

Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2023/2024

4/10

„Protein bioinformatics III“

Retrieving protein sequences from databases (Uniprot: FASTA formate)

Computing amino-acids compositions, molecular weight, isoelectric point, and other parameters (SMS)

Prediction of proteases cutting (PeptideCutter)

Predicting elements of protein secondary structure, signal peptide, transmembrane helix

Finding 3-D structure

Finding all proteins that share a similar sequence

Finding evolutionary relationships between proteins, drawing proteins' family trees

Computing the optimal alignment between two or more protein sequences

...

Pairwise alignment

Global alignment – aligns full length sequence

Local alignment – aligns part of the sequences that fit best

(eg similar domains comparison, repetitive sequences...)

```
seq1  EARDF-NQYYSSIKRSGSIQ
      . : ..... . .
seq2  LPKLFIDQYYSSIKRTMG-H
```

global sequence alignment

```
seq1  NQYYSSIKRS
      . : .....
seq2  DQYYSSIKRT
```

local sequence alignment



V9PWX7	V9PWX7_SCHMA	1	MIESSTTIQVISAGLPRGTGKSLKNALEIIYHKPCYHMFEEIFNKQSDIIKWQNLIHDSH	60
V9PWX8	V9PWX8_SCHMA	1	MIESSTTIQVISAGLPRGTGKSLKNALEIIYHKP YHMFEEIFNKQSDIIKWQNLIHDSH	60
V9PWX7	V9PWX7_SCHMA	61	MITTPLLTTKRTIAIYDKLKEKLLDGYIATDLPCTCGFYKDLNMIYPNAKVLLTIRDKYDW	120
V9PWX8	V9PWX8_SCHMA	61	MITTPLLTTKRTIAIYDKLKEKLLDGYIATDLPCTCGFYKDLNMIYPNAKVLLTIRDKYDW	120
V9PWX7	V9PWX7_SCHMA	121	LHSLRKVVLPKSNDPWLKIEEGDKVGLNSDFYKLTEDSLKFAFQKDDLNFDDDDQVLE	180
V9PWX8	V9PWX8_SCHMA	121	LHSLRKVVLPKSNDPWLKIEEGDKVGLNSDFYKLTEDSLKFAFQKDDLNFDDDDQVLE	180
V9PWX7	V9PWX7_SCHMA	181	CYDEYNRLVQETVPSDRLLVLRIGDWEPLCKFLNVEIPNGIDYPCVNSHHQMTQLTEQL	240
V9PWX8	V9PWX8_SCHMA	181	CYDEYNRLVQETVPSDRLLVLRIGDWEPLCKFLNVEIPNGIDYPCVNSHHQMTQLTEQL	240
V9PWX7	V9PWX7_SCHMA	241	IKYKSLDAIHHMFPDLI	257
V9PWX8	V9PWX8_SCHMA	241	IKYKSLDAIHHMFPDLI	257

V9PWX7	V9PWX7_SCHMA	1	MIESSTTIQVISAGLPRGTGKSLKNALEIIYHKPCYHMFEEIFNKQSDIIKWQNLIHDSH	60
A0A183QDM9	A0A183QDM9_9TREM1		M ESS + VI AGLPRTGKSLKNALEIIYHKPCYHM EII + +DI KWQ L ++	60
V9PWX7	V9PWX7_SCHMA	61	MITTPLLTTKRTIAIYDKLKEKLLDGYIATDLPCTCGFYKDLNMIYPNAKVLLTIRDKYDW	120
A0A183QDM9	A0A183QDM9_9TREM1		KMEP-----TNELMINDGLKEIILMNYGAVTDVPACGFYKELMNIYPNAKVLLTIRDKYDW	115
V9PWX7	V9PWX7_SCHMA	121	LHSLRKVVLPKSNDPWLKIEEGDKV-----	146
A0A183QDM9	A0A183QDM9_9TREM16		LHSLRKVVLPKSNDPWLKIEEGDKVILTIRNKYDWLSFRQTLMPKSNDSNRITIDEAD	175
V9PWX7	V9PWX7_SCHMA	147	--LGLNSDFYKLTEDSLKFAFQKDDLNFDDDDQVLECYDEYNRLVQETVPSDRLLVLRIG	204
A0A183QDM9	A0A183QDM9_9TREM176		L L F K+ DS+K AF+K D + D+D +L+C+DEYNR V ETVPS+RLL+ +LG	235
V9PWX7	V9PWX7_SCHMA	205	DGWEPLCKFLNVEIPNGIDYPCVN	228
A0A183QDM9	A0A183QDM9_9TREM236		DGWEPLC+FLNV++P G+ YP +N	259

Pairwise alignment- Global



Job Dispatcher Help & Privacy Input form

Welcome to the new Job Dispatcher website. We'd love to hear your [feedback](#) about the new webpages!

EMBOSS Needle reads two input sequences and writes their optimal global sequence alignment to file.

Input sequence ⓘ

Sequence type

Protein DNA

Paste your sequence here - or use the example sequence

>Homo sapiens
 MVGRRALIVLAHSERTSFNYAMKEAAAAALKKGWEVWESDLYAMNFPVISRKDITGKLDKDPANFQYPAESVLAYKEGH
 LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRRKKAVALSITGGSGSMYSLQG
 IHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK
 KEVQDEKNKFGLSVGHHLGKSIPDNQIKARK

Zvolit soubor Nevybrán Zádný soubor

Paste your sequence here - or use the example sequence

>Sus scrofa
 MAVRKALILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFPVISRKDITGKLDKDPGNFQYPAETALAYKEGR
 LSPDIVAEQKKVEAADLVIFQFPLQWFGVPAILKGWFERVLIQEFAYTYAAMYDKGPFRRKKAVALSITGGSGSMYSLQG
 IHGDMNILLWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK
 KQVQDEQKSNKFGLSVGHHLGKSIPDNQVQKARK

Parameters

OUTPUT FORMAT ⓘ	MATRIX ⓘ	GAP OPEN ⓘ	GAP EXTEND ⓘ	END GAP ⓘ	END GAP OPEN ⓘ
pair	BLOSUM62	10	0.5	false	10

Less options ^

Submit

Title

Submit

```
#####
#
# Aligned_sequences: 2
# 1: Homo
# 2: Sus
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 274
# Identity: 245/274 (89.4%)
# Similarity: 260/274 (94.9%)
# Gaps: 0/274 ( 0.0%)
# Score: 1320.0
#
#
#####
```

Homo	1	MVGRRALIVLAHSERTSFNYAMKEAAAAALKKGWEVWESDLYAMNFPVISRKDITGKLDKDPANFQYPAESVLAYKEGH	50
Sus	1	MAVRKALILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFPVISRKDITGKLDKDPGNFQYPAETALAYKEGR	50
Homo	51	ISRKDITGKLDKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIF	100
Sus	51	ISRKDITGKLDKDPGNFQYPAETALAYKEGRSPDIVAEQKKVEAADLVIF	100
Homo	101	QFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRRKKAVALSITGGSGSMYSLQG	150
Sus	101	QFPLQWFGVPAILKGWFERVLIQEFAYTYAAMYDKGPFRRKKAVALSITGGSGSMYSLQG	150
Homo	151	GSMSYSLQGIHGD MNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK	200
Sus	151	GSMSYSLQGIHGD MNILLWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK	200
Homo	201	RIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEKNK	250
Sus	201	RIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKQVQDEQKSN	250
Homo	251	KFGLSVGHHLGKSIPDNQIKARK	274
Sus	251	KFGLSVGHHLGKSIPDNQVQKARK	274

Pairwise alignment- Local

EMBL-EBI home Services Research Training About us EMBL-EBI

Lalign

Pairwise Sequence Alignment (PSA)

Job Dispatcher Help & Privacy Input form

Welcome to the new Job Dispatcher website. We'd love to hear your [feedback](#) about the new webpages!

LALIGN finds internal duplications by calculating non-intersecting local alignments (

Input sequence ⓘ

Sequence type

Protein DNA

Paste your sequence here - or use the example sequence

[Zvolit soubor](#) [Nevybrán žádný soubor](#)

