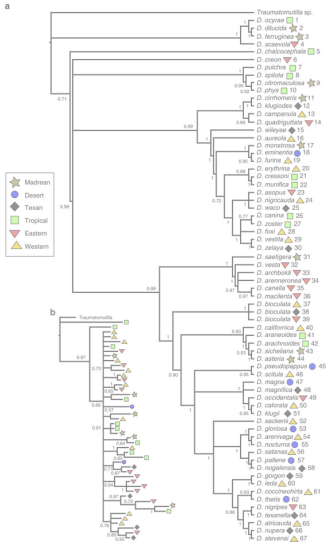
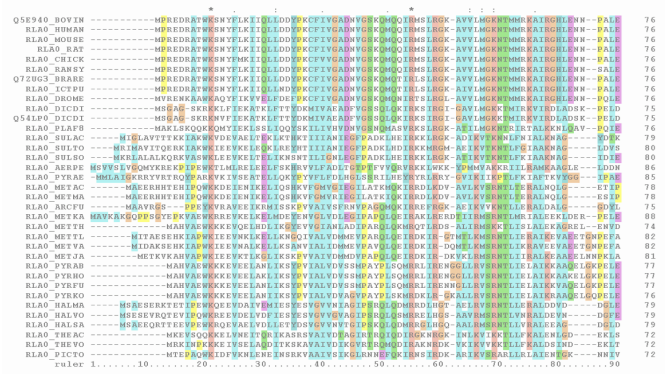
Praktične vaje pri predmetu

Biokemijska informatika

9. vaja

Prileganje zaporedij in filogenetska drevesa





Miha Pavšič
maj 2014

Prileganja večih zaporedij in na njihovi osnovi zgrajena filogenetska drevesa nam omogočajo:

- razumevanje bioloških procesov
- ugotovavljanje povezav/podobnosti/sorodnosti med proteini
- sklepanje na biološko funkcijo proteinov

Koraki:

1. Identifikacija problema ter nukleotidnega/aminokislinskega zaporedja.
2. Iskanje/identifikacija drugih zaporedij, ki so sorodna našemu.
3. Izračun prileganja zaporedij in analiza prileganja.
4. Priprava filogenetskega drevesa in analiza.

Iskanje/identifikacija sorodnih zaporedij

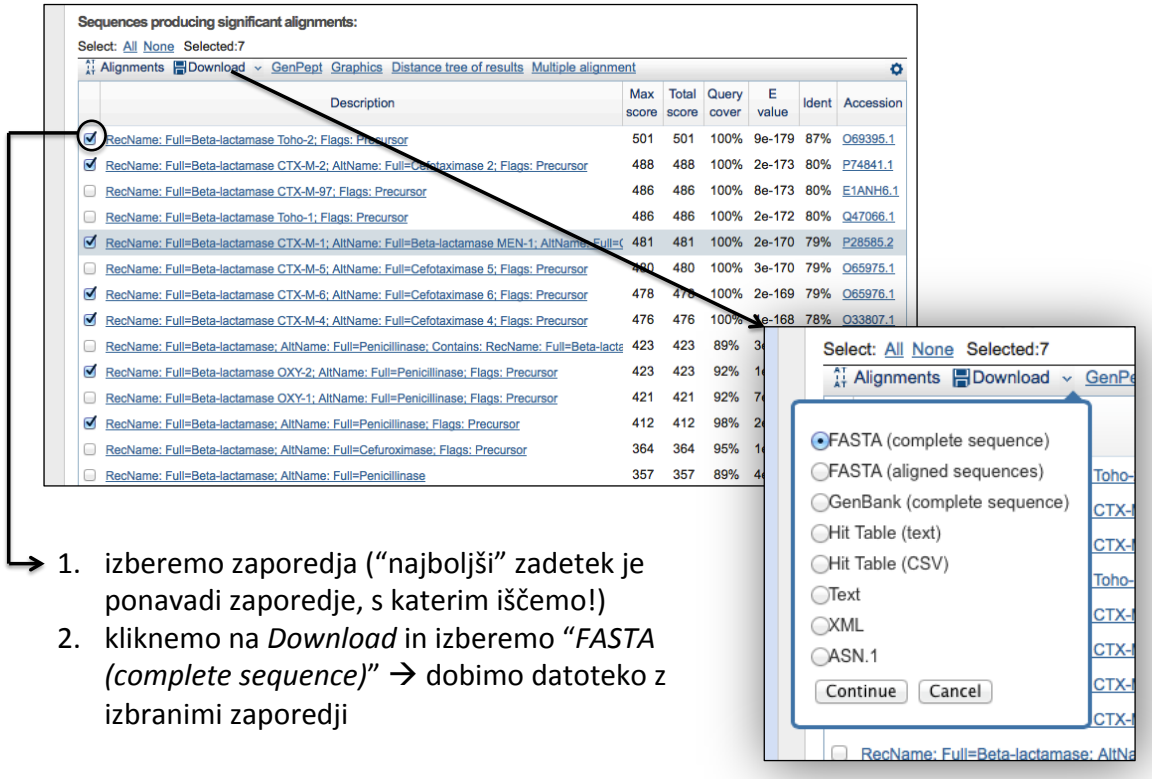
2 osnovna načina:

- **iskanje "na roko"**, na primer direktno v bazi UniProt ali GenBank (predvsem takrat, ko točno vemo, kaj iščemo)
- **iskanje podobnih zaporedij z BLAST** (ko bomo v prileganje vključili veliko število zaporedij in ko ne vemo, koliko/katera zaporedja v bazi so podobna našemu)
  - pozor pri večih vnosih v bazo:
    - če želimo samo reprezentativna zaporedja (brez recimo variacij znotraj iste vrste) izberemo zgolj eno
    - če želimo obširen nabor zaporedij izberemo tudi posamezne variante oz. zapise v bazi za isto zaporedje
  - dolžina prileganja (*coverage*)
  - vrednost E

Izbor sorodnih zaporedij za vključitev v skupno prileganje je osnovan na namenu izdelave le-tega!

Iskanje/identifikacija sorodnih zaporedij: BLAST

Primer rezultatov iskanja z BLAST:



1. izberemo zaporedja ("najboljši" zadetek je ponavadi zaporedje, s katerim iščemo!)
2. kliknemo na *Download* in izberemo "*FASTA (complete sequence)*" → dobimo datoteko z izbranimi zaporedji

Iskanje/identifikacija sorodnih zaporedij

Priprava ustrezno oblikovane vhodne datoteke

- Vsi programi sprejmejo vhodne podatke v obliki FASTA.
- Tudi izbrana zaporedja, dobljena pri iskanju z BLAST, so v formatu FASTA.
- Datoteko oblikujemo z enim od orodij za enostavno urejanje besedil (npr. Beležnica/ Notepad) in si jo shranimo, saj predstavlja naše vhodne podatke.

Ponavadi prilagodimo **IMENA ZAPOREDIJ** (to, kar piše za znakom >), na primer:

- ko primerjamo homologe proteina iz različnih organizmov ponavadi uporabimo kar imena organizmov:
  - human, mouse, rat, cow, ...
  - Hsapiens, Mmusculus, Rnorvegicus, Btaurus, ...
  - Hsap, Mmus, Rnor, Btau, ...
- lahko navedemo krajše ime proteina (kratico) in skrajšano ime organizma:
  - FHL2human, FHL2mouse, FHL2rat, ...
  - FHL2_human, FHL2_mouse, FHL2rat, ...
  - FHL2_Hs, FHL2_Mm, FHL2_Rn, ...

Ponavadi v imenih ne vključimo presledkov (lahko pa so) in si prizadevamo, da so **čim krajša in jedrnata** (vsebujejo podatke, ki jih potrebujemo (odvisno od analize)) ter da nista dve imeni enaki!

