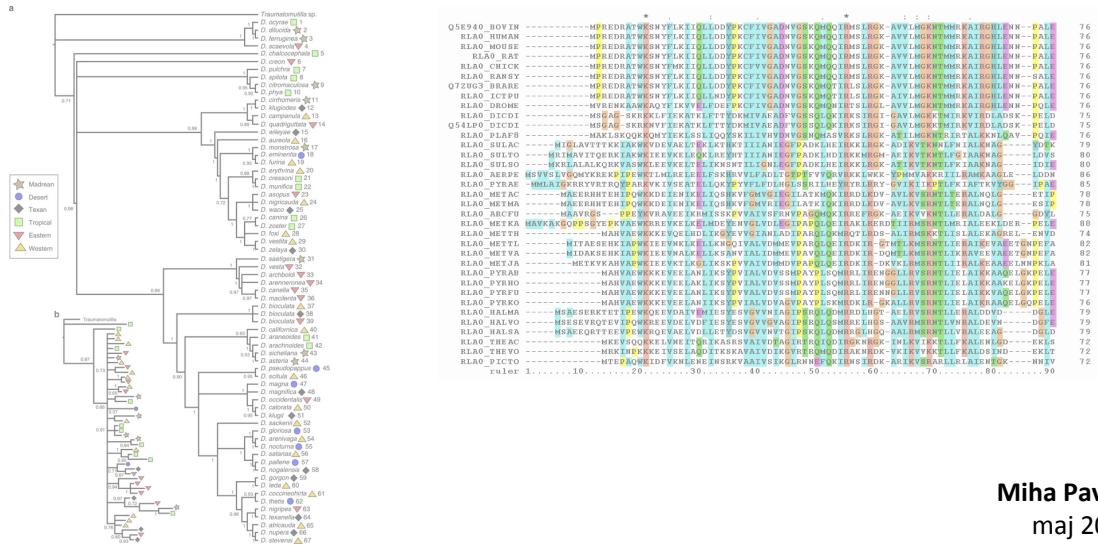


## 9. vaja

# Prileganje zaporedij in filogenetska drevesa



Miha Pavšič  
maj 2014

9. vaja: Prileganje zaporedij in filogenetska drevesa

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

- razumevanje bioloških procesov
- ugotavljanje 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:

The screenshot shows the BLAST search results interface. The table lists sequences with columns for Description, Max score, Total score, Query cover, E value, Ident, and Accession. Several sequences are checked, including those with accessions O69395.1, P74841.1, E1ANH6.1, Q47066.1, P28585.2, O65975.1, O65976.1, and Q33807.1. A 'Download' button is visible, and a dropdown menu is open, showing options like FASTA (complete sequence), FASTA (aligned sequences), GenBank (complete sequence), Hit Table (text), Hit Table (CSV), Text, XML, and ASN.1. The 'FASTA (complete sequence)' option is selected.

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input checked="" type="checkbox"/> RecName: Full=Beta-lactamase Toho-2; Flags: Precursor	501	501	100%	9e-179	87%	O69395.1
<input checked="" type="checkbox"/> RecName: Full=Beta-lactamase CTX-M-2; AltName: Full=Cefotaximase 2; Flags: Precursor	488	488	100%	2e-173	80%	P74841.1
<input type="checkbox"/> RecName: Full=Beta-lactamase CTX-M-97; Flags: Precursor	486	486	100%	8e-173	80%	E1ANH6.1
<input type="checkbox"/> RecName: Full=Beta-lactamase Toho-1; Flags: Precursor	486	486	100%	2e-172	80%	Q47066.1
<input checked="" type="checkbox"/> RecName: Full=Beta-lactamase CTX-M-1; AltName: Full=Beta-lactamase MEN-1; AltName: Full=Cefotaximase 1; Flags: Precursor	481	481	100%	2e-170	79%	P28585.2
<input type="checkbox"/> RecName: Full=Beta-lactamase CTX-M-5; AltName: Full=Cefotaximase 5; Flags: Precursor	480	480	100%	3e-170	79%	O65975.1
<input checked="" type="checkbox"/> RecName: Full=Beta-lactamase CTX-M-6; AltName: Full=Cefotaximase 6; Flags: Precursor	478	478	100%	2e-169	79%	O65976.1
<input checked="" type="checkbox"/> RecName: Full=Beta-lactamase CTX-M-4; AltName: Full=Cefotaximase 4; Flags: Precursor	476	476	100%	1e-168	78%	Q33807.1
<input type="checkbox"/> RecName: Full=Beta-lactamase; AltName: Full=Penicillinase; Contains: RecName: Full=Beta-lactamase; Flags: Precursor	423	423	89%	3e-168	78%	O33807.1
<input checked="" type="checkbox"/> RecName: Full=Beta-lactamase OXY-2; AltName: Full=Penicillinase; Flags: Precursor	423	423	92%	1e-168	78%	O33807.1
<input type="checkbox"/> RecName: Full=Beta-lactamase OXY-1; AltName: Full=Penicillinase; Flags: Precursor	421	421	92%	7e-168	78%	O33807.1
<input checked="" type="checkbox"/> RecName: Full=Beta-lactamase; AltName: Full=Penicillinase; Flags: Precursor	412	412	98%	2e-168	78%	O33807.1
<input type="checkbox"/> RecName: Full=Beta-lactamase; AltName: Full=Cefuroximase; Flags: Precursor	364	364	95%	1e-168	78%	O33807.1
<input type="checkbox"/> RecName: Full=Beta-lactamase; AltName: Full=Penicillinase	357	357	89%	4e-168	78%	O33807.1