Paste your sequence here - or use the example sequence

Parameters More options ▼

Submit

Title

Lalign's job

Submit

Results for Job ID: lalign-I20240306-175812-0116-49168171-p1m

Tool Output **Result Files** Submission Details

```
# /fasta/bin/lalign36 -m 91 lalign-I20240306-175812-0116-49168171-p1m.asequence lalign-I20240306-175812-0116-49168171-p1m.bsequence -p -s BL50 -f -12 -g -2 -E 10.0 -m 0 -m "P11 lalign-I20240306-175812-0116-49168171-p1m.output.lav"
LALIGN finds non-overlapping local alignments
version 36.3.8h May, 2020
Please cite:
X. Huang and W. Miller (1991) Adv. Appl. Math. 12:373-381

Query: lalign-I20240306-175812-0116-49168171-p1m.asequence
1>>>Homo sapiens - 274 aa
Library: lalign-I20240306-175812-0116-49168171-p1m.bsequence
274 residues in 1 sequences

Statistics: (shuffled [500]) MLE statistics: Lambda= 0.1905; K=0.04079
statistics sampled from 1 (1) to 500 sequences
Threshold: E() < 10 score: 50
Algorithm: Smith-Waterman (1982, Michael Farrar 2004) (7.2 Nov 2010)
Parameters: BL50 matrix (15:-5), open/ext: -12/-2
Scan time: 0.010

The best non-identical alignments are: ls-w bits E(1) S_id S_sim aLen
Sus scrofa (274) 1659 468.9 5.34-157 0.594 0.978 274
++ 39 13.3 0.24 0.556 0.559 9
++ 36 14.5 0.26 0.400 0.525 25
++ 36 14.5 0.26 0.239 0.630 46
++ 35 14.2 0.26 0.269 0.654 26
++ 32 13.4 1 0.264 0.518 11
++ 31 13.1 1 0.261 0.696 23

>>>Homo, 274 aa vs lalign-I20240306-175812-0116-49168171-p1m.bsequence library

>>>Sus scrofa (274 aa)
MATCHES: E=6827 score: 1659; 468.9 bits; E(1) < 5.34-157
99.4% identity (97.5% similar) in 274 aa overlap (1-274:1-274)

10 20 30 40 50 60
Homo MVGRRALIVLHSEKTSFNAMKEAAALAKKGGHEVVEESLVANNPFIISKDIITGKL
Sus MAVRKALILAHSEKTSFNAMKEAAVEALKRGGHEVAVSOLVANNPFIISKDIITGKL

70 80 90 100 110 120
Homo KDRANFQYPAESVLAEGHLSPOIVAEQKLEAADLVZFPPLQWGFVPAILKQWPERV
Sus KDRGNFQYPAETALAYEGRLSPOIVAEQKVEAADLVZFPPLQWGFVPAILKQWPERV

130 140 150 160 170 180
Homo FIFGFAYTAAAYDKGPPRKKAVLSITGGSGSHYSLQIQHGDNNVILWPIQGIILHFC
Sus LIIGFAYTAAAYDKGPPRKKAVLSITGGSGSHYSLQIQHGDNNVILWPIQGIILHFC
```

uence [More example inputs](#)

Pairwise alignment- Local (visualization)

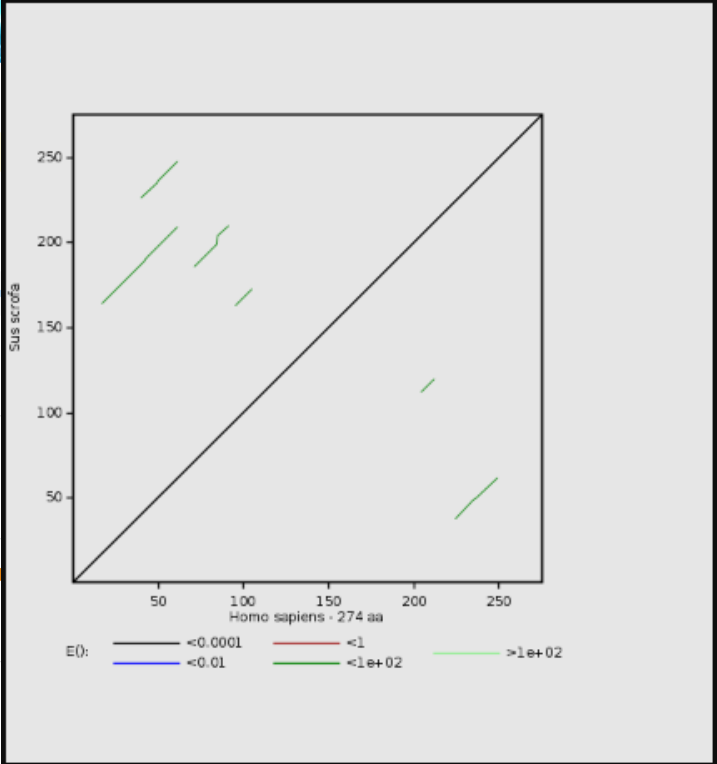
Lalign
Pairwise Sequence Alignment (PSA)

Job Dispatcher Help & Privacy **Input form**

Welcome to the new **Job Dispatcher** website. We'd love to hear your [feedback](#) about the new webpages!

Results for Job ID: lalign-I20240306-175812-0116-49168171-p1m

Tool Output	Result Files	Download
Tool Output	lalign-I20240306-175812-0116-49168171-p1m.out	Download
Visual Result (SVG)	lalign-I20240306-175812-0116-49168171-p1m.visual-svg	Download
Visual Result (PNG)	lalign-I20240306-175812-0116-49168171-p1m.visual-png	Download
Visual Result (JPEG)	lalign-I20240306-175812-0116-49168171-p1m.visual-jpg	Download
First Input Sequence	lalign-I20240306-175812-0116-49168171-p1m.asequence	Download



Practical part

Try pairwise alignment.
(global and local)

Hw: Compare „your“ sequence (human) with
sequence from mouse (*Mus musculus*).

How similar are these proteins?

Multiple sequence alignment (MSA)

=The alignment of more than two sequences

The goal of MSA is twofold:

- Aligning corresponding regions of the sequences
- Revealing positions that are conserved

The main steps to a useful MSA require

- Choosing the right sequences
- Choosing the right MSA method
- Interpreting the alignment

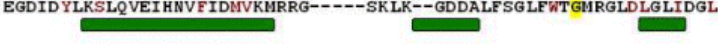
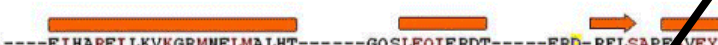
An Example of Conserved Positions: (The Serine Proteases Active Site)

```
CLPP_ECOLI E.col (40) ERVIFLTGQV---EDHMANLIVAQMLFLEAENPEKDIYLYIINSPGGVITAGMSIYDTMQFIKPD---VSTIC (105)
CLP1_MYXXA M.xan (26) DRIIMLGTPV---NDDVANIIVAQLLFESEDPDKGINLYIINSPGGVITAGLAIYDTMQYVKCP---VSTIC ( 91)
21228980 M.maz (27) MISLFLPAYQSIEDAEQVLRWIRKY---RDYPLELILHTPGGQLHASIQIARALKNHFKK---TRVLI ( 92)
15643678 T.mar (58) SISLFGFPVRRYIDIEDSEEILRAIKLTP---SDMPIDLILHTPGGLVLAAEQIARALKMHKGGK---VTVEV (123)
15668307 M.jan (64) SIGLFGIPVYKFITIEDSEEILRAIRAAP---KDKPIDLIHTPGGLVLAATQIARALKAHFAE---TRVIV (129)
18976612 P.fur (59) SIGFFGIPVYKFIISIEDSEEVLRAIRMAP---KDKPIDLIHTPGGLVLAATQIARALKDHPAE---TRVIV (124)
22972030 C.aur (53) TMSLLGFPVLRVINIEDSEAVLRAIKMTD---RDIPIIDLILHTPGGLVLAAEQIARALTKHAAK---VTVEV (118)
23050732 M.bar (75) AISLFGIPAYQYIDEEDAEQILRWIRKY---KDYPLELILHTPGGQLHSSIQIARALRRHSGKN---TKVII (139)
15964138 S.mel (50) HVARVAVTGLIQ---DDRELVERLERIADN---QSVKALIVTISSPGGTTYGGEVIYKAIKRVAEKKP---VVSVD (116)
17934547 A.tum (27) AIMAGGNQFRPALNLASYPALLEKFAVKDA---PAVAISLWSPGGSPVQARMIYNRIKRLAEKDKK---EIFV ( 96)
```

```
CLPP_ECOLI E.col (106) MGQAASMGAFLLTAGAKGKRFCLPNSRVMIHQPLGGY-----QGQATDI----- (147)
CLP1_MYXXA M.xan ( 92) VGQAASMGALLLLAGAKGKRYALPNSRIMIHPPLGGA-----QGQATDI----- (133)
21228980 M.maz ( 93) PHYSMGGTIIALA ADE-IVMDKDAVIGPID---VVGDPVIRGVFPAPSWIHAEETKK-EDADDSTLVMS----- (156)
15643678 T.mar (124) PHYAMSGTIIALA ADE-IVMDENAVLGPVD---PQLGN---MPAPSEAAVKKRDVNEVDDQTLILA--- (184)
15668307 M.jan (130) PHYAMSGTIIALA ADK-IVMDENAVLGPVD---PQLGQ---YPAFISIVKAVEQKQADKADDQTLILA--- (190)
18976612 P.fur (125) PHYAMSGTIIALA ADR-IVMDPHAVLGPVD---PQLGQ---YPAFISIKAVEQKGAEKVDDQTLILA--- (185)
22972030 C.aur (119) PHYAMSGTIIALA ADE-IVMDENAVLGPVD---PQLGQ---HPAASISLVLERKPLSEIDDETLMMA--- (179)
23050732 M.bar (140) PHYSMGGTIIALA ANE-IVMDRDAVIGPID---VIGDFIRGMYPAPSWIYAAETKK-EKADDTTLVMS----- (204)
15964138 S.mel (117) RTLAASAGYLIALAGDR-IVAGETSITGSIG-VIFQY---PQVKTLMDKLGVSLSEIKSRPLKAPSPFFHPPS (184)
17934547 A.tum ( 97) EDVAASGGYMIALAGDE-IIADPTSIVGSIQ-VVSGG---FGFPEMLRKIGVERRYVTAGENWILDFQPEK (164)
```

```
CLPP_ECOLI E.col (148) ---EIHAREILKVKGRMNMELMALHT-----GQSLEQIERDT-----ERD-RFLSAPAEVVEY (196)
CLP1_MYXXA M.xan (134) ---DIQAKEILRLRSYINGLIVKHT-----GHTIERIEKDT-----ERD-YFMSAPDARQY (182)
21228980 M.maz (157) ---DISRKALRLRNVAKELELLEGKIQPD-GKEDRLEEVEKLVSG-EMISTPLSAREAKEL (213)
15643678 T.mar (185) ---DIAEKAIQVKKEFVVEILSDKV---SKEKAEKIADKLCSSG-YWTDYPIIYKELREM (237)
15668307 M.jan (191) ---DIAKKAINQVQNFVYNLLKDKY---GEEKAKELSKILTEG-RWTDYPIITVEAEKEL (243)
18976612 P.fur (186) ---DVAKKAIKQVQDFLYDLLKDKY---GEEKARELAQILTEG-RWTDYPIITVEHAREL (238)
22972030 C.aur (180) ---DIAEKAIQVKRTVCELLRDKM---PVERAEVAHTLASG-VWTDYPIITVSEAREL (232)
23050732 M.bar (205) ---DVSRKALKFTRNVAKELELLEGKIQPGPAGESRLDEVVEKLVSG-EMISTPLSAGEAKKI (262)
15964138 S.mel (185) DEARAMIQAMIDDSYGWFDLVAERRK-----LPRPEALALADGRIFTGRQALEGKLVDEL (240)
17934547 A.tum (165) EGDIDYLSLQVEIHNVFIDNVKMRRG-----SKLKG--GDDALFSGLFWTGRGLDLGLIDGL (220)
```

