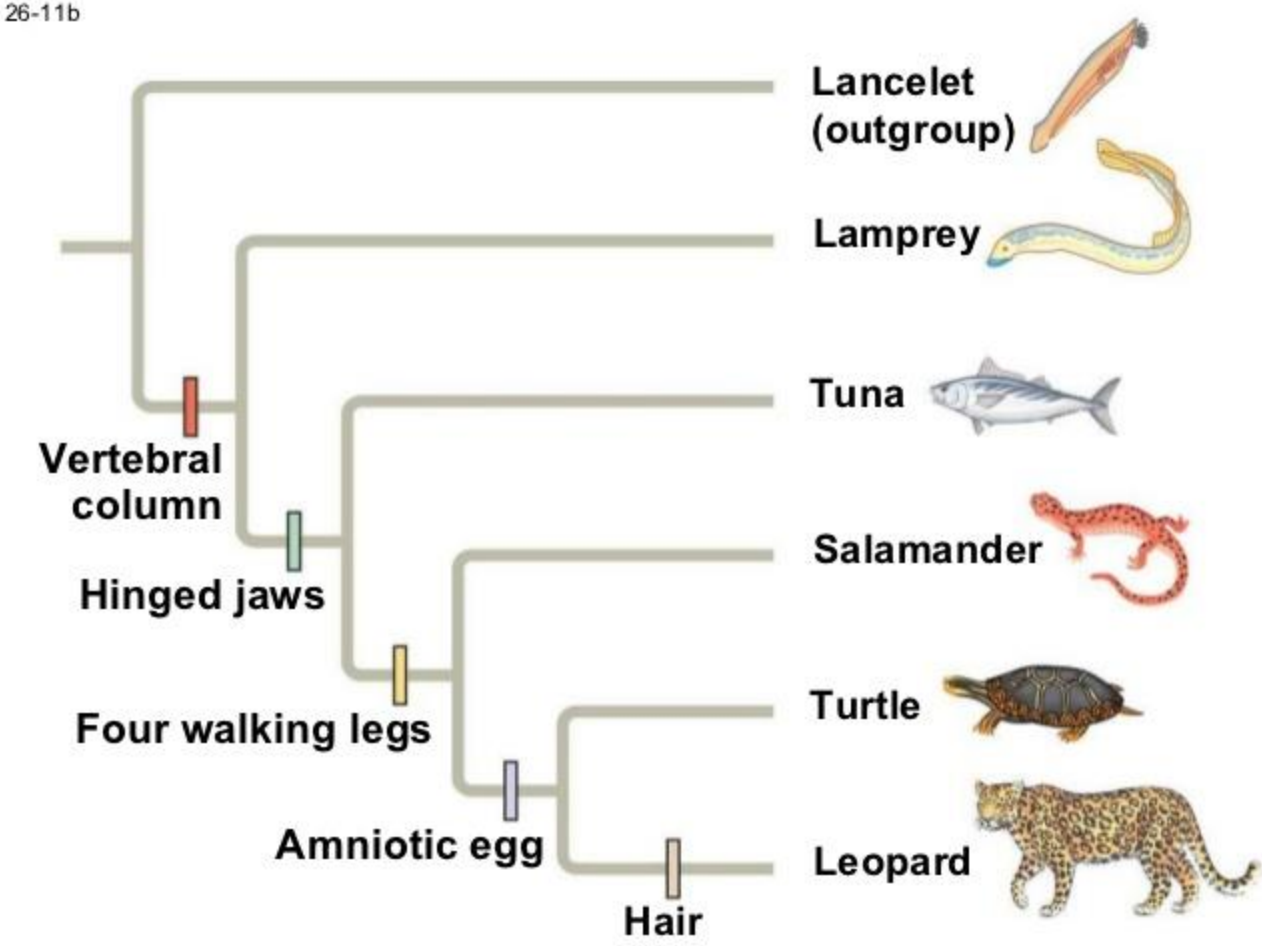
- Smithův-Watermannův algoritmus.
- Upravený S-W algoritmus, který vrací průměr nejvyšších hodnot v tabulce.
- Lokální alignment, jenž má sníženou penalizaci za díry velikosti násobku tří (inspirováno chybějícím kodonem). Opět používáme část nejvyšších hodnot v tabulce.
- Ke každému alignmentu jsme zkoušeli i zarovnání s otočenou sekvencí (mohlo dojít k inverzím částí sekvencí).

Fylogenetický strom - opakování

Neighbor joining tree

- Input: Srovnané sekvence nebo matice vzdáleností
- Postupné seskupování nejbližších dvojic
  - Známe z domácí úlohy
- Output strom příbuzností
- Používáme dva způsoby určení nejbližší dvojice
  - Připojujeme vždy dle nejvyšší hodnoty v tabulce, která vede mezi různými komponentami vytvářeného stromu.
  - Při spojení komponent upravíme hodnoty na průměr přes všechny prvky v komponentě, následně použijeme předchozí krok.
    - Funguje znatelně lépe

