Základy praktické bioinformatiky

Téma 4/10

Proteinová bioinformatika III

Cíle:

Student bude schopen porovnat sekvence dvou či více proteinů, určit míru identity, zobrazit jednoduchý fylogenetický strom vybraných proteinů a najít 3D strukturu proteinu.

Porovnávání sekvencí

Porovnávání sekvencí spočívá v nalezení nejlepšího skóre při jejich zarovnání. Skóre vychází z algoritmů obsažených „uvnitř“ programů, a je dáno tím jaké aminokyseliny jsou v zarovnání pod sebou (hodnocení dle substituční matice BLOSUM62, PAM…) a kolik a jak dlouhých mezer se ve výsledném zarovnání nachází.

Párové porovnání

Při párovém porovnávání porovnáváme dvě sekvence. Rozlišujeme porovnání lokální (Laling) nebo globální (Needle).

Globální porovnání

(https://www.ebi.ac.uk/Tools/psa/emboss_needle/)

Zde jsou sekvence porovnány globálně. Všimněte si, že jsou porovnané v celé délce, což zahrnuje 342 pozic (aminokyselin), včetně přesahujícího N-konce první sekvence. Identita obou sekvencí je 40,4%.

```
#=======================================
#
# Aligned_sequences: 2
# 1: AAH09679.1
# 2: AAI47025.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 342
# Identity:     138/342 (40.4%)
# Similarity:   196/342 (57.3%)
# Gaps:          51/342 (14.9%)
# Score: 646.5
#
#
#=======================================

AAH09679.1         1 MVSPATRKSLPKVKAMDFITSTAILPLLFGCLGVFGLFRLLQWVR---GK     47
                                          .:..:||||.  ||:.||..:.|.|.   .|
AAI47025.1         1 -------------------MAMLMLPLLL--LGISGLLFIYQEVSRLWSK     29

AAH09679.1        48 AYLRNAVVVITGATSGLGKECAKVFYAAGAKLVLCGRNGGALEELIRELT     97
                     :.::|.|||||.|.|||||||:||:..||:|||||:|...||.|...|
AAI47025.1        30 SAVQNKVVVITDAISGLGKECARVFHTGGARLVLCGKNWERLENLYDAL-     78

AAH09679.1        98 ASHATKVQTHKPYLVTFDLTDSGAIVAAAAEILQCFGYVDILVNNAGISY    147
                     .|.|...:|..|.||..||:|...:...|.|:|.|:|.||||:|||.:..
AAI47025.1        79 ISVADPSKTFTPKLVLLDLSDISCVPDVAKEVLDCYGCVDILINNASVKV    128

AAH09679.1       148 RGTIMDTTVDVDKRVMETNYFGPVALTKALLPSMIKRRQGHIVAISSIQG    197
                     :|.....::::||::|:.|||||:.|||||||:||.||.|.||.:::|||
AAI47025.1       129 KGPAHKISLELDKKIMDANYFGPITLTKALLPNMISRRTGQIVLVNNIQG    178

AAH09679.1       198 KMSIPFRSAYAASKHATQAFFDCLRAEMEQYEIEVTVISPGYIHTNLSVN    247
                     |..||||:.|||||||...|||||||:|:|:::.::.:||.:|.   |.:
AAI47025.1       179 KFGIPFRTTYAASKHAALGFFDCLRAEVEEYDVVISTVSPTFIR---SYH    225

AAH09679.1       248 AITADGS-------------RYGVMDTTTAQGRSPVEVAQDVLAAVGKKK    284
                     .....|:             .|||         .|||||::|:..|.:||
AAI47025.1       226 VYPEQGNWEASIWKFFFRKLTYGV---------HPVEVAEEVMRTVRRKK    266

AAH09679.1       285 KDVILADLLPSLAVYLRTLAPGLFFSLMASRARKERKSKNS-    325
                     ::|.:|:.:|..|||:||..|..||:::|...::.....
AAI47025.1       267 QEVFMANPIPKAAVYVRTFFPEFFFAVVACGVKEKLNVPEEG    308
```