Active Site



„Evolution in a Nutshell“

Amino acids mutate randomly

Mutations are then selected (accepted) or counter-selected (rejected)

If a mutation is harmful, it is counter-selected

- It disappears from the genome
- You never see it

Mutations of important positions (such as active sites) are almost always harmful

You can recognize important positions because they never mutate!

MSAs reveal these *conserved* positions

An Example of Conserved Positions: (The Serine Proteases Active Site)

```
CLPP_ECOLI E.col (40) ERVIFLTGGV---EDHMANLIVAQMLFLEAENPEKDIYLYIINSPGGVITAGMSIYDTMQFIKPD---VSTIC (105)
CLP1_MYXXA M.xan (26) DRIIMLGTPV---NDDVANIIVAQLLFESEDPDKGINLYIINSPGGVITAGLAIYDTMQYVKCP---VSTIC ( 91)
21228980 M.maz (27) MISLFLPAYQSIEDAEQVLRWIRKY---RDYPLELILHTPGGQLHASIQIARALKNHFKK---TRVLI ( 92)
15643678 T.mar (58) SISLFGFPVRRYIDIEDSEEILRAIKLTP---SDMPIDLILHTPGGLVLAAEQIARALKMHKGGK---VTVEV (123)
15668307 M.jan (64) SIGLFGIPVYKFITIEDSEEILRAIRAAP---KDKPIDLIHTPGGLVLAATQIARALKAHFAE---TRVIV (129)
18976612 P.fur (59) SIGFFGIPVYKFIISIEDSEEVLRAIRMAP---KDKPIDLIHTPGGLVLAATQIARALKDHPAE---TRVIV (124)
22972030 C.aur (53) TMSLLGFPLVRYINIEDSEAVLRAIKMTD---RDIPIIDLILHTPGGLVLAAEQIARALTKHAAK---VTVEV (118)
23050732 M.bar (75) AISLFGIPAYQYIDEEDAEQILRWIRKY---KDYPLELILHTPGGQLHSSIQIARALRRHSKN---TKVII (139)
15964138 S.mel (50) HVARVAVTGLIQ---DDRELVERLERIADN---QSVKALIVTISSPGGTTYGGEVIYKAIRKVAEKKP---VVSVD (116)
17934547 A.tum (27) AIMAGGNQFRPALNLASYPALLEKAFVKDA---PAVAISLWSPGGSPVQARMIYNRIQLAAEKDKK---EIFV ( 96)
```

```
CLPP_ECOLI E.col (106) MGQAASMGAFLLTAGAKGKRFCLPNSRVMIHQPLGGY-----QGQATDI----- (147)
CLP1_MYXXA M.xan ( 92) VGQAASMGALLLLAGAKGKRYALPNSRIMIHPPLGGA-----QGQATDI----- (133)
21228980 M.maz ( 93) PHYSMGGTIIALA ADE-IVMDKDAVIGPID---VVGDPPIRGVFPAPSWIHAEETKK-EDADDSTLVMS----- (156)
15643678 T.mar (124) PHYAMSGTIIALA ADE-IVMDENAVLGPVD---PQLGN---MPAPSEAAVKKRDVNEVDDQTLILA--- (184)
15668307 M.jan (130) PHYAMSGTIIALA ADK-IVMDENAVLGPVD---PQLGQ---YPAFISIVKAVEQKGADKADDQTLILA--- (190)
18976612 P.fur (125) PHYAMSGTIIALA ADR-IVMDPHAVLGPVD---PQLGQ---YPAFSIIVKAVEQKGAEKVDDQTLILA--- (185)
22972030 C.aur (119) PHYAMSGTIIALA ADE-IVMDENAVLGPVD---PQLGQ---HPAASISLVLERKPLSEIDDETLMMA--- (179)
23050732 M.bar (140) PHYSMGGTIIALA ANE-IVMDRDAVIGPID---VIGDFIRGMYPAPSWIYAAETKK-EKADDTTLVMS----- (204)
15964138 S.mel (117) RTLAASAGYLIALAGDR-IVAGETSITGSIG-VIFQY---PQVKTLMDKLGVSLSEIKSRPLKAPSPFFHPPS (184)
17934547 A.tum ( 97) EDVAASGGYMIALAGDE-IIADPTSIVGSIQ-VVSGG---FGFPEMLRKI GVERRVYTAGENW---ILDFFQPEK (164)
```

```
CLPP_ECOLI E.col (148) ---EIHAREILKVKGRMNEMLALHT-----GQSLEQIERDT-----ERD-RFLSAPFARVEY (196)
CLP1_MYXXA M.xan (134) ---DIQAKEILRLRSYINGLIVKHT-----GHTIERIEKDT-----ERD-YFMSAPARQY (182)
21228980 M.maz (157) ---DISRKALRLTRNVAKELELLEGKI QPD-GKEDRLEEVEKLVSG-EMI STPLSAREAKEL (213)
15643678 T.mar (185) ---DIAEKAI RQVKEFVVEILSDKV-----SKEKAEKIADKLCSG-YWTDYPIIYKELREM (237)
15668307 M.jan (191) ---DIAKKAINQVQNFVYNLLKDKY-----GEEKAKELSKILTEG-RWTDYPIITVEAEKEL (243)
18976612 P.fur (186) ---DVAKKAI KQVQDFLYDLLKDKY-----GEEKARELAQILTEG-RWTDYPIITVEHAREL (238)
22972030 C.aur (180) ---DIAEKAI RQVKRTVCELLRDKM-----PVERAEVAHTLASG-VWTDYPIITVSEAREL (232)
23050732 M.bar (205) ---DVSRRKALKFTRNVAKELELLEGKI QPGPAGESRLDEVVEKLVSG-EMI STPLSAGEAKKI (262)
15964138 S.mel (185) DEARAMIQAMIDDSYGWFDLVAERRK-----LPRPEALALADGRIFTGRQALEGKLVDEL (240)
17934547 A.tum (165) EGDIDYKSLQVEIHNVFIDNVKMRRG-----SKLK--GDDALFSGLFWT GHRGLDLGLIDGL (220)
```

Active Site



Using MSA:

<i>Application</i>	<i>Procedure</i>
Extrapolation	Determine the function of your protein
Phylogenetic analysis	Build a Phylogenetic tree
Pattern identification	Discover important positions
Domain identification	Turn your alignment into a domain profile

Gathering Sequences with BLAST

The most convenient way to select your sequences for comparison is to use a BLAST server

➤ Homework 3. : 5) Find and download five similar sequences.

```
>[Pongo pygmaeus]  
MDHRKARVLPAGHYCPSLGIWSSQVGSVRSSVPPSIR  
RHERLREKMRRRLESGDKWFSLEFFPPRTAEGAVNLI  
GLETILHMTCCCHQRLEEITGHLHKAKQLGLKNIMALR
```

remove brackets []

Gathering Sequences with BLAST

→ change sequences (FASTA) names into organism only

The screenshot shows a BLAST search results page with the following elements:

- Navigation tabs:** Descriptions (selected), Graphic Summary, Alignments, Taxonomy.
- Section:** Sequences producing significant alignments.
- Actions:** Download (dropdown menu), Select columns, Show.
- Download Menu:** FASTA (complete sequence), FASTA (aligned sequences), GenBank (complete sequence), Hit Table (text), Hit Table (CSV).
- Sequence List:** A table with checkboxes and descriptions of sequences, including *stearoyl-CoA desaturase* from *Homo sapiens*, *acyl-CoA desaturase* from *Gorilla gorilla gorilla*, and *acyl-CoA desaturase* from *Pan troglodytes*.
- Preview Window:** A window titled "seqdump (1).txt - Poznámkový blok" showing the FASTA format of the selected sequences, with the organism name in brackets at the end of the header line (e.g., >gi|13435426|gb|AAH04579.1| Nqo1 protein [Mus musculus]).