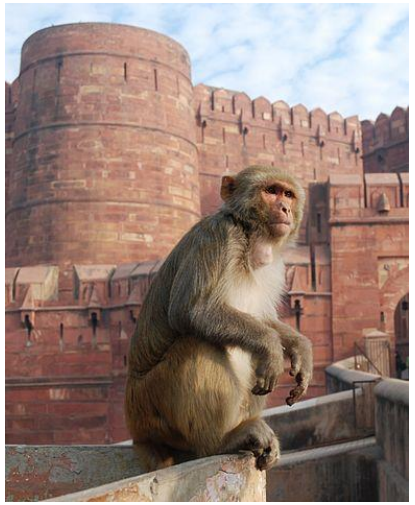
Fig. 26-11b



(b) Phylogenetic tree

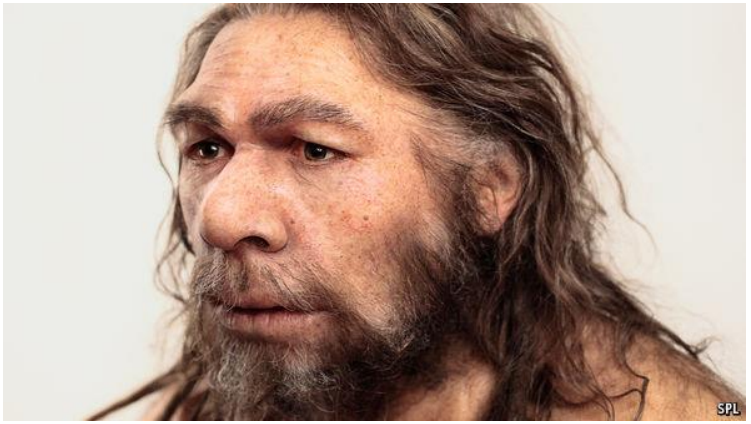
Použitá vstupní data

Canis lupus Vlk obecný



Macaca mulatta Makak rhesus

Homo sapiens
(2 chromosomy) Člověk moudrý



Gorilla gorilla Gorila nížinná

Aquila chrysaetos Orel skalní

Passer domesticus Vrabec domácí

Gallus gallus Kur bankivský

Bombus terrestris Čmelák zemní

Apis mellifera Včela medonosná

Apis cerana Včela východní

Agaricus bisporus Pečárka dvouvýtrusá

Trametes versicolor Outkovka pestrá

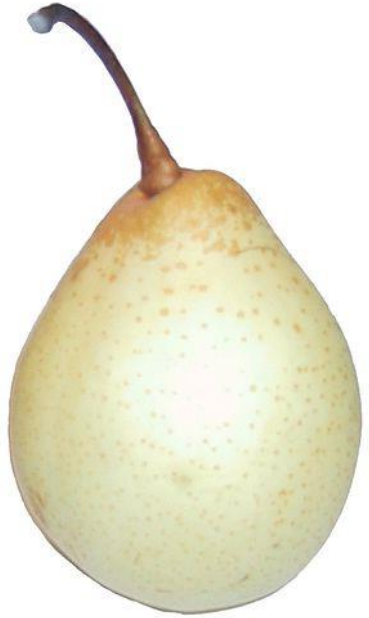
Lactuca sativa Locika setá

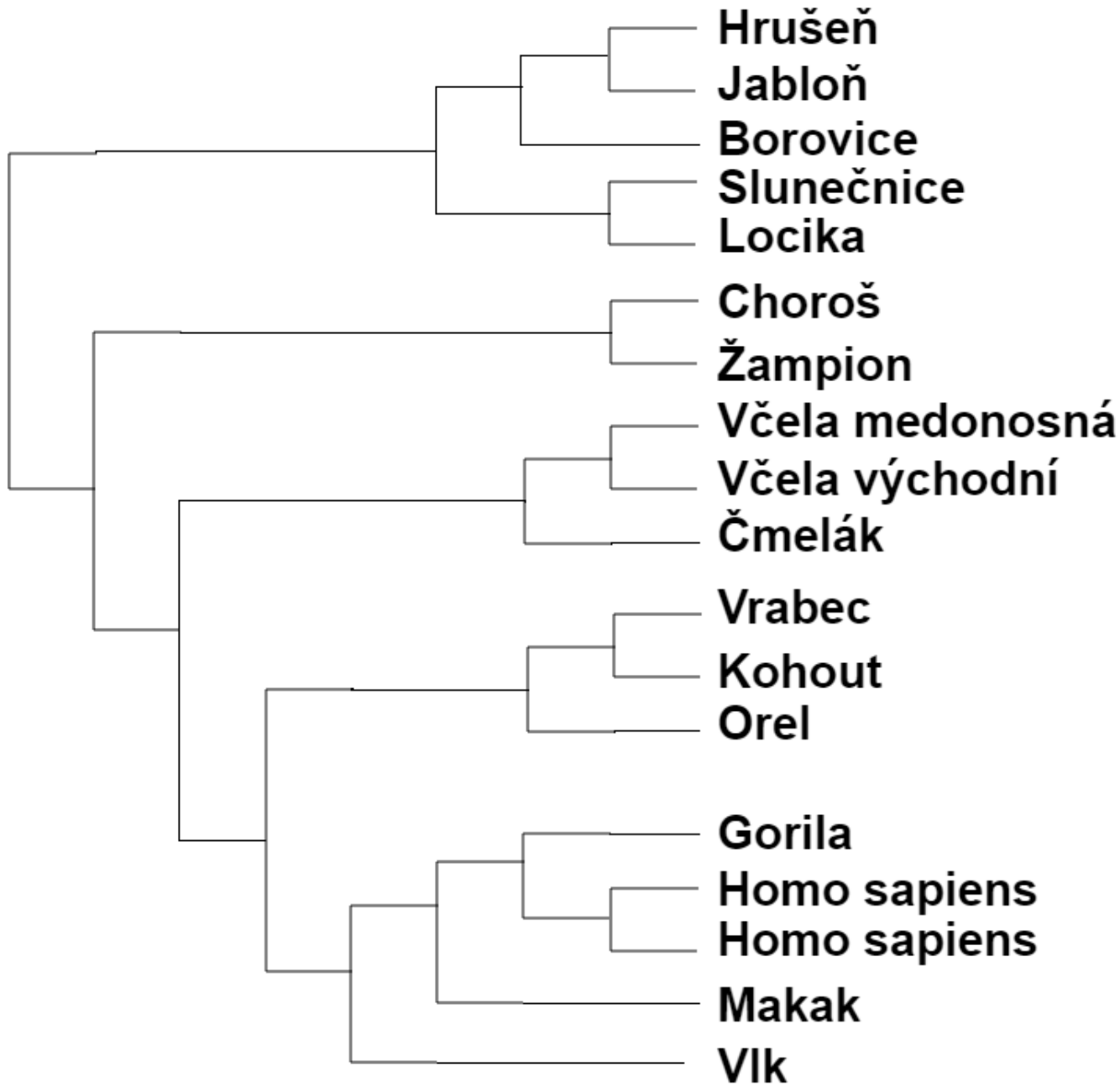
Helianthus annuus Slunečnice roční

Pinus taeda Borovice taeda

Malus domestica Jabloň

Pyrus bretschneideri Hrušeň




Hrušeň
Jabloň
Borovice
Slunečnice
Locika
Choroš
Žampion
Včela medonosná
Včela východní
Čmelák
Vrabec
Kohout
Orel
Gorila
Homo sapiens
Homo sapiens
Makak
Vlk