Iskanje/identifikacija sorodnih zaporedij

Priprava ustrezno oblikovane vhodne datoteke - primer

```
>bBlac_Toho-2
MVTKRVQRMMSAAAACIPLLLGSPTLYAQTSAVQQKLAALEKSSGGRLGVALIDTADNTQVLYRGDERFPMCSTSKVMAA
AAVLKQSETQKQLLNQPVEIKPADLVNYNPIAEKHVNGTMTLAELSAAALQYSDNTAMNKLIAQLGGPGGVTAFARAIGD
ETFRLDRTEPTLNTAIPGDPRDTTTARAGADVASLRWVMRWAKPSGAVGDVAQRQYDRAAGIRAGLPTSWTVGDKTGSGD
YGTTNDIAVIWPQGRAPLVLVTYFTQPQQNAESRRDVLASAARIIAEGL
>Blac_CTX-M-2
MMTQSIRRSMLTVMATLPLLFSSATLHAQANSVQQQLEALEKSSGGRLGVALINTADNSQILYRADERFAMCSTSKVMAA
AAVLKQSESDKHLLNQRVEIKKSDLVNYNPIAEKHVNGTMTLAELGAAALQYSDNTAMNKLIAHLGGPDKVTAFARSLGD
ETFRLDRTEPTLNTAIPGDPRDTTTPLAMAQTLKNLTLGKALAETQRAQLVTWLKGNTTGSASIRAGLPKSWVVGDKTGS
GDYGTTNDIAVIWPENHAPLVLVTYFTQPEQKAESRRDILAAAAKIVTHGF
>Blac_CTX-M-1
MVKKSLRQFTLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRLGVALINTADNSQILYRADERFAMCSTSKVMAV
AAVLKKSESEPNLLNQRVEIKKSDLVNYNPIAEKHVDGTMSLAELSAAALQYSDNVAMNKLISHVGGPASVTAFARQLGD
ETFRLDRTEPTLNTAIPGDPRDTTSPRAMAQTLRNLTLGKALGDSQRAQLVTWMKGNTTGAASIQAGLPASWVVGDKTGS
GDYGTTNDIAVIWPKDRAPLILVTYFTQPQPKAESRRDVLASAAKIVTNGL
>Blac_CTX-M-6
MMTQSIRRSMLTVMATLPLLFSSATLHAQANSVQQQLEALEKSSGGRLGVALINTADNSQILYVADERFAMCSTSKVMAA
AAVLKQSESDKHLLNQRVEIRASDLVNYNPIAEKHVNGTMTLAQLGAGALQYSDNTAMNKLIAHLGGPDKVTAFARSLGD
ETFRLDRTEPTLNSAIPGDPRDTTTPLAMAQTLKNLTLGKALAETQRAQLVTWLKGNTTGSASIRAGLPKSWGVGDKTGS
GDYGTTNDIAVIWPENHAPLVLVTYFTQPEQKAESRRDVLAAAAKIVTHGF
>Blac_OXY-2
MIKSSWRKIAMLAAAVPLLLASGALWASTDAIHQKLTDLEKRSGGRLGVALINTADNSQILYRGDERFAMCSTSKVMAAA
AVLKQSESNKEVVNKRLEINAADLVVWSPITEKHLQSGMTLAELSAATLQYSDNTAMNLIIGYLGGPEKVTAFARSIGDA
TFRLDRTEPTLNTAIPGDERDTSTPLAMAESLRKLTLGDALGEQQRAQLVTWLKGNTTGGQSIRAGLPESWVVGDKTGAG
DYGTTNDIAVIWPEDHAPLVLVTYFTQPQQDAKNRKEVLAAAAKIVTEGL
```

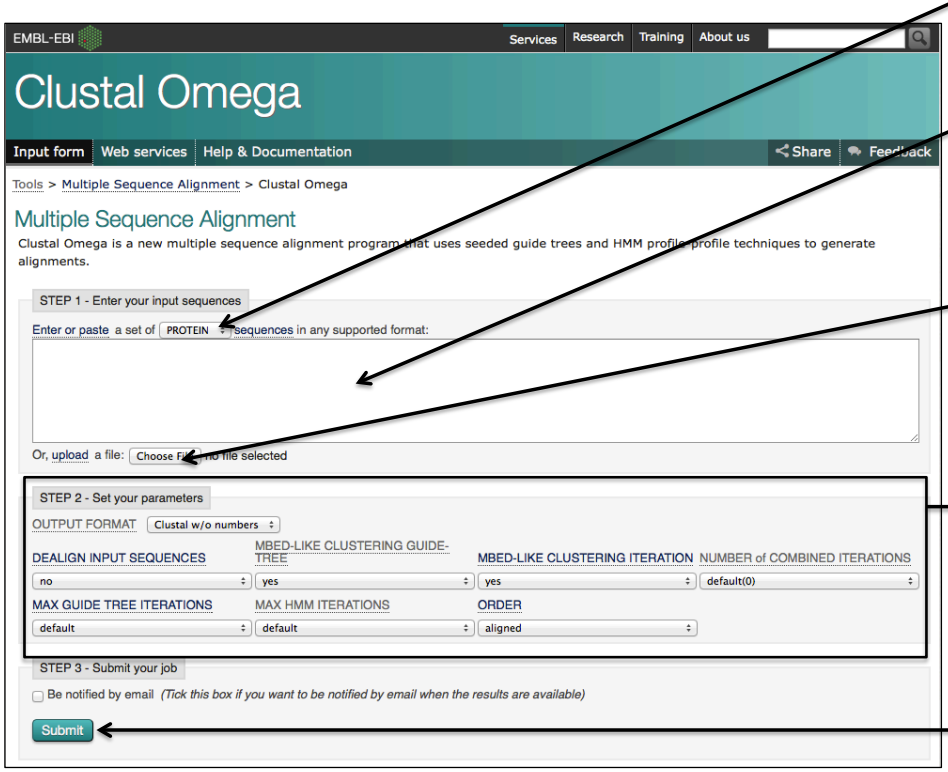
Izračun prileganja

Na voljo je več orodij, ki uporabljajo različne algoritme:

- **Clustal Omega** - novo orodje, primerno predvsem za (srednje) veliko število prileganih zaporedij
- **Clustal W2** - klasično orodje za prileganje večih zaporedij
- **DbClustal** - različica programa Clustal, ki omogoča pripravo prileganja večih zaporedij, pridobljenih preko iskanja s programom BLAST - kot vhodne podatke moramo podati naše iskalno zaporedje ter celoten rezultat iskanja z BLAST, ki ga opravimo posebej!
- **Kalign** - zelo hiter program za prileganje večih zaporedij, primeren predvsem za veliko število prileganih zaporedij
- **MAFFT** - program za prileganje večih zaporedij, ki uporablja hitro Fourierjevo transformacijo (FFT)
- **MUSCLE** - zelo natančno orodje, primerno predvsem za prileganje srednjega števila aminokislinskih zaporedij
- **MView** - orodje za preoblikovanje prikaza prileganja večih zaporedij (npr. barve) oz. spremembo formata
- **T-Coffee** - primerno predvsem za prileganje manjšega števila zaporedij
- **WebPRANK** - orodje, ki za umeščanje vrzeli upošteva evolucijske podatke

Izračun prileganja: Clustal Omega

http://www.ebi.ac.uk/Tools/msa/clustalo/



nastavimo tip zaporedij, ki jih prilegamo (ak ali nt)

polje za vnos zaporedij (format FASTA)

zaporedja, ki jih prilegamo, lahko prenesemo na strežnik tudi kot tekstno datoteko

možnosti za prileganje; do razlage pridemo s klikom na ime možnosti

s klikom na "Submit" pošljemo zahtevo na strežnik in počakamo...

Izračun prileganja: Clustal Omega

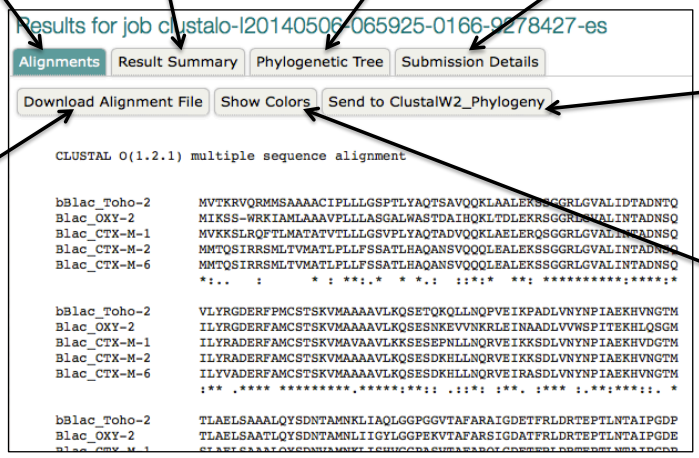
Rezultati prileganja

povzetek rezultatov – povezave do datotek:
- vhodni podatki
- prileganje
- podatki za izris filogenetskega drevesa
- matrika z odstotki identičnosti

podatki za izris filogenetskega drevesa ter enostavno filogenetsko drevo

vhodni parametri za prileganje, verzija programa; izpisan je tudi ekvivaleten ukaz za delo v ukazni vrstici

prikaz prileganja (privzeto se najprej prikaže le-to)



rezultate lahko direktno pošljemo v program ClustalW2_Phylogeny za izris filogenetskega drevesa

prenos prileganja v tekstni datoteki

vklop barvnega prikaza (**kisle**, **bazične**, **polarne**, **nepolarne**)