Aligning Your Sequences

Aligning sequences correctly is very difficult

- It's hard to align protein sequences with less than 25% identity (70% identity for DNA)

All methods are approximate

Alignment methods use the progressive algorithm

- Compares the sequences two by two
- Builds a guide tree
- Aligns the sequences in the order indicated by the tree

Alignment: MultAlin



Multiple sequence alignment by Florence Corpet

Published research using this software should cite:
"Multiple sequence alignment with hierarchical clustering"
F. CORPET, 1988, Nucl. Acids Res., 16 (22), 10881-10890



Sequence data

Cut and paste your sequences here below.

```
>gi|13435426|gb|AAH04579.1| Nqo1 protein [Mus musculus]
>gi|71059897|emb|CAJ18492.1| Nqo1 [Mus musculus]
MAARRALIVLAHSERTSPNYAMKEAAVEALKKRGWEVLESPLYAMNPNFIISRNDITGELKDSKNFQYPS
EESLAKKEGR
LSPDIVAEHKKLEAADLVIPQPLQWPGVPAILKGFVFLVAGFAYTYAAMYDNGFPQMKKILLISITG
GSGSMYSLQG
VHGDMVILNFIQSGILRFGFQVLEPOLVYSIGHTFPDARMQILEGWKKRLETVWEETPLYFAPSFLD
LNFQAGFLMK
KEVQEEQKKNKFGLSVGHHLGKSI PADNQIKARK
>gi|524939198|ref|XP_005071892.1| PREDICTED: NAD(P)H dehydrogenase
```

or select a file:

Sequence input format:

For nucleotidic sequences, you must change the Symbol comparison Table (see below)

Substitution matrix: PAM/BLOSUM

Optional Parameters

Result page format:

The sequence alignment will be displayed as

MultAlin

Multalin result page



[Go directly to Alignment](#)

Multalin version 5.4.1
Copyright I.N.R.A. France 1989, 1991, 1994, 1996
Published research using this software should cite
Multiple sequence alignment with hierarchical clustering
F. CORFET, 1988, Nucl. Acids Res., 16 (22), 10881-10890
Symbol comparison table: biosum62
Gap weight: 12
Gap length weight: 2
Consensus levels: high=90% low=50%
Consensus symbols:
! is anyone of IV
\$ is anyone of IM
* is anyone of FV
is anyone of NDQEBZ
MSF: 274 Check: 0
Name: gi1134354261gb|IABM04 Len: 274 Check: 4705 Weight: 1.23
Name: gi15249391981ref|XP_ Len: 274 Check: 6867 Weight: 1.23
Name: gi12274304031ref|NP_ Len: 274 Check: 6661 Weight: 0.89
Name: gi14262425031ref|XP_ Len: 274 Check: 6108 Weight: 0.89
Name: gi13867817031ref|NP_ Len: 274 Check: 4019 Weight: 0.89
Name: gi1302306851gb|IABP20 Len: 274 Check: 4190 Weight: 0.89
Name: Consensus Len: 274 Check: 4506 Weight: 0.00

```
//
1 10 20 30 40 50 60 70 80 90 100 110 120 130
g1134354261gb|IABM04 HARRRRLIVLASEKTSFYNAHKEHIVLKKRGAEVLESOLYAHNFPISRNDDTGLKDSKQFQYPSSESLAKREGALSPDIYVHNRKLEARDLVTFQPLQAFGVPALLKGAFFERYLVGFRITTYA
g15249391981ref|XP_ HARRRRLIVLASEKTSFYNAHKEHIVLKKRGAEVLESOLYAHNFPVTSRNDITGLKDSKQFQYPSSESLAKREGALSPDIYVHNRKLEARDLVTFQPLQAFGVPALLKGAFFERYLVGFRITTYA
g12274304031ref|NP_ HAYPKRLIVLASEKTSFYNAHKEHIVLKKRGAEVLESOLYAHNFPVTSRNDITGLKDSKQFQYPSSESLAKREGALSPDIYVHNRKLEARDLVTFQPLQAFGVPALLKGAFFERYLVGFRITTYA
g14262425031ref|XP_ HAYPKRLIVLASEKTSFYNAHKEHIVLKKRGAEVLESOLYAHNFPVTSRNDITGLKDSKQFQYPSSESLAKREGALSPDIYVHNRKLEARDLVTFQPLQAFGVPALLKGAFFERYLVGFRITTYA
g13867817031ref|NP_ HARRRRLIVLASEKTSFYNAHKEHIVLKKRGAEVLESOLYAHNFPISRNDDTGLKDSKQFQYPSSESLAKREGALSPDIYVHNRKLEARDLVTFQPLQAFGVPALLKGAFFERYLVGFRITTYA
g1302306851gb|IABP20 HARRRRLIVLASEKTSFYNAHKEHIVLKKRGAEVLESOLYAHNFPISRNDDTGLKDSKQFQYPSSESLAKREGALSPDIYVHNRKLEARDLVTFQPLQAFGVPALLKGAFFERYLVGFRITTYA
Consensus
narrfRLIVLASEKTSFYNAHKEHIVLKKRGAEVLESOLYAHNFPISRNDDTGLKDSKQFQYPSSESLAKREGALSPDIYVHNRKLEARDLVTFQPLQAFGVPALLKGAFFERYLVGFRITTYA

131 140 150 160 170 180 190 200 210 220 230 240 250 260
g1134354261gb|IABM04 HAYTNGGPFQNKSTLSTITGGSSQSYSLQVINGRQVLLNPITQSGLLNFQGFQVLEPQLVYSIGHTPDRNRKLEEDRQKLETVARETPLYYPSSFLQNFQAGFLMKKQVQEKQNRGLSVGHIL
g15249391981ref|XP_ HAYTNGGPFQNKSTLSTITGGSSQSYSLQVINGRQVLLNPITQSGLLNFQGFQVLEPQLVYSIGHTPDRNRKLEEDRQKLETVARETPLYYPSSFLQNFQAGFLMKKQVQEKQNRGLSVGHIL
g12274304031ref|NP_ HAYTNGGPFQNKSTLSTITGGSSQSYSLQVINGRQVLLNPITQSGLLNFQGFQVLEPQLVYSIGHTPDRNRKLEEDRQKLETVARETPLYYPSSFLQNFQAGFLMKKQVQEKQNRGLSVGHIL
g14262425031ref|XP_ HAYTNGGPFQNKSTLSTITGGSSQSYSLQVINGRQVLLNPITQSGLLNFQGFQVLEPQLVYSIGHTPDRNRKLEEDRQKLETVARETPLYYPSSFLQNFQAGFLMKKQVQEKQNRGLSVGHIL
g13867817031ref|NP_ HAYTNGGPFQNKSTLSTITGGSSQSYSLQVINGRQVLLNPITQSGLLNFQGFQVLEPQLVYSIGHTPDRNRKLEEDRQKLETVARETPLYYPSSFLQNFQAGFLMKKQVQEKQNRGLSVGHIL
g1302306851gb|IABP20 HAYTNGGPFQNKSTLSTITGGSSQSYSLQVINGRQVLLNPITQSGLLNFQGFQVLEPQLVYSIGHTPDRNRKLEEDRQKLETVARETPLYYPSSFLQNFQAGFLMKKQVQEKQNRGLSVGHIL
Consensus
HAYTNGGPFQNKSTLSTITGGSSQSYSLQVINGRQVLLNPITQSGLLNFQGFQVLEPQLVYSIGHTPDRNRKLEEDRQKLETVARETPLYYPSSFLQNFQAGFLMKKQVQEKQNRGLSVGHIL

261 270 274
g1134354261gb|IABM04 GKSIPRNDLKKRK
g15249391981ref|XP_ GKSIPRNDLKKRK
g12274304031ref|NP_ GKSIPRNDLKKRK
g14262425031ref|XP_ GKSIPRNDLKKRK
g13867817031ref|NP_ GKSIPRNDLKKRK
g1302306851gb|IABP20 GKSIPRNDLKKRK
Consensus
GKSIPRNDLKKRK
```

- Available files:
- Sequence input file
 - Cluster file
 - Results as a fasta file
 - Results as a text page (msf)
 - Results as postscript page(s) with ESPript (protein only)
 - Alignment and tree description (frd) Get a better view of your protein family : phylogenetic tree, pruned tree and subtrees, summarised coloured alignment and subalignments.
 - Results as an html page (needs to enable style sheets)
 - Results as a text page with colour indications (need a text editor)
 - Results as a gif image

Amino Acid	Characteristic
W, Y, F	It is common to find conserved tryptophans. Tryptophan is a large hydrophobic residue that sits deep in the core of proteins. It plays an important role in their stability and is therefore difficult to mutate. When tryptophan mutates, it is usually replaced by another aromatic amino acid, such as phenylalanine or tyrosine. Patterns of conserved aromatic amino acids constitute the most common signatures for recognizing protein domains.
G, P	It is common to find conserved columns with a glycine or a proline in a multiple alignment. These two amino acids often coincide with the extremities of well-structured beta strands or alpha helices. (For more on these structures, see Chapter 11.)
C	Cysteines are famous for making C-C (disulphide) bridges. Conserved columns of cysteines are rather common and usually indicate such bridges. Columns of conserved cysteines with a specific distance provide a useful signature for recognizing protein domains and folds.
H, S	Histidine and serine are often involved in catalytic sites, especially those of proteases. Conserved histidine or a conserved serine are good candidates for being part of an active site.
K, R, D, E	These charged amino acids are often involved in ligand binding. Highly conserved columns can also indicate a salt bridge inside the core of the protein.
L	Leucines are rarely very conserved unless they're involved in protein-protein interactions such as a leucine zipper.