Testovací data

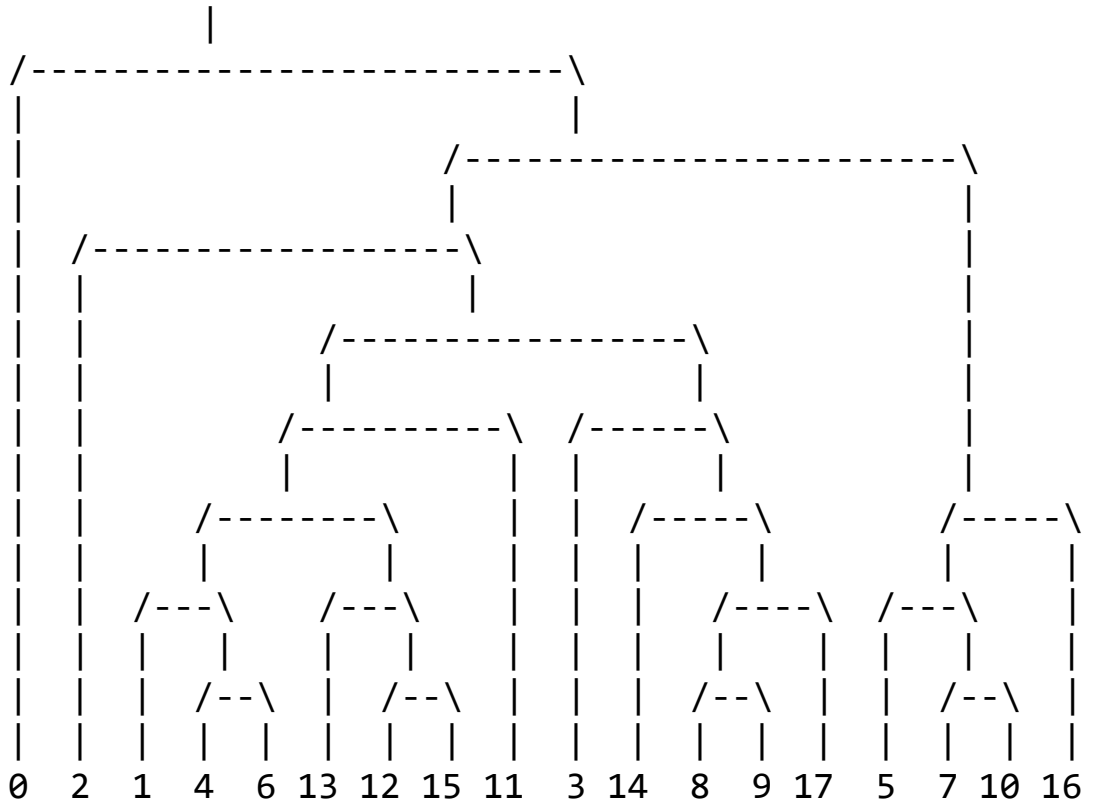
- Zkoušeli jsme dva druhy vstupů:
  - Náhodně vybrané sekvence 30 000 bází z jednotlivých organismů.
  - Dlouhé sekvence různé délky (stovky tisíc bází).
- Pro kratší variantu jsme vytvořili tabulku alignmentem kompletních sekvencí.
- Pro delší variantu jsme pro každou dvojici organismů vybírali sadu náhodných kratších vzorků.
- Z principu fungování algoritmu dosahujeme lepších výsledků při alignmentu delších sekvencí.

Porovnávací metoda

- Jako skóre algoritmů jsme použili míru shody s očekávaným stromem
  - Všechna spojení kromě nejvyššího získala skóre $(1 - \#chyb/max.\#chyb)^2$ pro nejlepší z netriviálních podstromů očekávaného stromu.
  - Zcela náhodný strom získával kolem 4 ze 16 možných.

Naše výsledky

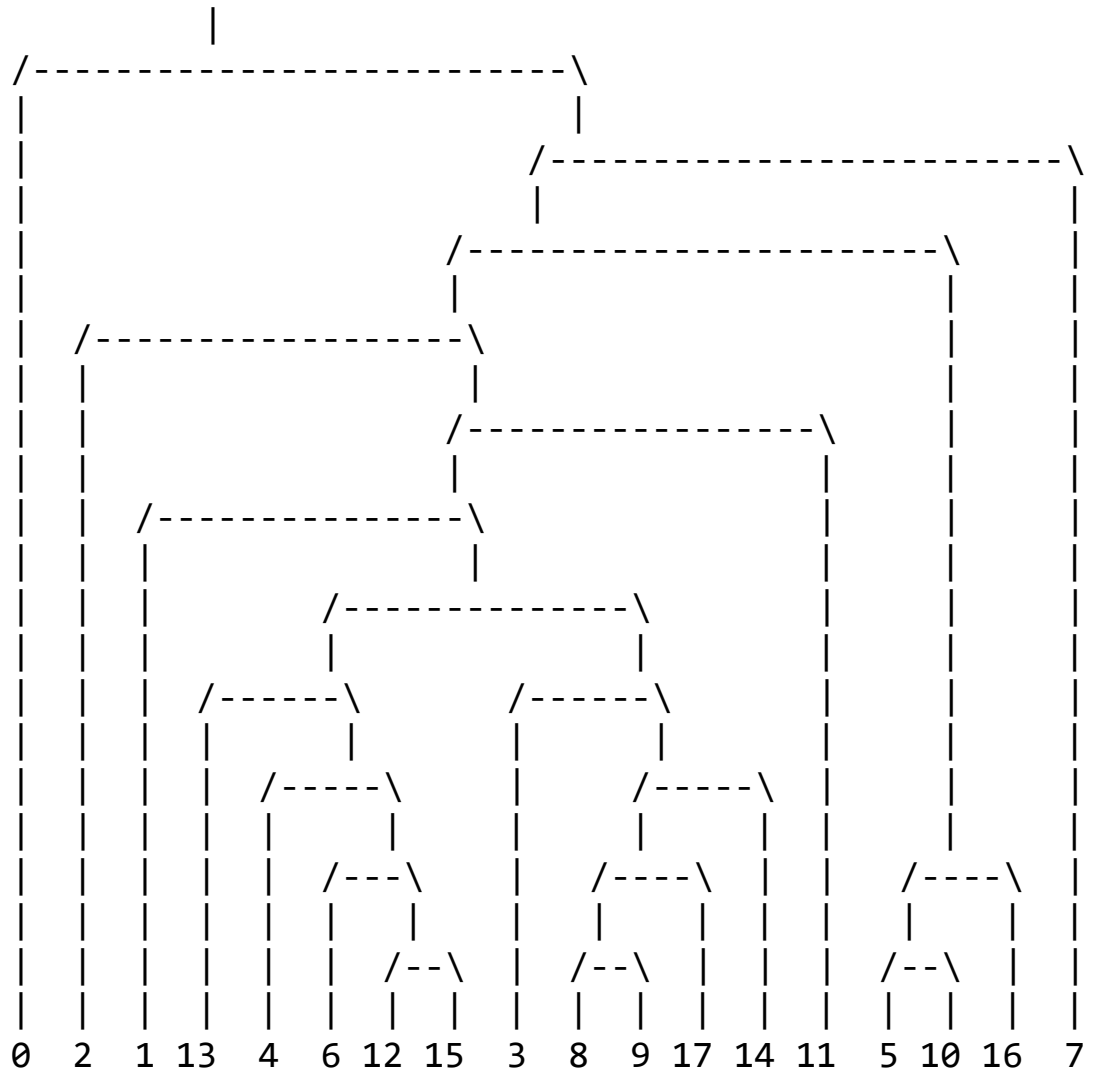
- 30000 bází
- Běžný lokální alignment s použitím inverze
- Průměrovací strom
- Match score: 12/16
- Perfect hits: 8/16