Izračun prileganja: Clustal Omega

Rezultati prileganja – podatki za izris enostavnega filogenetskega drevesa

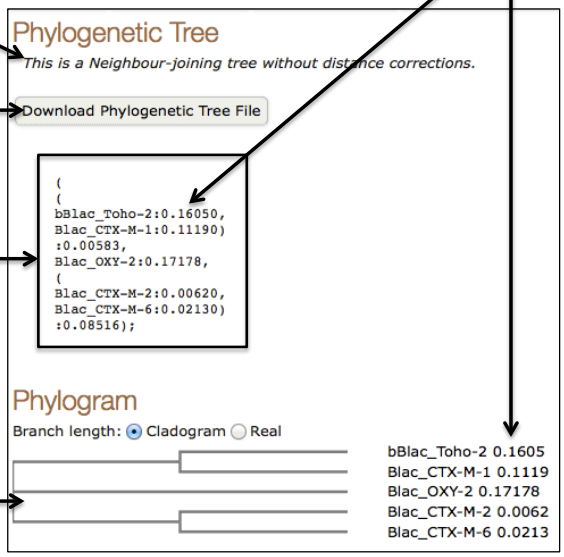
"without distance corrections" – program ne upošteva večih subtitucij na istem mestu v zaporedjih (predvsem pomembno pri bolj oddaljenih zaporedjih)

te številke so "razdalje" in so povezane s podobnostjo

način priprave drevesa

prenos tekstne datoteke s podatki za izris filogenetskega drevesa

podatki za izris filogenetskega drevesa



filogenetsko drevo na osnovi zgornjih podatkov

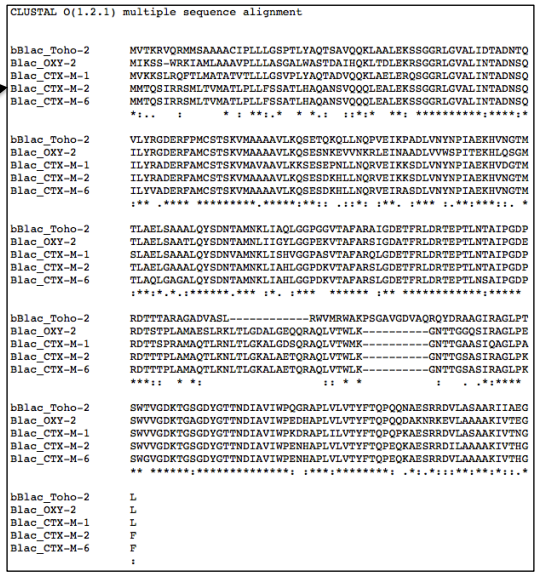
Priprava filogenetskega drevesa: ClustalW2 Phylogeny

ClustalW2 Phylogeny omogoča nastavitev parametrov za izris drevesa, ki jih ni na voljo med rezultati prileganja.

Podatke lahko v ClustalW2 Phylogeny vnesemo na več načinov:

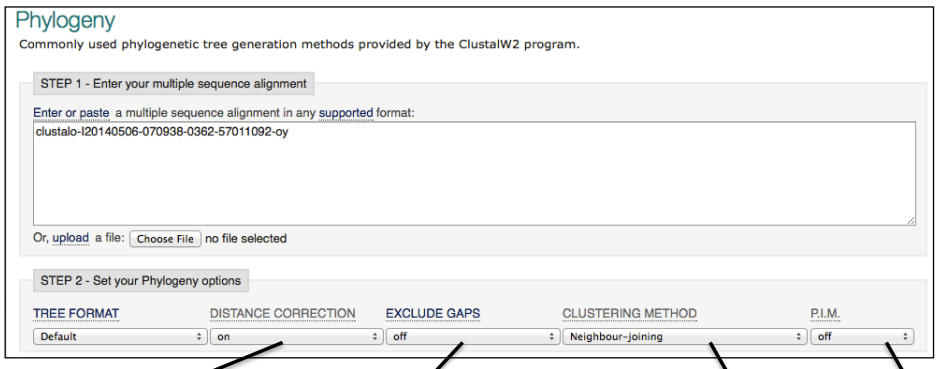
- direktno s klikom na 

pri rezultatih prileganja
- rezultat prileganja (enaka vsebina kot je v *Alignment File*) prilepimo v ustrezno okno programa ClustalW2 Phylogeny
- v ustrezno okno programa ClustalW2 Phylogeny prilepimo kodo prileganja, ki smo ga pripravili s Clustal (v obliki **`clustalo-I20140506-070938-0362-57011092-oy`**)



```
CLUSTAL O(1.2.1) multiple sequence alignment


bBlac_Toho-2       MVTKRVQRMMSAAAACIPLLLGSPTLYAQTSAVQQKLAALEKSSGGRLGVALIDTADNTQ
Blac_OXY-2         MIKSS-WRKIAMLAAAVPLLLASGALWASTDAIHQKLTDLEKRSGGRLGVALINTADNSQ
Blac_CTX-M-1       MVKKSLRQFTLMATATVTLLLGSVPLYAQTADVQQKLAELERQSGGRLGVALINTADNSQ
Blac_CTX-M-2       MMTQSIRRSMLTVMATLPLLFSSATLHAQANSVQQQLEALEKSSGGRLGVALINTADNSQ
Blac_CTX-M-6       MMTQSIRRSMLTVMATLPLLFSSATLHAQANSVQQQLEALEKSSGGRLGVALINTADNSQ
                   *:..   :      * : **:.*  * *.:  ::*:*  **: **********:****:*

bBlac_Toho-2       VLYRGDERFPMCSTSKVMAAAAVLKQSETQKQLLNQPVEIKPADLVNYNPIAEKHVNGTM
Blac_OXY-2         ILYRGDERFAMCSTSKVMAAAAVLKQSESNKEVVNKRLEINAADLVVWSPITEKHLQSGM
Blac_CTX-M-1       ILYRADERFAMCSTSKVMAVAAVLKKSESEPNLLNQRVEIKKSDLVNYNPIAEKHVDGTM
Blac_CTX-M-2       ILYRADERFAMCSTSKVMAAAAVLKQSESDKHLLNQRVEIKKSDLVNYNPIAEKHVNGTM
Blac_CTX-M-6       ILYVADERFAMCSTSKVMAAAAVLKQSESDKHLLNQRVEIRASDLVNYNPIAEKHVNGTM
                   :** .**** *********.*****:**:: .::*: :**. :*** :.**:***::. *

bBlac_Toho-2       TLAELSAAALQYSDNTAMNKLIAQLGGPGGVTAFARAIGDETFRLDRTEPTLNTAIPGDP
Blac_OXY-2         TLAELSAATLQYSDNTAMNLIIGYLGGPEKVTAFARSIGDATFRLDRTEPTLNTAIPGDE
Blac_CTX-M-1       SLAELSAAALQYSDNVAMNKLISHVGGPASVTAFARQLGDETFRLDRTEPTLNTAIPGDP
Blac_CTX-M-2       TLAELGAAALQYSDNTAMNKLIAHLGGPDKVTAFARSLGDETFRLDRTEPTLNTAIPGDP
Blac_CTX-M-6       TLAQLGAGALQYSDNTAMNKLIAHLGGPDKVTAFARSLGDETFRLDRTEPTLNSAIPGDP
                   :**:*.*.:******.*** :*. :***  ****** :** ************:*****

bBlac_Toho-2       RDTTTARAGADVASL------------RWVMRWAKPSGAVGDVAQRQYDRAAGIRAGLPT
Blac_OXY-2         RDTSTPLAMAESLRKLTLGDALGEQQRAQLVTWLK----------GNTTGGQSIRAGLPE
Blac_CTX-M-1       RDTTSPRAMAQTLRNLTLGKALGDSQRAQLVTWMK----------GNTTGAASIQAGLPA
Blac_CTX-M-2       RDTTTPLAMAQTLKNLTLGKALAETQRAQLVTWLK----------GNTTGSASIRAGLPK
Blac_CTX-M-6       RDTTTPLAMAQTLKNLTLGKALAETQRAQLVTWLK----------GNTTGSASIRAGLPK
                   ***::  * *:                  :: * *          :   . .*:****

bBlac_Toho-2       SWTVGDKTGSGDYGTTNDIAVIWPQGRAPLVLVTYFTQPQQNAESRRDVLASAARIIAEG
Blac_OXY-2         SWVVGDKTGAGDYGTTNDIAVIWPEDHAPLVLVTYFTQPQQDAKNRKEVLAAAAKIVTEG
Blac_CTX-M-1       SWVVGDKTGSGDYGTTNDIAVIWPKDRAPLILVTYFTQPQPKAESRRDVLASAAKIVTNG
Blac_CTX-M-2       SWVVGDKTGSGDYGTTNDIAVIWPENHAPLVLVTYFTQPEQKAESRRDILAAAAKIVTHG
Blac_CTX-M-6       SWGVGDKTGSGDYGTTNDIAVIWPENHAPLVLVTYFTQPEQKAESRRDVLAAAAKIVTHG
                   ** ******:**************: :***:********: .*:.*:::**:**:*::.*

bBlac_Toho-2       L
Blac_OXY-2         L
Blac_CTX-M-1       L
Blac_CTX-M-2       F
Blac_CTX-M-6       F
                   :
```