Alignment: NCBI/COBALT



COBALT

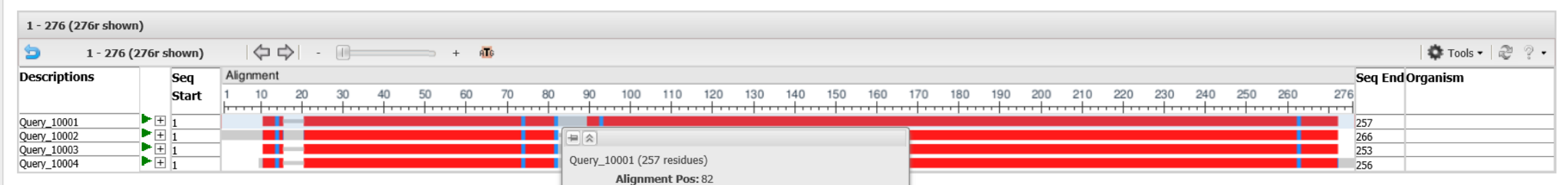
Constraint-based Multiple Alignment Tool

[Home](#) [Recent Results](#) [Help](#)

[Phylogenetic Tree](#) [Edit and Resubmit](#) [Download](#)

- Cobalt RID DK0EGH2J212 (4 seqs)

Graphical Overview



Descriptions Select All [Re-align](#) [Alignment parameters](#)

Accession	Links
<input checked="" type="checkbox"/> Icl Query_10001	tr V9PWX7 V9PWX7_SCHMA Sulfotransferase [unclassified] 3_07706 PE=4 SV=1
<input checked="" type="checkbox"/> Icl Query_10002	tr A0A094ZWWQ2 A0A094ZWWQ2_SCHHA Uncharacterized protein um PE=2 SV=1
<input checked="" type="checkbox"/> Icl Query_10003	tr C1LER5 C1LER5_SCHJA Uncharacterized protein
<input checked="" type="checkbox"/> Icl Query_10004	tr C1LS15 C1LS15_SCHJA Cell wall integrin

Alignment: Clustal Omega

Clustal Omega

Multiple Sequence Alignment (MSA)

[Job Dispatcher](#) [Help & Privacy](#) [Input form](#)

[Feedback](#)

Welcome to the new **Job Dispatcher** website. We'd love to hear your [feedback](#) about the new webpages!

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Input sequence ⓘ

Sequence Type

Protein DNA RNA

Paste your sequence here - or use the example sequence

[Zvolit soubor](#) [Nevybrán žádný soubor](#)

[Use the example](#)

[Clear sequence](#)

[More example inputs](#)

Parameters

OUTPUT FORMAT ⓘ

ClustalW with character counts

All sequences in fasta format



Alignment: Clustal Omega

Results for Job ID: clustalo-I20240306-172217-0875-65691984-p1m

Alignments | Tool Output | Guide Tree | Phylogenetic Tree | Results Viewers | Result Files | Submission Details

Nightingale

COLOR SCHEME: clustal2

LEGEND: ARND CQEGHILKMF PSTWYVBXZ

6 sequences

SU
CROCUT
LEMU
MACAC
[HOM
PA

Sequence	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260
SU	M	A	V	R	K	A	L	I	I	L	A	H	S	E	K	T	S	F	N	Y	A	M	K	E	A	A	V	E	A	L	K	R	G	W	E	V	A	V	S	D	L	Y	A	M	N	F	N	P	V	I	S	R	K	D	I	T	G	K	L	K	D	P	G	N	F	Q	Y	P	A	E	T	A	L	A	Y	K	E	G	R	L	S	P	D	I	V	A	E	Q	K	K	V	E	A	A	D	L	V	I																																																																																																																																																																		
CROCUT	-	I	A	R	R	A	L	I	V	L	A	H	A	E	T	T	S	F	N	H	A	M	K	E	A	A	V	E	A	L	K	S	K	G	W	E	V	T	V	S	D	L	Y	A	M	N	F	N	P	V	I	S	R	K	D	I	T	G	K	L	K	D	P	P	E	N	F	Q	Y	P	A	E	S	V	L	A	Y	K	E	G	R	L	S	P	D	I	V	A	E	Q	K	K	L	E	A	A	D	L	V	I																																																																																																																																																																
LEMU	M	A	A	R	K	A	L	I	V	L	A	H	S	E	R	T	S	F	N	H	A	M	K	E	E	A	A	V	A	A	L	K	K	K	G	W	E	V	A	V	S	D	L	Y	A	M	N	F	N	P	I	I	S	R	K	D	I	T	G	K	L	K	D	P	A	N	F	Q	Y	A	A	E	S	T	L	A	Y	K	E	G	R	L	S	P	D	I	V	A	E	Q	K	K	L	E	A	A	D	L	V	I																																																																																																																																																																
MACAC	M	V	G	R	R	A	L	I	V	L	A	H	S	E	R	T	S	F	N	Y	A	M	K	E	E	A	A	V	A	A	L	K	K	K	G	W	E	V	V	E	S	D	L	Y	A	M	N	F	N	P	I	I	S	R	K	D	I	T	G	K	L	K	D	P	A	N	F	Q	Y	P	A	E	S	V	L	A	Y	K	E	G	H	L	S	P	D	I	V	A	E	Q	K	K	L	E	A	A	D	L	V	I																																																																																																																																																																
[HOM	M	V	G	R	R	A	L	I	V	L	A	H	S	E	R	T	S	F	N	Y	A	M	K	E	E	A	A	V	A	A	L	K	K	K	G	W	E	V	V	E	S	D	L	Y	A	M	N	F	N	P	I	I	S	R	K	D	I	T	G	K	L	K	D	P	A	N	F	Q	Y	P	A	E	S	V	L	A	Y	K	E	G	H	L	S	P	D	I	V	A	E	Q	K	K	L	E	A	A	D	L	V	I																																																																																																																																																																
PA	M	V	G	R	R	A	L	I	V	L	A	H	S	E	R	T	S	F	N	Y	A	M	K	E	E	A	A	V	A	A	L	K	K	K	G	W	E	V	V	E	S	D	L	Y	A	M	N	F	N	P	I	I	S	R	K	D	I	T	G	K	L	K	D	P	A	N	F	Q	Y	P	A	E	S	V	L	A	Y	K	E	G	H	L	S	P	D	I	V	A	E	Q	K	K	L	E	A	A	D	L	V	I																																																																																																																																																																

(*) conserved amino acids
(:) amino acids with similar size and hydrophobicity
(.) amino acids with similar size or hydrophobicity

Alignment: Clustal Omega

Results for Job ID: clustalo-I20240306-175517-0875-65691984-p1m 34-p1m

Alignments | Tool Output | Guide Tree | Phylogenetic Tree | Results Viewers | Result Files | Submission Details