0-žampion
1-jabloň
2-čmelák
3-vlk
4-včela mednonosná
5-Orel
6-včela východní
7-vrabec
8,9-člověk
10-kohout
11-hrušeň
12-locika
13-borovice
14-makak
15-slunečnice
16-choroš
17-gorila

Naše výsledky

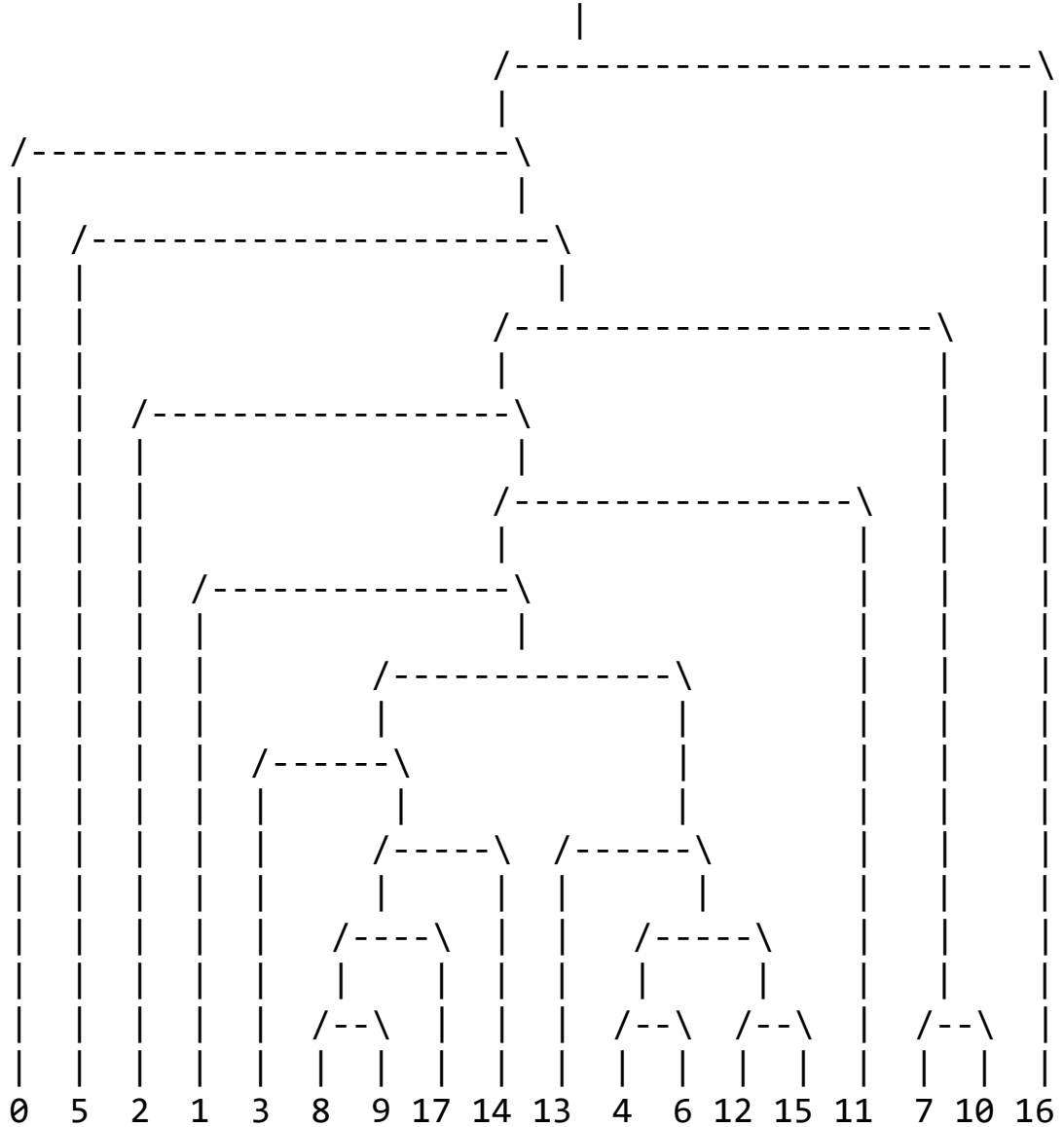
- 30000 bází
- Vícehodnotový alignment s použitím inverze
- Maxiamlizační strom
- Match score: 9.15/16
- Perfect hits: 5/16



0-žampion
1-jabloň
2-čmelák
3-vlk
4-včela mednonosná
5-Orel
6-včela východní
7-vrabec
8,9-člověk
10-kohout
11-hrušeň
12-locika
13-borovice
14-makak
15-slunečnice
16-choroš
17-gorila