1. izberemo zaporedja (“najboljši” zadelek 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
MVTKRVRMMSAAAACIPLLLGSPTLYAQTSAVQQKLAALEKSSGGRLGVALIDTADNTQVLYRGDERFPMCSTSKVMAA
AAVLKQSETQKQLLNQPVETKPADLVNYPNPIAEKHVNGTMTLAELSAAALQYSDNTAMNKLIHQGGPGGVTAFAAIGD
ETFRLDRTEPTLNTAIPGDRDTRARAGADVASLRWVMRWAKPSGAVGDVAQRQYDRAAGIRAGLPTSWTVGDKTGSGD
YGTNTNDIAVIWPQGRAPLVLVVYFTQPPQNAESRRDVLASAARIIEEGL
>Blac_CTX-M-2
MMTQSIIRRSMLTVMATLPLLFSSATLHAQANSVQQQLEALEKSSGGRLGVALINTADNSQILYRADERFAMCSTSKVMAA
AAVLKQSESDKHLNQRVEIKKSDLVNYNPIAEKHVNGTMTLAEELGAAALQYSDNTAMNKLIHLLGGPDKVTAFARSLGD
ETFRLDRTEPTLNTAIPGDRDTRPLAMAQTLKNTLTKALAEQRAQLVTWLGKNTTGSASIRAGLPKSWVVGDKTGS
GDYGTNTNDIAVIWPENHAPLVLVVYFTQPEQKAESRRDILAAAIAKIVTHGF
>Blac_CTX-M-1
MVKSLRQFTLMATATVTLGSLVPLAQTADVQQKLAELERQSGGRLGVALINTADNSQILYRADERFAMCSTSKVMAV
AAVLKQSESEPNLLNQRVEIKKSDLVNYNPIAEKHVNGTMTLAEELGAAALQYSDNVAMNKLIHLLGGPDKVTAFARSLGD
ETFRLDRTEPTLNTAIPGDRDTRPRAMAQTLRNLTKALGDSQRAQLVTWLMKNTTGAASIQAGLPASVWVGDKTGS
GDYGTNTNDIAVIWPDKRAPLILVYFTQPPQKAESRRDVLASAARIIVTNGL
>Blac_CTX-M-6
MMTQSIIRRSMLTVMATLPLLFSSATLHAQANSVQQQLEALEKSSGGRLGVALINTADNSQILYVADERFAMCSTSKVMAA
AAVLKQSESDKHLNQRVEIRASDLVNYNPIAEKHVNGTMTLAQLGAGALQYSDNTAMNKLIHLLGGPDKVTAFARSLGD
ETFRLDRTEPTLNTAIPGDRDTRPLAMAQTLKNTLTKALAEQRAQLVTWLGKNTTGSASIRAGLPKSWVVGDKTGS
GDYGTNTNDIAVIWPENHAPLVLVVYFTQPEQKAESRRDVLAAAIAKIVTHGF
>Blac_OXY-2
MIKSWRKIAMLAAAVPLLLASGALWASTDAIHQKLTDLKRSGGRLGVALINTADNSQILYRGDERFAMCSTSKVMAAA
AVLKQSESNKEVVNKRLEINAADLVVWSPITEKHLQSGMTLAELSAATLQYSDNTAMNLIIGYLLGGPEKVTAFARSIGDA
TFRLDRTEPTLNTAIPGDRDTRPLAMAESLRKLTLDALGEGQRAQLVTWLGKNTTGGQSIRAGLPESVWVGDKTGAG
DYGTNTNDIAVIWPEHAPLVLVVYFTQPPQDAKNRKEVLAAAIAKIVTEGL
```

## 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/>

The screenshot shows the Clustal Omega web interface. It is divided into three main steps:

- STEP 1 - Enter your input sequences:** This section contains a text area for pasting sequences and a file upload button. An arrow points to the text area with the label "polje za vnos zaporedij (format FASTA)".
- STEP 2 - Set your parameters:** This section contains several dropdown menus for configuring the alignment process. An arrow points to the "OUTPUT FORMAT" dropdown with the label "nastavimo tip zaporedij, ki jih prilegamo (ak ali nt)". Another arrow points to the "DEALIGN INPUT SEQUENCES" dropdown with the label "zaporedja, ki jih prilegamo, lahko prenesemo na strežnik tudi kot tekstno datoteko".
- STEP 3 - Submit your job:** This section contains a "Submit" button. An arrow points to this button with the label "s klikom na 'Submit' pošljemo zahtevo na strežnik in počakamo...".

Additional annotations on the right side of the image include:

- "možnosti za prileganje; do razlage pridemo s klikom na ime možnosti" (options for alignment; we get the explanation by clicking on the name of the option) pointing to the parameter dropdowns.

## Izračun prileganja: Clustal Omega

### Rezultati prileganja

- pozveteq 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 ekvivalenten ukaz za delo v ukazni vrstici

prikaz prileganja (privzeto se najprej prikaže le-to)

prenos prileganja v tekstni datoteki

Results for job clustalo-I20140506-065925-0166-9278427-es

Alignments Result Summary Phylogenetic Tree Submission Details

Download Alignment File Show Colors Send to ClustalW2\_Phylogeny

CLUSTAL O(1.2.1) multiple sequence alignment