Nightingale

CLUSTAL O(1.2.4) multiple sequence alignment

```

Sus      MAVRKAL IILAHSEKTSFNHYAMKEAAVEALKRRGWEVAVSDLYAMNPNVISRKDITGKL 60
Crocota  -IARRALIVLHAETTSFNHAMKEAAVEALKKSGWEVTVSDLYAMNPNVISRRDITGTL 59
Lemur    MAARKALIVLHSSERTSFNHAMKDAALAEALKKGGWEVAVSDLYAMNPNPIISKDITGKL 60
Macaca   MVGKRALIVLHSSERTSFNYAMKEAAVAALKKGGWEVAVSDLYAMNPNPIISKDITGKL 60
[Homo    MVGRRALIVLHSSERTSFNYAMKEAAAAALKKGGWEVAVSDLYAMNPNPIISKDITGKL 60
Pan      MVGRRALIVLHSSERTSFNYAMKEAAAAALKKGGWEVAVSDLYAMNPNPIISKDITGKL 60
:::***:*.:*****:*****:*****:*****:*****:*****:*****:

Sus      KDPGNFQYPAETALAYKEGRSPDIVAEQKKVEAADLVIPFPLQWFGVPAILKGFPERV 120
Crocota  KDPGNFQYPAESVLAYKEGRSPDIVAEQKLEAADLVIPFPLQWFGVPAILKGFPERV 119
Lemur    KDPENFQYPVESVLAYKEGRSPDIVAEQKLEAADLVIPFPLQWFGVPAILKGFPERV 120
Macaca   KDPANFYAAESTLAYKEGRSPDIVAEQKLEAADLVIPFPLQWFGVPAILKGFPERV 120
[Homo    KDPANFYPAESVLAYKEGRSPDIVAEQKLEAADLVIPFPLQWFGVPAILKGFPERV 120
Pan      KEPANFYPAESVLAYKEGRSPDIVAEQKLEAADLVIPFPLQWFGVPAILKGFPERV 120
*: *  ***  .*:*****:*****:*****:*****:*****:*****:

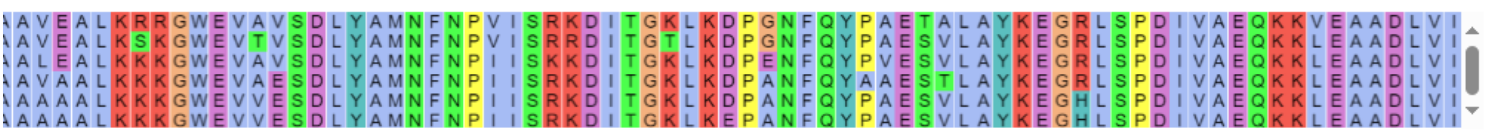
Sus      LIGEFAYTYAAMYDKGPRFKKAVLSITTTGGSGSMYSLQGIGHGDMVILWPIQSGILHFC 180
Crocota  LIGGFAYTYAAMYDNGPFRNKKTVLSITTTGGSGSMYSLQGIGHGDMVILWPIQSGTLHFC 179
Lemur    LIGEFAYSYAAMYDKGPRNKKTLVLSITTTGGSGSMYSLQGIGHGDMVILWPLQSGTLHFC 180
Macaca   FVGEFAYTLAAMYDKGPRFSKAVLSITTTGGSGSMYSLQGIGHGDMVILWPIQSGILHFC 180
[Homo    FIGEFAYTYAAMYDKGPRFSKAVLSITTTGGSGSMYSLQGIGHGDMVILWPIQSGILHFC 180
Pan      FIGEFAYTYAAMYDKGPRFSKAVLSITTTGGSGSMYSLQGIGHGDMVILWPIQSGILHFC 180
::*  ***:  *****:***:*****:*****:*****:*****:*****:

Sus      GFQVLEPQLTYSIGHTPADARIQILEGKKRLENINWDETPLYFAPSSLPDLNFQAGFLMK 240
Crocota  GFQVLEPQLTYSIGHTPDARIQILEGKKRLENINWDETPLYFAPSSLPDLNFQAGFLMK 239
Lemur    GFQVLEPQLTYSIGHTPADARIQILEGKKRLENINWDETPLYFAPSSLPDLNFQAGFVYMK 240
Macaca   GFQVLEPQLTYSIGHTPADARIQILEGKKRLENINWDETPLYFAPSSLPDLNFQAGFLMK 240
[Homo    GFQVLEPQLTYSIGHTPADARIQILEGKKRLENINWDETPLYFAPSSLPDLNFQAGFLMK 240
Pan      GFQVLEPQLTYSIGHTPADARIQILEGKKRLENINWDETPLYFAPSSLPDLNFQAGFLMK 240
*****:*****:*****:*****:*****:*****:*****:

Sus      KQVQDEQKSKKFGLSVGHHLGKSIPTDNQIKARK 274
Crocota  KEVQDEQKSKKFGLSVGHHLGKSIPTDNQIKARK 273
Lemur    KEVQDEQKSKKFGLSVGHHLGKSIPTDNQIKARK 274
Macaca   KEVQDEEKNKFGLSVGHHLGKSIPTDNQIKARK 274
[Homo    KEVQDEEKNKFGLSVGHHLGKSIPTDNQIKARK 274
Pan      KEVQDEEKNKFGLSVGHHLGKSIPTDNQIKARK 274
*:***:*.:*****:*****:*****:*****:*****:

```

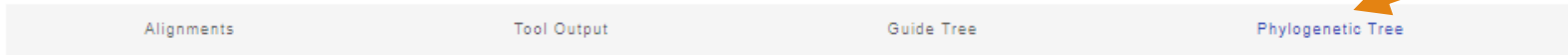
240 260



(*) conserved amino acids
 (:) amino acids with similar size and hydrophobicity
 (.) amino acids with similar size or hydrophobicity

Alignment: Clustal Omega

Results for Job ID: clustalo-I20240306-181352-0108-64863764-p1m

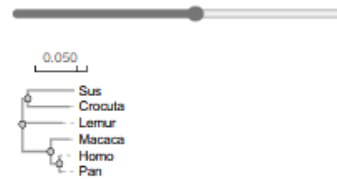


Phylogenetic Tree

```
(  
(  
Sus:0.05566,  
Crocuta:0.05057)  
:0.00639,  
Lemur:0.04885,  
(  
Macaca:0.02251,  
(  
Homo:0.00000,  
Pan:0.00365)  
:0.01083)  
:0.03377);
```

„Newick“ formate

Phylogram



Phylogenetic tree

Practical part

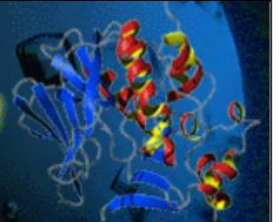
Try multiple alignment.

Download five similar sequences from different organisms. (Hw3)

„advanced“ phylogeny analysis



Information
Génomique et
Structurale



Home

Phylogeny Analysis

"One Click"

"Advanced"

"A la Carte"

"One Click" Mode

Alignment
MUSCLE

Curation
Gblocks

Phylogeny
PhyML

Tree Rendering
TreeDyn

1. Overview

2. Data & Settings

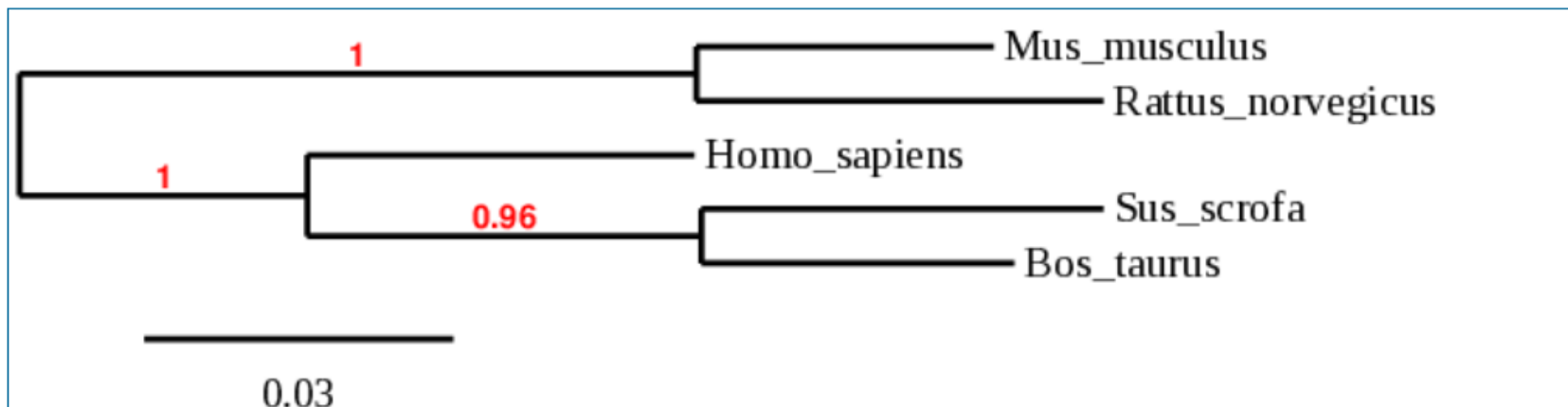
3. Alignment

4. Curation

5. Phylogeny

6. Tree Rendering

Tree Rendering results



Practical part

Try building the phylogeny tree using
[phylogeny.org](https://www.phylogeny.org)

Compare the trees



3-D protein structure: PDB

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB Login

RCSB PDB PROTEIN DATA BANK An Information Portal to 128330 Biological Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands Go

Advanced Search | Browse by Annotations

PDB-101 Worldwide PDB EMDataBank NUCLEIC ACID DATABASE Structural Biology Knowledgebase Worldwide Protein Data Bank Foundation

14 Structures 2 Unreleased Structures 10 Citations 12 Ligands

Search Parameter:

Refine Search Save Search to MyPDB

Text Search for: nqo1 and TAXONOMY is just Homo sapiens (human)

Refinements



Currently showing 1 - 14 of 14

Displaying 25 Results

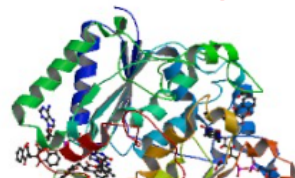
ORGANISM
Homo sapiens only (14)

UNIPROT MOLECULE NAME
NAD(P)H dehydrogenase [qu ... (13)
Ribosyldihyronicotinamid ... (1)
Refine Query

TAXONOMY

View: Detailed Reports: Select a Report Sort: Release Date: Newest to Oldest Download Files

practical example



5FUQ

CRYSTAL STRUCTURE OF THE H80R VARIANT OF NQO1 BOUND TO DICOUMAROL

Medina-Carmona, E., Fuchs, J.E., Gavira, J.A., Salido, E., Palomino-Morales, R., Mesa-Torres, N., Timson, D.L., Roy, A.J.

Practical part

Try PDB.

Find out if your sequence has a 3D structure.

Enzyme database: Brenda

go to...   login  history  all enzymes  Contact

HOME
Classic view

1987-2019
The Comprehensive Enzyme Information System

 Technische Universität Braunschweig

A new class EC 7, Translocases, is available, now. Read more about EC 7 at the IUBMB.