Naše výsledky

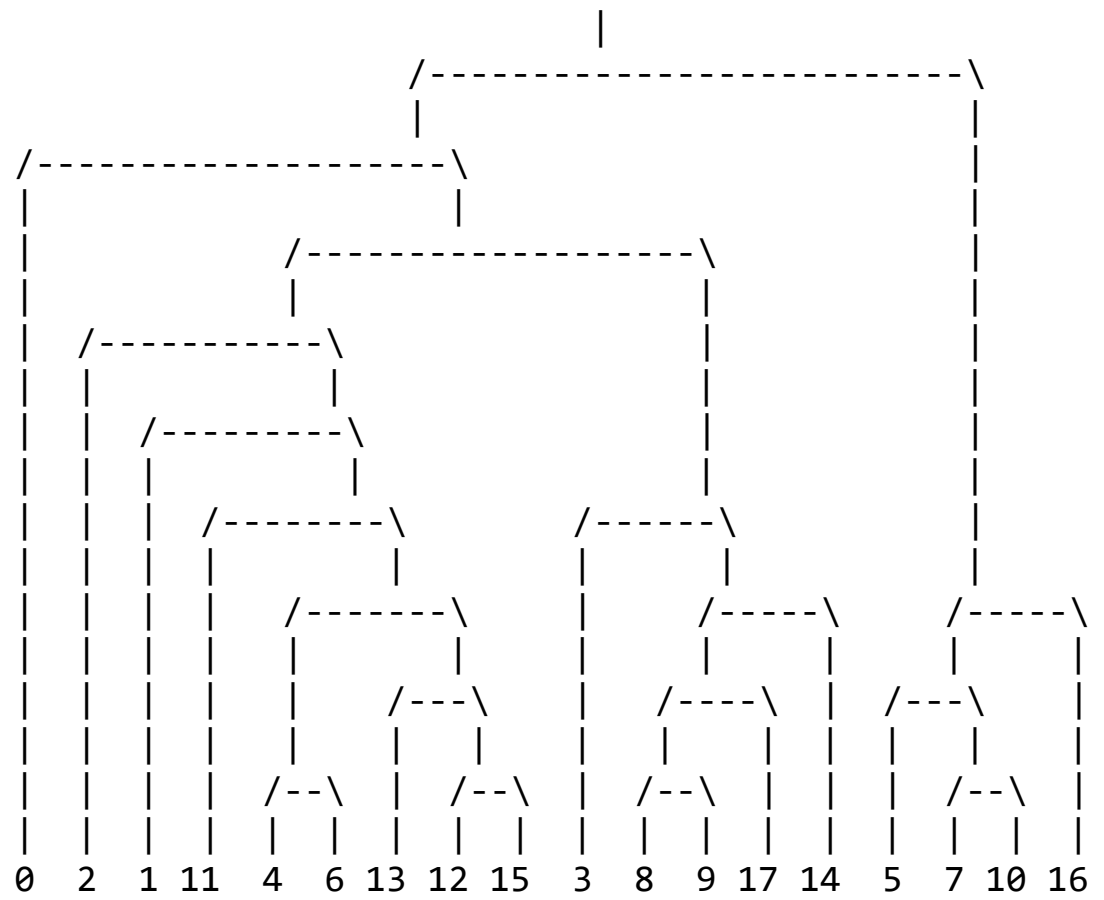
- 30000 bází
- Kodonový alignment s použitím inverze
- Maximalizační strom
- Match score: 9.71/16
- Perfect hits: 7/16



0-žampion
1-jabloň
2-čmelák
3-vlk
4-včela mednonosná
5-Orel
6-včela východní
7-vrabec
8,9-člověk
10-kohout
11-hrušeň
12-locika
13-borovice
14-makak
15-slunečnice
16-choroš
17-gorila

Naše výsledky

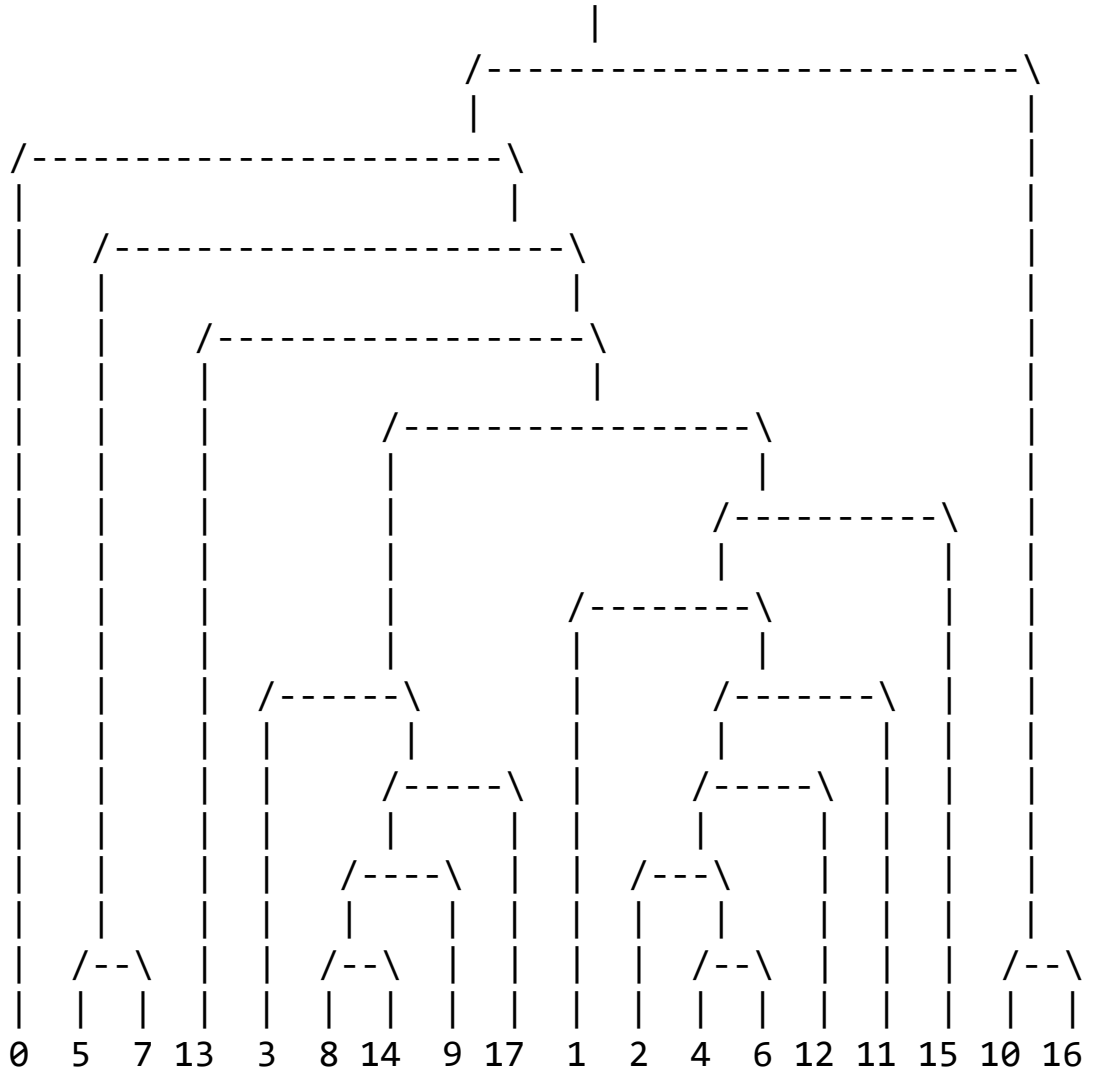
- 30000 bází
- Kodonový alignment s použitím inverze
- Průměrovací strom
- Match score: 12.1/16
- Perfect hits: 8/16



0-žampion
1-jabloň
2-čmelák
3-vlk
4-včela mednonosná
5-Orel
6-včela východní
7-vrabec
8,9-člověk
10-kohout
11-hrušeň
12-locika
13-borovice
14-makak
15-slunečnice
16-choroš
17-gorila