```

bBlac_ToHo-2      MVTKRVQRMSAAAACIFLLLSGFTLYAQTSAVQKRLAALERSGGRLGVALIDTADNTQ
Blac_OXY-2        MKSS-WRKIAMLAARVFLLLASGALWASTDAIHQKRLDLEKRSGGRLGVALINTADNSQ
Blac_CTX-M-1      MVKSLRQFTLMATAFVFLLLGVSFLYAQTADVQKRLAELERQSGGRLGVALINTADNSQ
Blac_CTX-M-2      NMTQSIIRRSMLTVMATLPLFFSATLHAQANSVQQOLEALEKSSGGRLGVALINTADNSQ
Blac_CTX-M-6      NMTQSIIRRSMLTVMATLPLFFSATLHAQANSVQQOLEALEKSSGGRLGVALINTADNSQ
*..  :          * : **:* * *.: :*: * ** : *****:****:*

bBlac_ToHo-2      VLYRGDERFPMCSTSKVMAAAVLKQSETKQKQLNQPVEIKPADLVNYPNIAEKHVNGTM
Blac_OXY-2        ILYRGDERFAMCSTSKVMAAAVLKQSESNKEVNVNKRLEINAADLVNVPNITEKHLQSGM
Blac_CTX-M-1      ILYRADERFAMCSTSKVMAAAVLKQSESNKLNORVEIKKSDLVNYPNIAEKHVNGTM
Blac_CTX-M-2      ILYRADERFAMCSTSKVMAAAVLKQSESDKHLNORVEIKKSDLVNYPNIAEKHVNGTM
Blac_CTX-M-6      ILYVADERFAMCSTSKVMAAAVLKQSESDKHLNORVEIRASDLVNYPNIAEKHVNGTM
:* ..** * ..** * ..** * ..** * ..** * ..** * ..** * ..** * ..** *

```

bBlac\_ToHo-2 TLAELSAALQYSDNTAMNKLIAQLGGPGGVAFARAIIGDTEFRDLRTEPTLNTAIPGDP
Blac\_OXY-2 TLAELSAATLQYSDNTAMNLIIGYLGPEKVTAFARSIGDTEFRDLRTEPTLNTAIPGDE
Blac\_CTX-M-1 TLAELSAALQYSDNTAMNKLIAQLGGPGGVAFARAIIGDTEFRDLRTEPTLNTAIPGDP
Blac\_CTX-M-2 TLAELSAATLQYSDNTAMNLIIGYLGPEKVTAFARSIGDTEFRDLRTEPTLNTAIPGDE
Blac\_CTX-M-6 TLAELSAALQYSDNTAMNKLIAQLGGPGGVAFARAIIGDTEFRDLRTEPTLNTAIPGDP

rezultate lahko direktno pošljemo v program ClustalW2\_Phylogeny za izris filogenetskega drevesa

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 substitucij 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

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Download Phylogenetic Tree File

```

(
  (
    bBlac_ToHo-2:0.16050,
    Blac_CTX-M-1:0.11190)
  :0.00583,
  Blac_OXY-2:0.17178,
  (
    Blac_CTX-M-2:0.00620,
    Blac_CTX-M-6:0.02130)
  :0.08516);

```

Phylogram

Branch length:  Cladogram  Real

bBlac\_ToHo-2 0.1605  
Blac\_CTX-M-1 0.1119  
Blac\_OXY-2 0.17178  
Blac\_CTX-M-2 0.0062  
Blac\_CTX-M-6 0.0213



## Priprava filogenetskega drevesa: ClustalW2 Phylogeny


### Phylogenetic Tree

View Phylogenetic Tree File