Please enter a search term

Enzyme, Ligand contains

add search field delete search field start search

Text-based queries

- Full-text Search
- Advanced Search
- Enzyme & Disease

Structure-based queries

- Ligand Structure Search
- Metabolic Pathways
- Enzyme Structures

Explorer

- Enzyme Classification
- TaxTree
- Protein folding: CATH / SCOPe

Translocases (900 organisms)

EC class 7


These enzymes catalyse the movement of ions or molecules across membranes or their separation within membranes, the reaction is designated as a transfer from side 1 to side 2 because the designations in and out, which had previously been used, can be ambiguous. The subclasses designate the types of components transferred and the sub-sub-classes indicate the reaction processes that provide the driving force for the translocation.

EC Browser

- 1 Oxidoreductases (9651 organisms)
- 2 Transferases (6622 organisms)
- 3 Hydrolases (10604 organisms)
- 4 Lyases (5111 organisms)
- 5 Isomerases (2083 organisms)
- 6 Ligases (1547 organisms)
- 7 Translocases (966 organisms)




Protein interactions

Version: 11.0 LOGIN | REGISTER

 Search Download Help My Data

There are several matches for 'NQO1'.
Please select one from the list below and press Continue to proceed. [<- BACK](#) [CONTINUE ->](#)

organism	protein
<input checked="" type="checkbox"/> Homo sapiens	NQO1 - NAD(P)H dehydrogenase [quinone] 1; The enzyme apparently serves as a quinone reductase in connection with conjugation reactions of hydroquinons involved in detoxification pathways as well as in biosynthetic processes such as the vitamin K-dependent gamma-carboxylation of glutamate residues in prothrombin synthesis; Belongs to the NAD(P)H dehydrogenase (quinone) family
<input type="checkbox"/> Homo sapiens	TCF7L1 - Transcription factor 7-like 1; Participates in the Wnt signaling pathway. Binds to DNA and acts as a repressor in the absence of CTNNB1, and as an activator in its presence. Necessary for the terminal differentiation of epidermal cells, the formation of keratohyalin granules and the development of the barrier function of the epidermis (By similarity). Down-regulates NQO1 , leading to increased mitomycin c resistance; TCF/LEF transcription factor family [a.k.a. <i>TCF3</i> , <i>Hs.516297</i> , <i>transcription factor 7 like 1</i>]

© STRING CONSORTIUM 2020	ABOUT	INFO	ACCESS	CREDITS
 SIB - Swiss Institute of Bioinformatics	Content	Scores	Versions	Funding
 CPR - Novo Nordisk Foundation Center Protein Research	References	Use scenarios	APIs	Datasources
 EMBL - European Molecular Biology Laboratory	Contributors	FAQs	Licensing	Partners
	Statistics	Cookies/Privacy	Usage	Software

<input type="checkbox"/> Balaenoptera acutorostrata	NQO1 - NAD(P)H dehydrogenase [quinone] 1
--	---

Look into the specific databases

Does your protein have any interaction partners?

Is your protein an enzyme? Find E.C. (**Hw**)

„Protein bioinformatics III“

Retrieving protein sequences from databases (Uniprot: FASTA formate)

Computing amino-acids compositions, molecular weight, isoelectric point, and other parameters (SMS)

Prediction of proteases cutting (PeptideCutter)

Predicting elements of protein secondary structure, signal peptide, transmembrane helix

Finding 3-D structure

Finding all proteins that share a similar sequence

Finding evolutionary relationships between proteins, drawing proteins' family trees

Computing the optimal alignment between two or more protein sequences

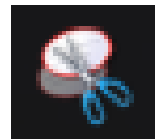
...

Homework 4

Work with „your“ protein.

- 1) Compare your sequence with the „same“ sequence from mouse, how identical are they?
- 2) Prepare multiple alignment of the five sequences from Hw3, snip the phylogeny tree.
- 3) Does your sequence have any isoforms (search in Uniprot)? Align them.
- 4) Is there a 3D structure? Snip one figure.
- 5) Is your protein an enzyme? Find E.C.

E.g use „výstřížky“



„snipping tool“

➤ Compile in „one note“ (or word, or pdf)

Homework 4:example

DÚ4

```
>>sp|Q64669|MQ01_MOUSE_NAD(P)H_dehydrogenase_[quinone]_1_(274_aa)
Weighted Score: 1626; 421.3 bits; E(1) < 1.1e-132
86.5% identity (97.8% similar) in 274 aa overlap (1-274:1-274)
```

i:86,5% Porovni probhlo v celém rozsahu obou proteinů.

1)

```

      10      20      30      40      50      60
sp|P15 MVGRRALIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRKIDITGKL
      . . . . .
sp|Q64 MAARRALIVLAHSEKTSYNYAMKEAAVERLEKKGWVLESDDLYAMNFWPIISRNDITGEL
      10      20      30      40      50      60

      70      80      90      100     110     120
sp|P15 KDFANFQYPAESVLAAYKRGHLSPDIVAEQKRLAADLVIPQFPLQWFGVPAILKGFPERV
      . . . . .
sp|Q64 KDSKNFCYPSSESLAYKRGRLSPDIVAEHKKLEAADLVIQFPLQWFGVPAILKGFPERV
      70      80      90      100     110     120

      130     140     150     160     170     180
sp|P15 FIGEFAYTYAAMYDKGPFASKAVLSITTGSGSMSYSLQGIHGDMNVILWPIQSGILRPF
      . . . . .
sp|Q64 LVAGFAYTYAAMYDNGPFQNKRTLLSITTGSGSMSYSLQGVHGDMNVILWPIQSGILRPF
      130     140     150     160     170     180

      190     200     210     220     230     240
sp|P15 GFQVLEPQLTYSIGHTPADARIQILEGWKKRLEINWDETPLYFAPSSLEFDLNFQAGFLMK
      . . . . .
sp|Q64 GFQVLEPQLVYSIGHTFPDARMQILEGWKKRLETVMEETPLYFAPSSLEFDLNFQAGFLMK
      190     200     210     220     230     240

      250     260     270
sp|P15 KEVQDEEKKKKFGLSVGHHLGKSIPTDQIKARK
      . . . . .
sp|Q64 KEVQEQKKNKFGLSVGHHLGKSIPTDQIKARK
      250     260     270
```

2)

```

1  10  20  30  40  50  60  70  80  90  100 110 120 130
Macaca MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRKIDITGEL
Sus MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRNDITGEL
Mus MFAKRLIVLAHSERTSYNYAMKEAAVERLEKKGWVLESDDLYAMNFWPIISRNDITGEL
Alligator MFAKRLIVLAHSERTSYNYAMKEAAVERLEKKGWVLESDDLYAMNFWPIISRNDITGEL
Consensus MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRNDITGEL

131 140 150 160 170 180 190 200 210 220 230 240 250 260
Macaca MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRKIDITGEL
Sus MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRNDITGEL
Mus MFAKRLIVLAHSERTSYNYAMKEAAVERLEKKGWVLESDDLYAMNFWPIISRNDITGEL
Alligator MFAKRLIVLAHSERTSYNYAMKEAAVERLEKKGWVLESDDLYAMNFWPIISRNDITGEL
Consensus MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRNDITGEL

261 270 274
Macaca KEVQDEEKKKKFGLSVGHHLGKSIPTDQIKARK
Sus KEVQDEEKKKKFGLSVGHHLGKSIPTDQIKARK
Mus KEVQEQKKNKFGLSVGHHLGKSIPTDQIKARK
Alligator KEVQEQKKNKFGLSVGHHLGKSIPTDQIKARK
Consensus KEVQEQKKNKFGLSVGHHLGKSIPTDQIKARK
```



3)

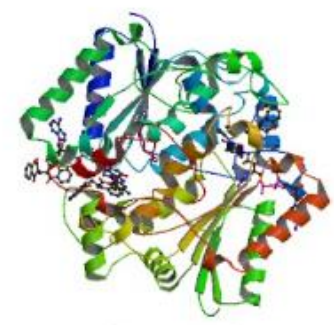
```

1  10  20  30  40  50  60  70  80  90  100 110 120 130
Isofar2 MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRKIDITGEL
NBD1 MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRNDITGEL
Isofar3 MFAKRLIVLAHSERTSYNYAMKEAAVERLEKKGWVLESDDLYAMNFWPIISRNDITGEL
Consensus MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRNDITGEL

131 140 150 160 170 180 190 200 210 220 230 240 250 260
Isofar2 MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRKIDITGEL
NBD1 MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRNDITGEL
Isofar3 MFAKRLIVLAHSERTSYNYAMKEAAVERLEKKGWVLESDDLYAMNFWPIISRNDITGEL
Consensus MFAKRLIVLAHSERTSYNYAMKEAAAAALEKKGWVVESDLYAMNFWPIISRNDITGEL

261 270 274
Isofar2 KEVQDEEKKKKFGLSVGHHLGKSIPTDQIKARK
NBD1 KEVQDEEKKKKFGLSVGHHLGKSIPTDQIKARK
Isofar3 KEVQEQKKNKFGLSVGHHLGKSIPTDQIKARK
Consensus KEVQEQKKNKFGLSVGHHLGKSIPTDQIKARK
```

4)



Protein bioinformatics I-III

SUMMARY AND EXAMPLES



Ex1: DHRS7

Find two human DHRS7 sequences: DHRS7B (AAH09679.1) and DHRS7C (AAI47025.1)

Run pairwise alignment. How identical are these two proteins?