Lokální porovnání

(https://www.ebi.ac.uk/Tools/psa/lalign/)

Zde jsou stejné sekvence porovnány lokálně. Toto porovnání nabízí další možnosti získání kratších porovnaných úseků (vhodné například pro nalezení repetitivních úseků v sekvenci). Všimněte si, že porovnávaný úsek je jen 300 aminokyselin, a identita v tomto úseku je 45%. Celkové skóre je 840. Druhé nejlepší skóre na obrázku je 34, porovnávaný úsek je jen 19 aminokyselin.

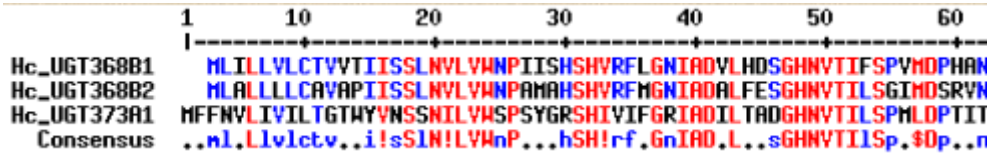
```
>>AAI47025.1 Dehydrogenase/reductase (SDR family) member  (308 aa)
 Waterman-Eggert score: 840; 253.1 bits; E(1) <  6.6e-72
45.0% identity (76.3% similar) in 300 aa overlap (24-319:5-301)

            30        40        50        60        70        80
AAH096 ILPLLFGCLGVFGLFRLLQWVR---GKAYLRNAVVVITGATSGLGKECAKVFYAAGAKLV
       .::::.  ::. ::. . : :    .:. ..: ::::: : :::::::.::...::.::
AAI470 MLPLLL--LGISGLLFIYQEVSRLWSKSAVQNKVVVITDAISGLGKECARVFHTGGARLV
           10          20        30        40        50        60

              90       100       110       120       130       140
AAH096 LCGRNGGALEELIRELTASHATKVQTHKPYLVTFDLTDSGAIVAAAAEILQCFGYVDILV
       :::.:   ::.:   :  : :   .:  : :: .::.: . .  .: :.:.:.: ::::.
AAI470 LCGKNWERLENLYDALI-SVADPSKTFTPKLVLLDLSDISCVPDVAKEVLDCYGCVDILI
            70        80        90       100       110       120

             150       160       170       180       190       200
AAH096 NNAGISYRGTIMDTTVDVDKRVMETNYFGPVALTKALLPSMIKRRQGHIVAISSIQGKMS
       :::... .:     ....::..:.::::::..::::::.::.:: :.:: ...:::::..
AAI470 NNASVKVKGPAHKISLELDKKIMDANYFGPITLTKALLPNMISRRTGQIVLVNNIQGKFG
            130       140       150       160       170       180

             210       220       230       240       250
AAH096 IPFRSAYAASKHATQAFFDCLRAEMEQYEIEVTVISPGYIHT-NLSVNAITADGSRYGVM
       :::::.:::::::. .:::::::::.:.:.. ....:: .:.. ..  .  . ..: .  .
AAI470 IPFRTTYAASKHAALGFFDCLRAEVEEYDVVISTVSPTFIRSYHVYPEQGNWEASIWKFF
            190       200       210       220       230       240

     260       270       280       290       300       310
AAH096 DTTTAQGRSPVEVAQDVLAAVGKKKKDVILADLLPSLAVYLRTLAPGLFFSLMASRARKE
          . :  :::::..:. .: .::..:.:. .:. :::.::. : .::...:  ....
AAI470 FRKLTYGVHPVEVAEEVMRTVRRKKQEVFMANPIPKAAVYVRTFFPEFFFAVVACGVKEK
            250       260       270       280       290       300

>--
 Waterman-Eggert score: 34; 14.4 bits; E(1) <  0.99
42.1% identity (63.2% similar) in 19 aa overlap (289-307:74-92)

      290       300
AAH096 LADLLPSLAVYLRTLAPGL
       : : : :.:   .:..: :
AAI470 LYDALISVADPSKTFTPKL
            80        90
```