Naše výsledky

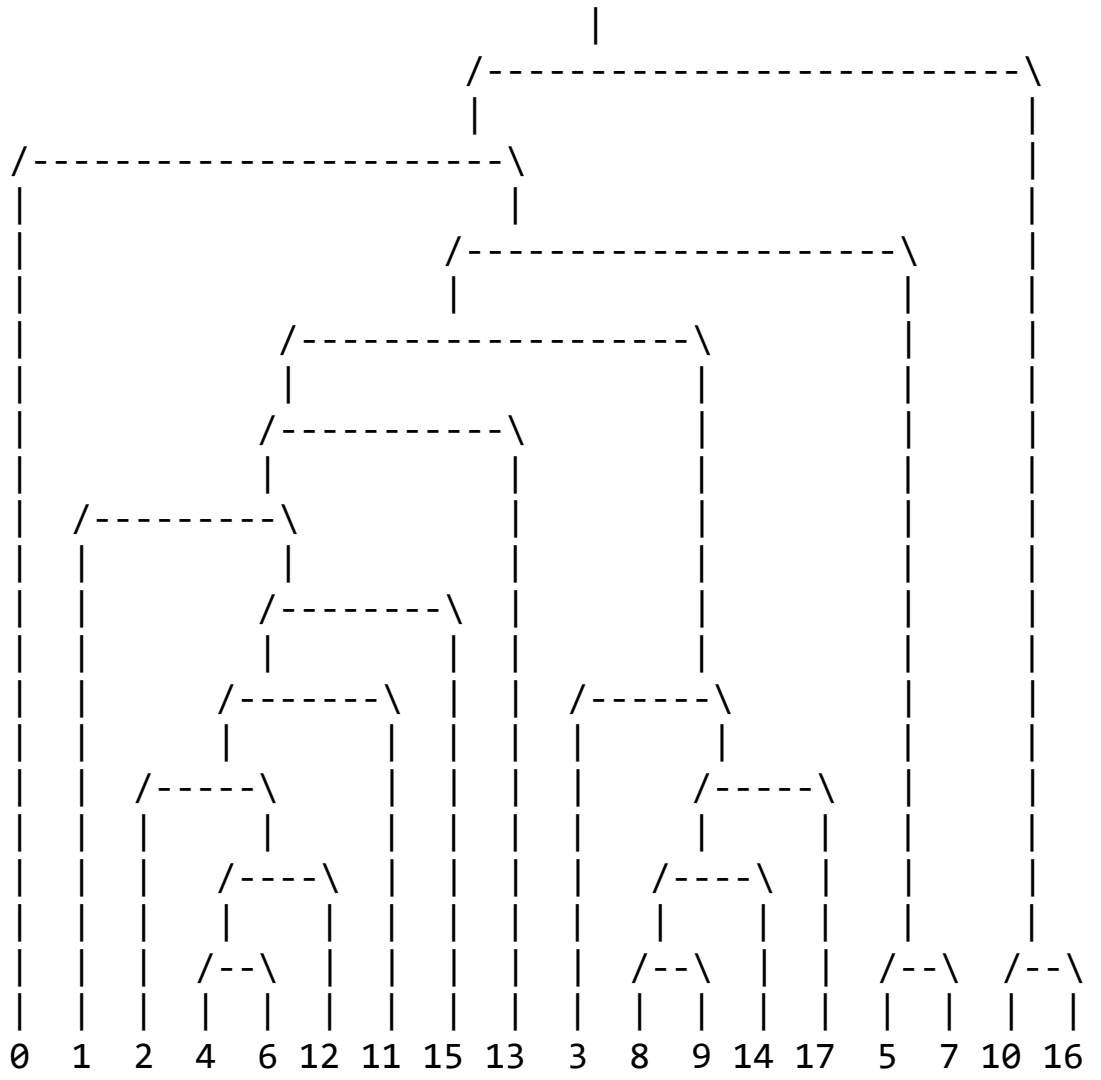
- Vzorky 100x3000 bází
- Vícehodnotový alignment
- Průměrovací strom
- Match score: 9.141/16
- Perfect hits: 4/16



0-žampion
1-jabloň
2-čmelák
3-vlk
4-včela mednonosná
5-Orel
6-včela východní
7-vrabec
8,9-člověk
10-kohout
11-hrušeň
12-locika
13-borovice
14-makak
15-slunečnice
16-choroš
17-gorila

Naše výsledky

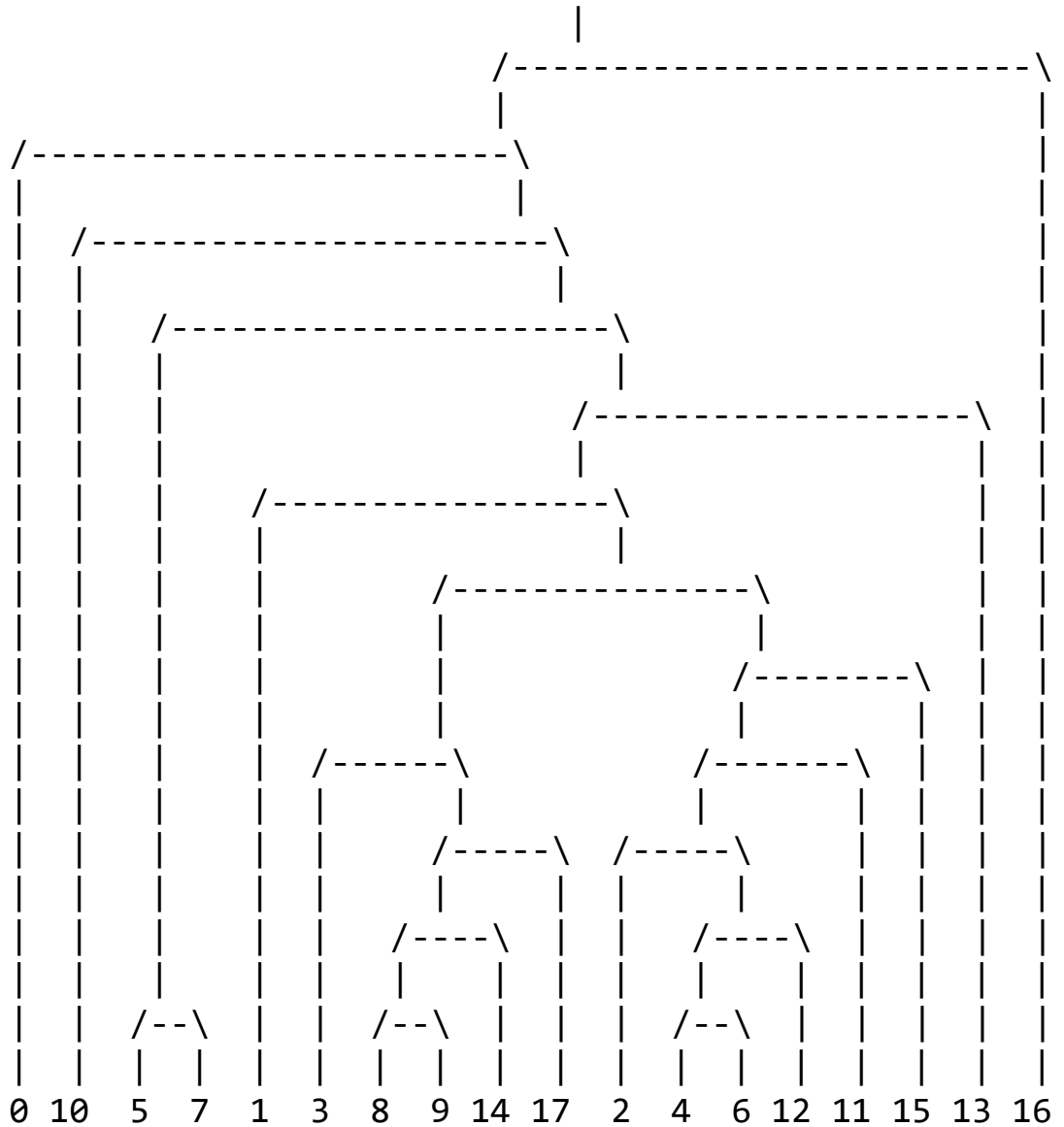
- Vzorky 100x3000 bází
- Kodonový alignment
- Průměrovací strom
- Match score: 9.632/16
- Perfect hits: 5/16