Ex1: DHRS7

Find two human DHRS7 sequences: DHRS7B (AAH09679.1) and DHRS7C (AAI47025.1)

Run pairwise alignment. How identical are these two proteins?

```
>>AAH09679.1 Dehydrogenase/reductase (SDR family) member (325 aa)
Waterman-Eggert score: 840; 246.0 bits; E(1) < 9e-70
45.0% identity (76.3% similar) in 300 aa overlap (5-301:24-319)
zde je požadovaná informace o identitě sekvencí (u
prvního náleženého porovnání)
      10      20      30      40      50      60
AAI470 MLPLLL--LGISGLLFYQEVSRWLSKSAVQNKVVVITDAISGLGKECARVFHTGGARLV
      . . . . . : : : : : : : : : : : : : : : : : : : : : : : : : : :
AAH096 ILPLLLFGCLGVFGLFRLLQWVR---GKAYLRNAVVVITGATSGLGKECAKVFYAAGAKLV
      30      40      50      60      70      80

      70      80      90      100     110     120
AAI470 LCGKNWERLENLYDALI-SVADPSKTFTPKLVLLDLSDISCPDVAKEVLDYCGVCDILI
      . . . . . : : : : : : : : : : : : : : : : : : : : : : : : : : :
AAH096 LCGRNGGALEELIRELTASHATKVQTHKPYLVTFDLTDSGAIVAAAEEILQCFGYVDILV
      90      100     110     120     130     140

      130     140     150     160     170     180
AAI470 NNASVKVKGPAHKISLELDKKIMDANYFGPITLTKALLPNMISRRTGQIVLVNNIQKFKG
      . . . . . : : : : : : : : : : : : : : : : : : : : : : : : : : :
AAH096 NNAGISYRGTIMDDTVVDKRVMETNYFGPVALTKALLPSMIKRRQGHIVAISISQKMS
      150     160     170     180     190     200

      190     200     210     220     230     240
AAI470 IPFRTTYAASKHAALGFDFCLRAEVEEYDVVISTVSPTFIRSYHYVPEQGNWEASIKWFF
      . . . . . : : : : : : : : : : : : : : : : : : : : : : : : : : :
AAH096 IPFRSAYAASKHATQAFDFCLRAEMEQQEIEVTVISPGYIHT-NLSVNAITADGSRVGM
      210     220     230     240     250

      250     260     270     280     290     300
AAI470 FRKLTYGVHPVEVAEEVMRTVRRKKQEVFMANPIPKAAVYVRTFFPEFFAVVACGVKEK
      . : : : : . . : : : . . : : : . . : : : . . : : : . . : : : . .
AAH096 DTTTAQGRSPVEVAQDVLAAVGGKKKDVILADLLPSLAVYLRLTAPGLFFSLMASRARKE
      260     270     280     290     300     310
```


Ex2: NQO1 isoforms

Find in Uniprot sequences of human NQO1 isoforms and align them.

How many isoforms are there?

Compare the output to description of each isoform, is it correct?

Ex2: NQO1 isoforms

Find in Uniprot sequences of human NQO1 isoforms and align them.

How many isoforms are there?

Compare the output to description of each isoform, is it correct?

- **Hint:** Uniprot ID: P15559, on the left sequence (3), download fasta formats and align them using eg.multalin, description is below each isoform (i2:140-173: Missing.)

```
1      10      20      30      40      50      60      70      80      90      100     110     120     130
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
sp|P15559|NQO1_HUMAN  MYGRRALIVLAHSERTSFNYAMKEAAAAALKKKGHEVVESDLYAMNFNPIISRKDIITGKLDKDPANFQYPAESVLAAYKEGHLSPDIYAEQKKLEADLVIFQFPLQWFGVPAILKGMFERVFIGEFAYTYR
sp|P15559-2|NQO1_HUM MYGRRALIVLAHSERTSFNYAMKEAAAAALKKKGHEVVESDLYAMNFNPIISRKDIITGKLDKDPANFQYPAESVLAAYKEGHLSPDIYAEQKKLEADLVIFQFPLQWFGVPAILKGMFERVFIGEFAYTYR
sp|P15559-3|NQO1_HUM MYGRRALIVLAHSERTSFNYAMKEAAAAALKKKGHEVVESDLYAMNFNPIISRKDIITGKLDKDPANFQYPAESVLAAYKEGHLSPDIYAEQKKLEADLVIFQ-----
Consensus            MYGRRALIVLAHSERTSFNYAMKEAAAAALKKKGHEVVESDLYAMNFNPIISRKDIITGKLDKDPANFQYPAESVLAAYKEGHLSPDIYAEQKKLEADLVIFQfplqwfvgvpailkgufervfigefaytya

131     140     150     160     170     180     190     200     210     220     230     240     250     260
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
sp|P15559|NQO1_HUMAN  ANYDKGPFRSKKAVLSITTTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFGVLEPQLTYSIGHTPADARIQILEGAKKRLLENIDETPLVFAPSSLFDLNFQAGFLMKKEVQDEEKKKFGLSYGHHL
sp|P15559-2|NQO1_HUM ANYDKGPFRS-----GILHFCGFGVLEPQLTYSIGHTPADARIQILEGAKKRLLENIDETPLVFAPSSLFDLNFQAGFLMKKEVQDEEKKKFGLSYGHHL
sp|P15559-3|NQO1_HUM -----SKKAVLSITTTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFGVLEPQLTYSIGHTPADARIQILEGAKKRLLENIDETPLVFAPSSLFDLNFQAGFLMKKEVQDEEKKKFGLSYGHHL
Consensus            anydkgpfRSkkavlsittggsgsmyslqgihgdmnvilwpiqsGILHFCGFGVLEPQLTYSIGHTPADARIQILEGAKKRLLENIDETPLVFAPSSLFDLNFQAGFLMKKEVQDEEKKKFGLSYGHHL

261     270     274
|-----|-----|
sp|P15559|NQO1_HUMAN  GKSIPDNQIKARK
sp|P15559-2|NQO1_HUM GKSIPDNQIKARK
sp|P15559-3|NQO1_HUM GKSIPDNQIKARK
Consensus            GKSIPDNQIKARK
```

Isoform 2 (identifier: **P15559-2**) [UniParc] [FASTA](#) [Added to basket](#)

The sequence of this isoform differs from the canonical sequence as follows:
140-173: Missing.

Ex3: sequence identification

What is the proposed function of unknown protein? (Ex3 in Moodle)

What organism does it come from?

Does the „unknown sequence“ have any transmembrane helices?

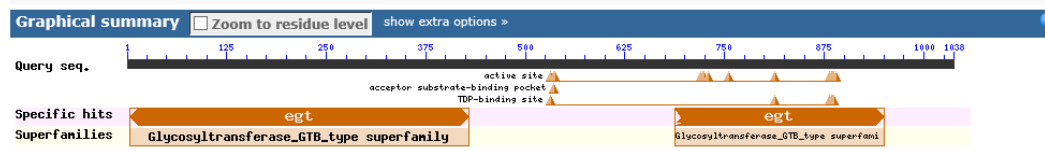
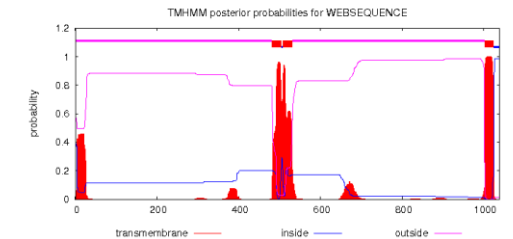
Ex3: sequence identification

What is the proposed function of unknown protein? (Ex3 in Moodle)

What organism does it come from?

Does the „unknown sequence“ have any transmembrane helices?

```
# WEBSEQUENCE Length: 1038
# WEBSEQUENCE Number of predicted TMs: 3
# WEBSEQUENCE Exp number of Aba in TMs: 72.47095
# WEBSEQUENCE Exp number, first 60 AA: 9.92926
# WEBSEQUENCE Total prob of N-in: 0.63176
WEBSEQUENCE TMHMM2.0 outside 1 481
WEBSEQUENCE TMHMM2.0 TMhelix 482 504
WEBSEQUENCE TMHMM2.0 inside 505 505
WEBSEQUENCE TMHMM2.0 TMhelix 509 531
WEBSEQUENCE TMHMM2.0 outside 532 1001
WEBSEQUENCE TMHMM2.0 TMhelix 1002 1024
WEBSEQUENCE TMHMM2.0 inside 1025 1038
```



[UDP-glucuronosyl UDP-glucosyltransferase domain containing protein \[Haemonchus contortus\]](#)

2152 2152 100% 0.0 100% [CDJ89503.1](#)

Version 07/04/2022

A) Work with the following sequence obtained after sequencing (also in Moodle):

```
TACTGTTTTTCGTACAGTTTTTGTAAATAAAAAAACCTATAAATATTCCGGATTATTCATACCGTCCCACCAT  
CGGGCGCGGATCTTTTTATCTAGCATAGCCAAAAAGAAAGAGCTTGCACATATGGAGAGATCAAACAGCA  
CAGCTTCTATGGCCGTGCAGAACTTCACCATGGAGCTATGGAGATTATGATAATTGCCAACAGGATCAT
```

B) Find a human protein sequence called FOX1

- What is the accession number and function of this protein?
- Does this protein have any transmembrane regions?
- How many cysteines does the sequence have?
- Compare how similar the protein is to the respective mouse homologue?
- Design primers to amplify the CDS of respective gene.
 - Compare how similar the protein is to the respective mouse homologue?
 - Design primers to amplify the CDS of respective gene.