Mnohonásobné porovnání (multiple alignment)

Umožňuje zarovnání více sekvencí najednou. Vstupní data pro programy představují jednotlivé sekvence ve fasta formátu, jejichž název (identifikátor) musí být odlišný. Výstupem těchto porovnání je většinou obrázek, který ukazuje všechny sekvence se zarovnanými pozicemi, které byly vyhodnoceny s nejlepším skóre. Zarovnání se u jednotlivých programů může trchu lišit, díky použitým maticím a výpočtovým parametrům.

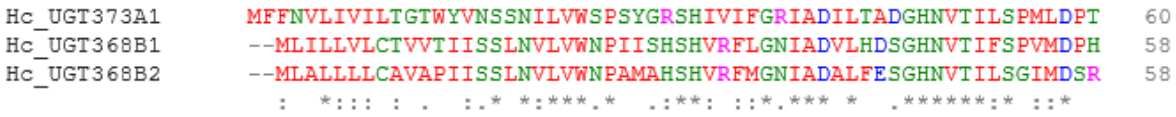
Multalin (http://multalin.toulouse.inra.fr/multalin/multalin.html)

Poskytuje graficky snadno pochopitelný výstup.



Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/)

Poskytuje grafický výstup, případně obsahuje možnost zobrazení ve formě fylogenetického stromu, kde jsou patrné příbuznosti sekvencí mezi sebou. Tento program také umožňuje stažení zarovnaného formátu sekvencí (Alignment file), který je možný převést pro úpravu do jiných programů (viz Boxshade)

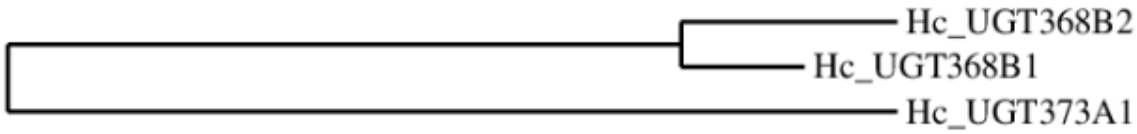


BoxShade

Tento program pomůže získat „pěkné“ zarovnání do publikace. Sám o sobě tento program sekvence neumí porovnat, ale z vloženého zarovnání vytvoří grafickou reprezentaci. Pro použití je nejprve nutné zadané sekvence porovnat jinde (např. Clustal Omega) a otevřít již zarovnané sekvence pomocí „Download alignment file“ a přenést celé zarovnání do okna v Boxshade. Na výběr je více typů výstupů. Pro klasické černobílé zobrazení pro publikace lze vybrat „RTF_new“.

Fylogenetické porovnání (http://www.phylogeny.fr/simple_phylogeny.cgi)

Neboli vytváření fylogenetických stromů, představuje složitější bioinformatický problém. Například fylogenetické porovnávání vybraných sekvencí ukazuje vzájemnou podobnost mezi nimi a vychází z mnohonásobného porovnání. Každá sekvence ve výsledném zobrazení představuje „konec větve“, dvě nejbližší větve představují dvě nejpodobnější si sekvence.



3D struktura proteinu (https://www.rcsb.org/)

3D struktury proteinu je možné získat z databáze PDB (Protein data bank). vyhledávat lze proteinovou zkratkou či přístupovým kódem proteinu nebo celým názvem. V databázi jsou k naleznutí veškeré struktury proteinů získané různými technikami, některé včetně ligandů, některé získané pouze z jednotlivých domén, apod. Struktury jsou k náhledu ve 3D s možností volně s nimi otáčet, upravovat náhled a typ zobrazení.