0-žampion
1-jabloň
2-čmelák
3-vlk
4-včela mednonosná
5-Orel
6-včela východní
7-vrabec
8,9-člověk
10-kohout
11-hrušeň
12-locika
13-borovice
14-makak
15-slunečnice
16-choroš
17-gorila

Naše výsledky

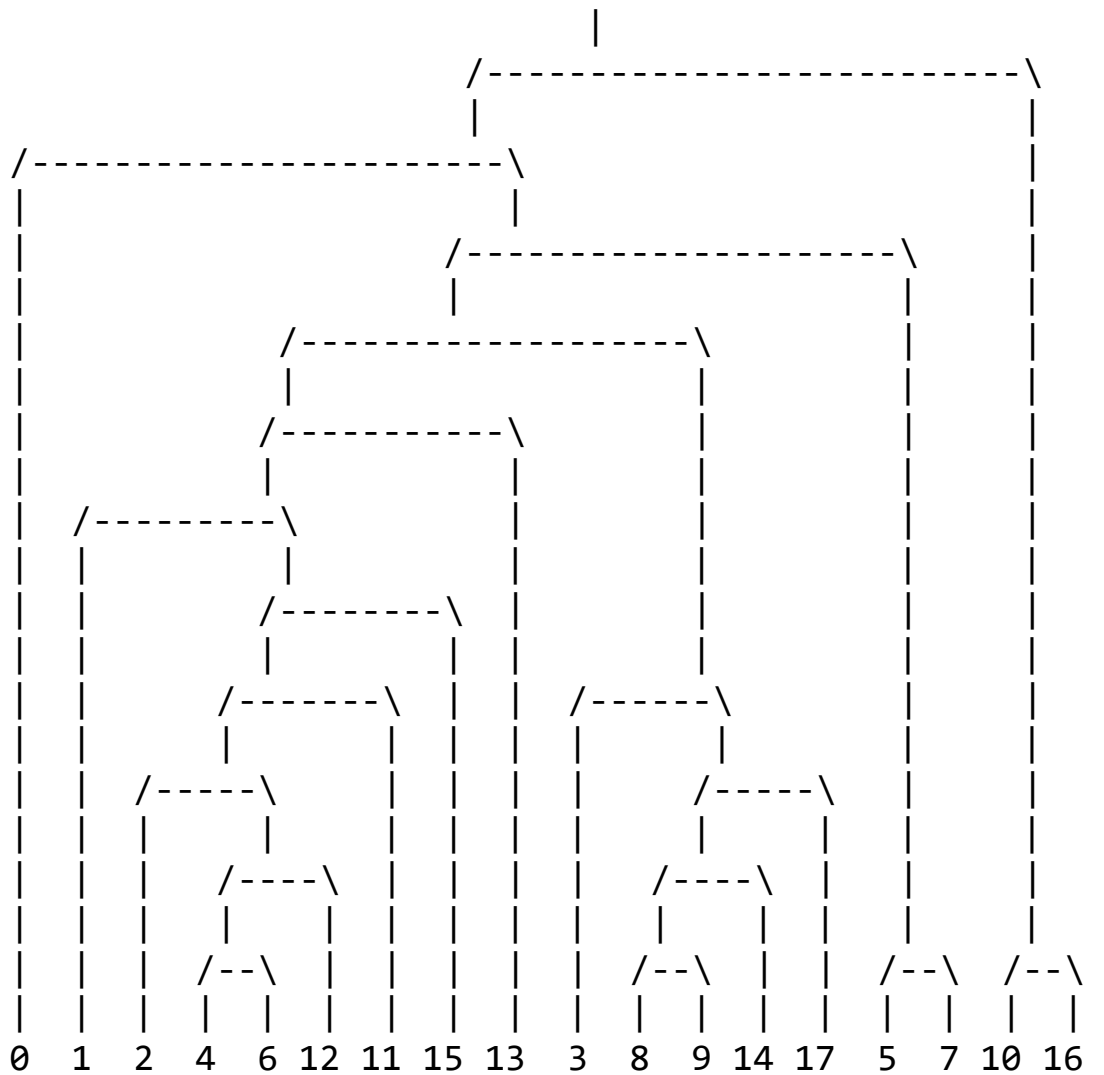
- Vzorky 100x3000 bází
- Kodonový alignment s použitím inverze
- Maximalizační strom
- Match score: 9.222/16
- Perfect hits: 4/16



0-žampion
1-jabloň
2-čmelák
3-vlk
4-včela mednonosná
5-Orel
6-včela východní
7-vrabec
8,9-člověk
10-kohout
11-hrušeň
12-locika
13-borovice
14-makak
15-slunečnice
16-choroš
17-gorila

Naše výsledky

- Vzorky 100x3000 bází
- Kodonový alignment s použitím inverze
- Průměrovací strom
- Match score: 9.582/16
- Perfect hits: 4/16