```
(
  (
    bBlac_Toho-2:0.16050,
    Blac_CTX-M-1:0.11190)
  :0.00583,
  Blac_OXY-2:0.17178,
  (
    Blac_CTX-M-2:0.00620,
    Blac_CTX-M-6:0.02130)
  :0.08516);
```

### Phylogram

Branch length:  Cladogram  Real



```
bBlac_Toho-2 0.1605
Blac_CTX-M-1 0.1119
Blac_OXY-2 0.17178
Blac_CTX-M-2 0.0062
Blac_CTX-M-6 0.0213
```

prenos tekstne datoteke s podatki za izris filogenetskega drevesa

↕ to so isti podatki!

podatki za izris filogenetskega drevesa

## Priprava filogenetskega drevesa: Phylodendron


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

izberemo tip drevesa

podatke za izris drevesa lahko prenesemo na strežnik tudi v obliki tekstne datoteke (View Phylogenetic Tree File pri Clustal Omega)

### Phylodendron Phylogenetic tree printer

Tree styles



Upload tree file:  no file selected or paste data or URL in box below

sem prilepimo podatke za izris

Title:

See sample [data 1](#) and [data 2](#)

Extra options

Output

Format:  width:  height:  (pixels)

For image maps, make hyperlinks to labels

Base URL for labels (URL's in node comments will be hyperlinked)

Font:  style:  size:

Tree growth

horizontal  vertical

use node lengths  fixed size

regular

Node position

intermediate  centered  V shaped

weighted  inner

format drevesa (PDF, GIF, ...)

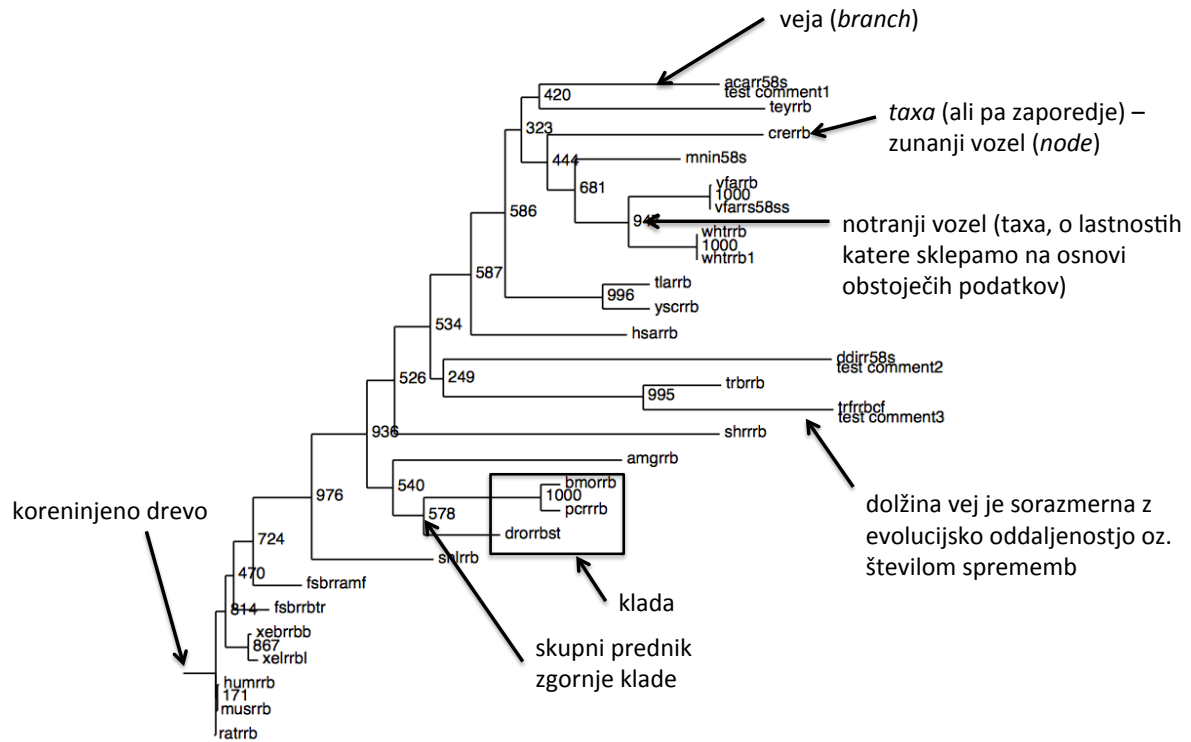
parametri za vozle

izgled drevesa

from Phylodendron, for drawing phylogenetic trees, by D.G. Gilbert version 0.8d  
Software at <http://iubio.bio.indiana.edu/soft/molbio/java/apps/trees/>



## Nekaj malega o filogenetskih drevesih...



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