Priprava filogenetskega drevesa: ClustalW2 Phylogeny



vklop/izklop korekcije zaradi večih zaporednih substitucij na istem mestu v zaporedjih

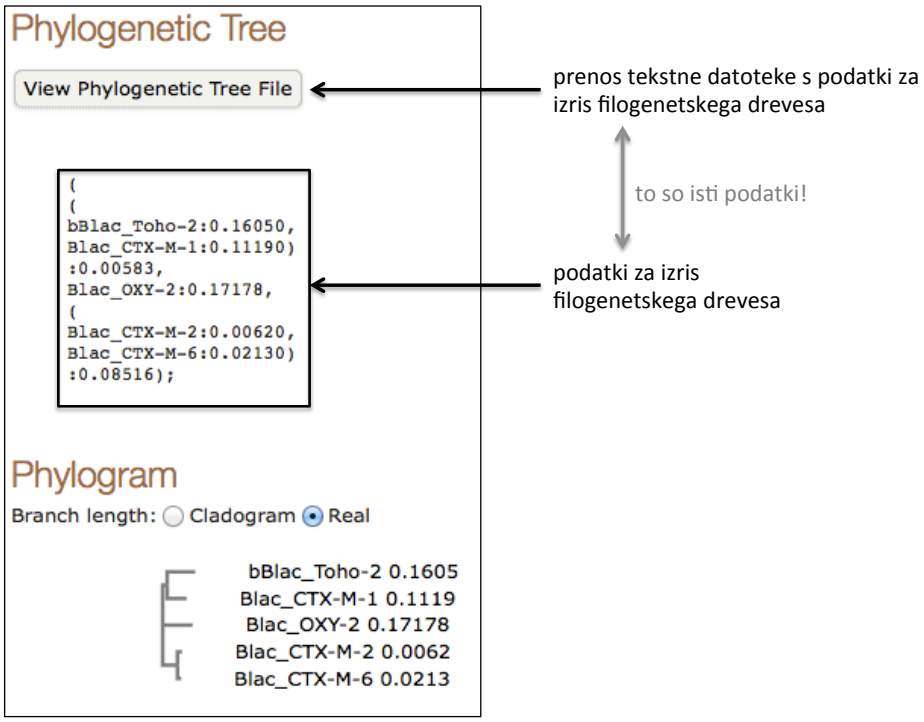
vklop/izklop (ne)upoštevanja vrzeli (če je to vklopljeno, potem program upošteva samo tiste dele prileganja, kjer imamo podatke z vvseh zaporedij)

metoda združevanja

vklop "percent identity matrix"