0-žampion
1-jabloň
2-čmelák
3-vlk
4-včela mednonosná
5-Orel
6-včela východní
7-vrabec
8,9-člověk
10-kohout
11-hrušeň
12-locika
13-borovice
14-makak
15-slunečnice
16-choroš
17-gorila

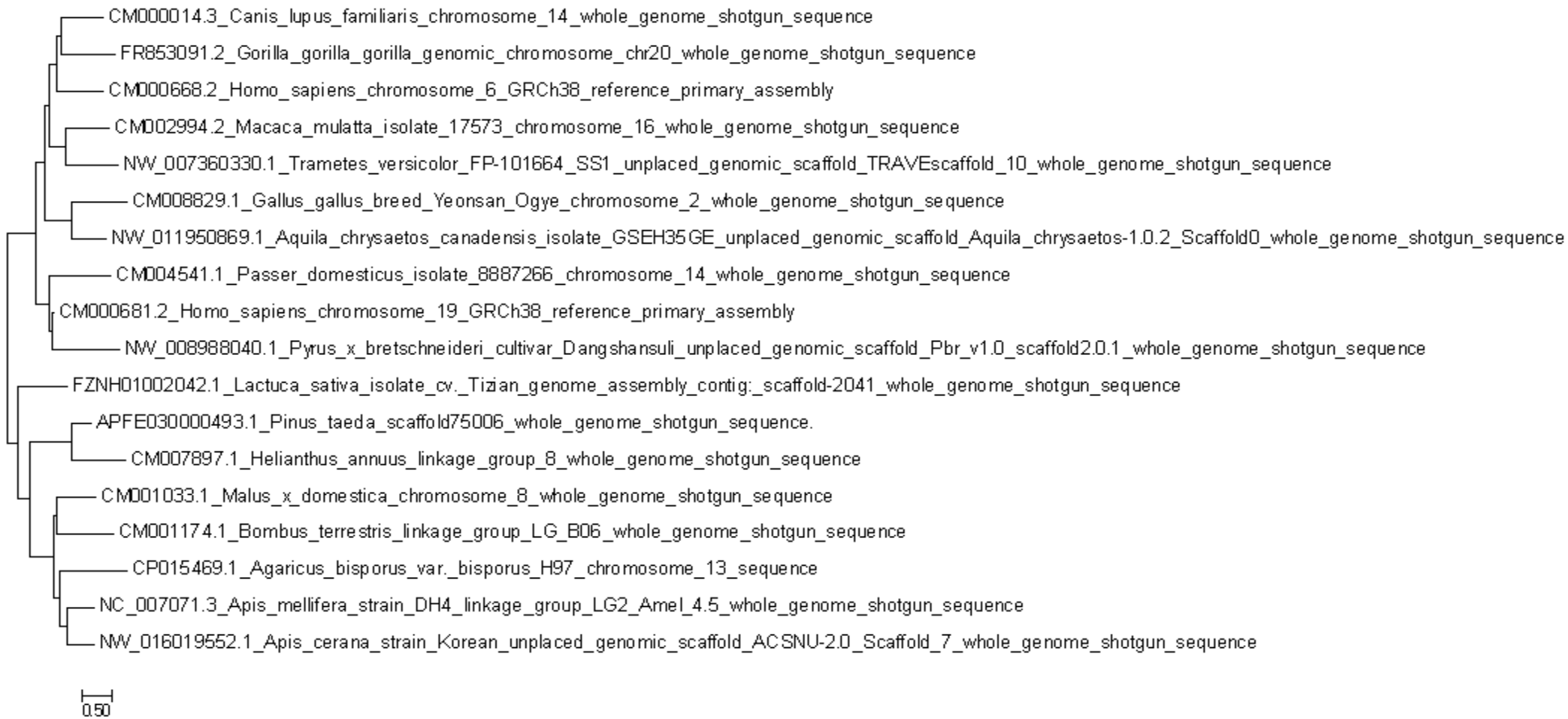
Shrnutí

- Ukázalo se, že použití průměrné hodnoty v podstromech funguje výrazně lépe, než maximální hodnoty.
- Na druhou stranu, výsledky různých typů algoritmů a použití i obrácené sekvence nemělo na výsledky statisticky významný vliv.
- Pro kompletní alignment sekvencí o délce 30 000 bází bylo průměrné skóre průměrovacího vytváření stromu 11,4.
- Pro vzorkování na plných sekvencích dosahovalo horších výsledků (průměrné skóre 9,2).
- Doba běhu cca 30 minut (bez invertované sekvence).

Použité komerční programy

- Mega7
  - ClustalW
    - Improved Clustal
  - MUSCLE
    - Draft progressive
    - Improved progressive
    - Refinment
    - Náhrada ClustalW

Mega - Clustal

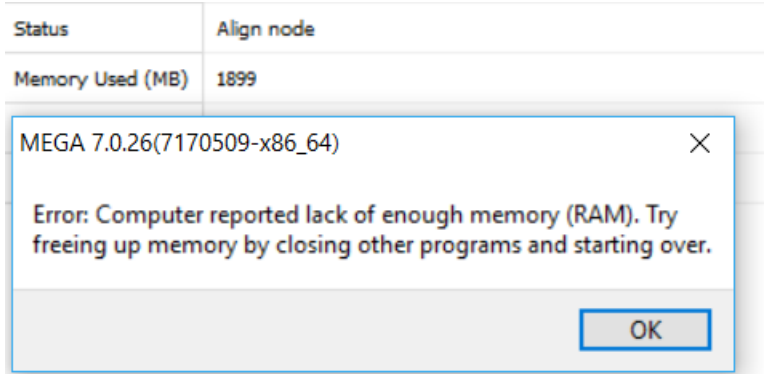


Mega Clustal

- Nepřesné
  - Pro krátké sekvence: Perfect hits: 1/16, match score: 7.777/16.
  - Pro dlouhé sekvence běhové problémy, skóre odpovídající náhodnému stromu.
- Dlouhé trvání
- Použito defaultní nastavení

Další výsledky

- Chyby
  - Mega v. 6, v. 7



  - www.ebi.ac.uk

We have not been able to format the results of this job (clustalo-I20180118-084546-0032-9047626-p1m). This could be because the job has failed to complete.

  - Genome.jp

The proxy server could not handle the request *POST /tools-bin/clustalw*.

Zdroje

- Wikipedia.org – obrázky i reference
- Clustal paper [*Higgins DG, Sharp PM (December 1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer".]*
- *Online clustal* - https://www.ebi.ac.uk/Tools/msa/clustalo/
- Sekvence - https://www.ncbi.nlm.nih.gov/nuccore