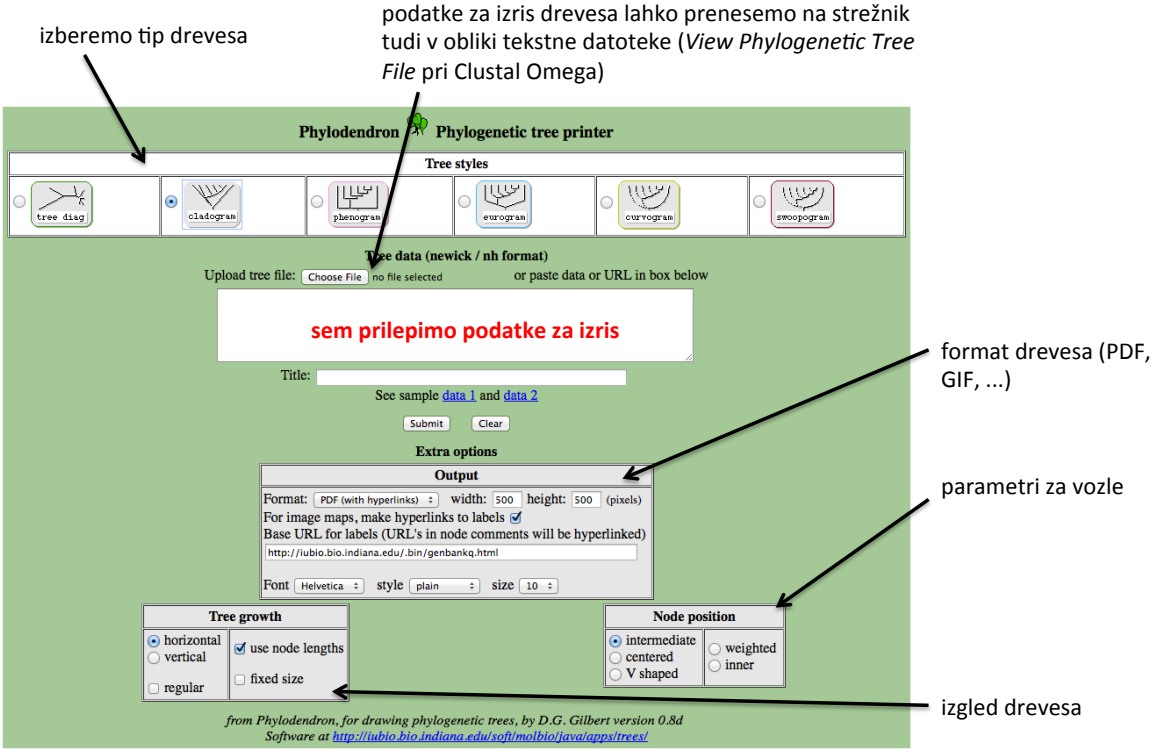
```
#
#
#  Percent Identity  Matrix - created by Clustal2.1
#
#

     1: bBlac_Toho-2  100.00   66.55   72.76   73.84   72.40
     2: Blac_OXY-2     66.55  100.00   70.69   73.45   72.41
     3: Blac_CTX-M-1   72.76   70.69  100.00   79.73   77.66
     4: Blac_CTX-M-2   73.84   73.45   79.73  100.00   97.25
     5: Blac_CTX-M-6   72.40   72.41   77.66   97.25  100.00
```

Priprava filogenetskega drevesa: ClustalW2 Phylogeny

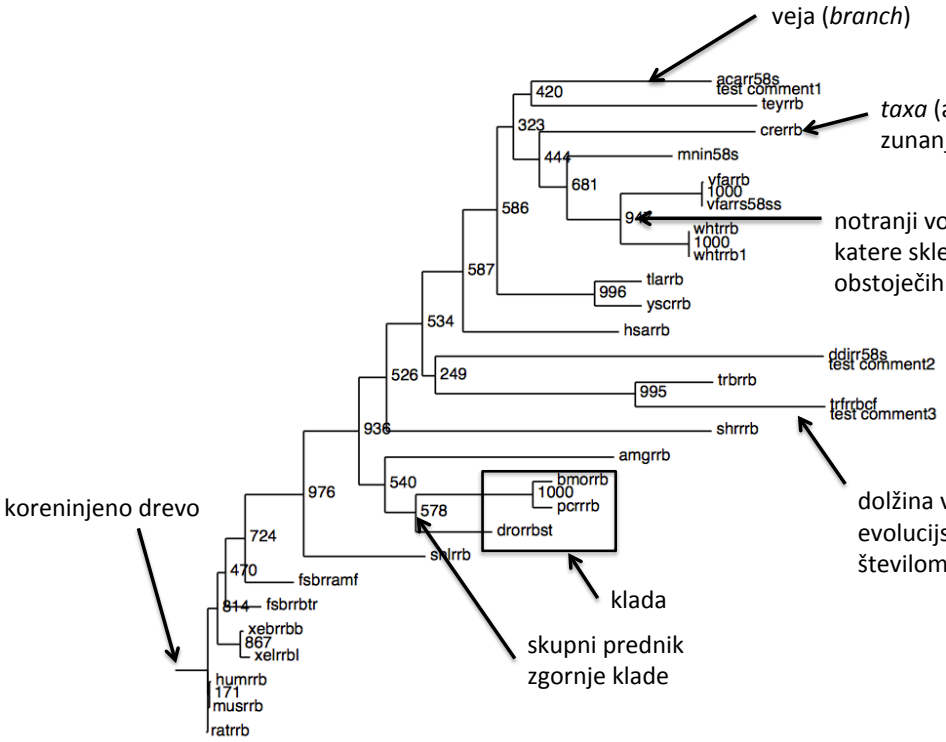


Priprava filogenetskega drevesa: Phylodendron

http://iubio.bio.indiana.edu/treeapp/treeprint-form.html



Nekaj malega o filogenetskih drevesih...



Kladogram (**cladogram**) – prikazuje le razvejitve (topologijo)
Filogram (**phylogram**) – prikazuje razvejitve in razdalje