Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2024/2025

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

…

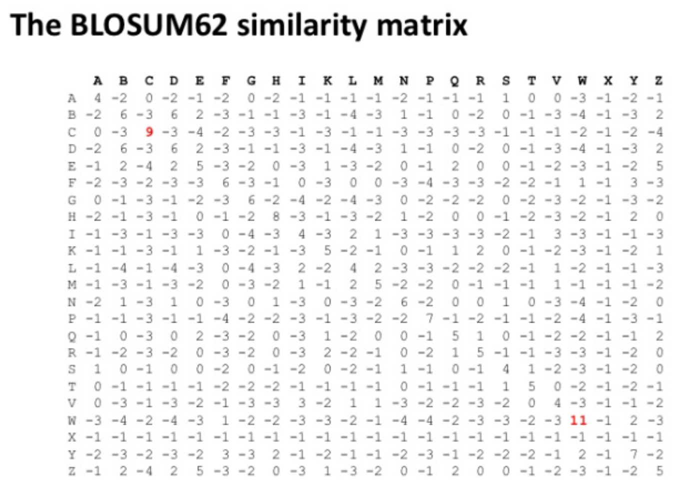
Searching for similar sequences

Similarity x Homology

BLAST: Basic Local Alignment and Search Tool

Finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Similarity matrix:



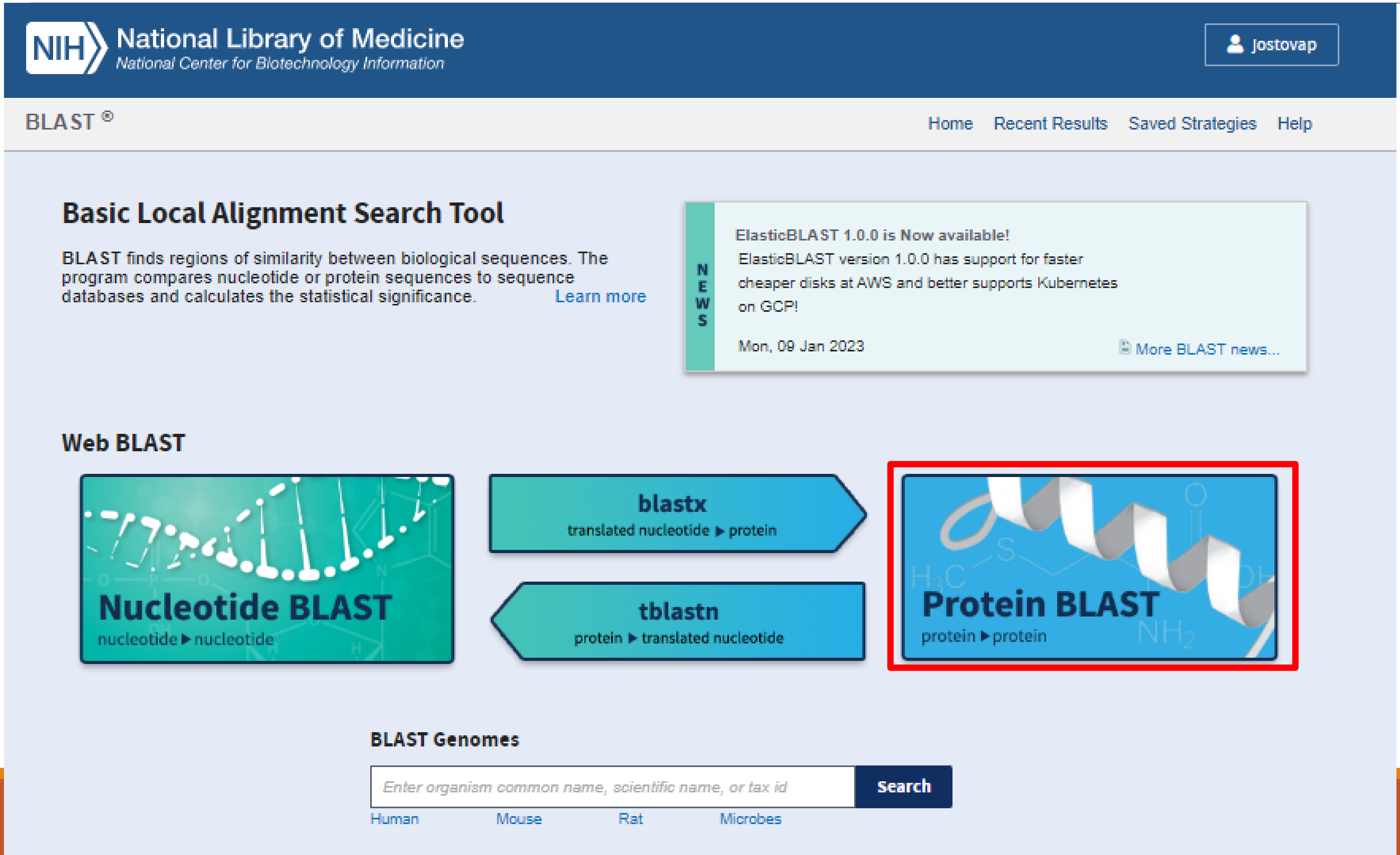
The BLOSUM62 similarity matrix

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
A	4	-2	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-1	-2	-1
B	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
C	0	-3	9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-1	-2	-4
D	-2	6	-3	6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-1	-3	2
E	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5
F	-2	-3	-2	-3	-3	6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	-1	3	-3
G	0	-1	-3	-1	-2	-3	6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-1	-3	-2
H	-2	-1	-3	-1	0	-1	-2	8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	-1	2	0
I	-1	-3	-1	-3	-3	0	-4	-3	4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1	-1	-3
K	-1	-1	-3	-1	1	-3	-2	-1	-3	5	-2	-1	0	-1	1	2	0	-1	-2	-3	-1	-2	1
L	-1	-4	-1	-4	-3	0	-4	-3	2	-2	4	2	-3	-3	-2	-2	-2	-1	1	-2	-1	-1	-3
M	-1	-3	-1	-3	-2	0	-3	-2	1	-1	2	5	-2	-2	0	-1	-1	-1	1	-1	-1	-1	-2
N	-2	1	-3	1	0	-3	0	1	-3	0	-3	-2	6	-2	0	0	1	0	-3	-4	-1	-2	0
P	-1	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7	-1	-2	-1	-1	-2	-4	-1	-3	-1
Q	-1	0	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5	1	0	-1	-2	-2	-1	-1	2
R	-1	-2	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5	-1	-1	-3	-3	-1	-2	0
S	1	0	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4	1	-2	-3	-1	-2	0
T	0	-1	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5	0	-2	-1	-2	-1
V	0	-3	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4	-3	-1	-1	-2
W	-3	-4	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	-1	2	-3
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Y	-2	-3	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	-1	7	-2
Z	-1	2	-4	2	5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-1	-2	5

„Leucine is more similar to Isoleucine than Histidine“

NCBI/BLAST

http://blast.ncbi.nlm.nih.gov/Blast.cgi



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

ElasticBLAST 1.0.0 is Now available!

ElasticBLAST version 1.0.0 has support for faster cheaper disks at AWS and better supports Kubernetes on GCP!

Mon, 09 Jan 2023

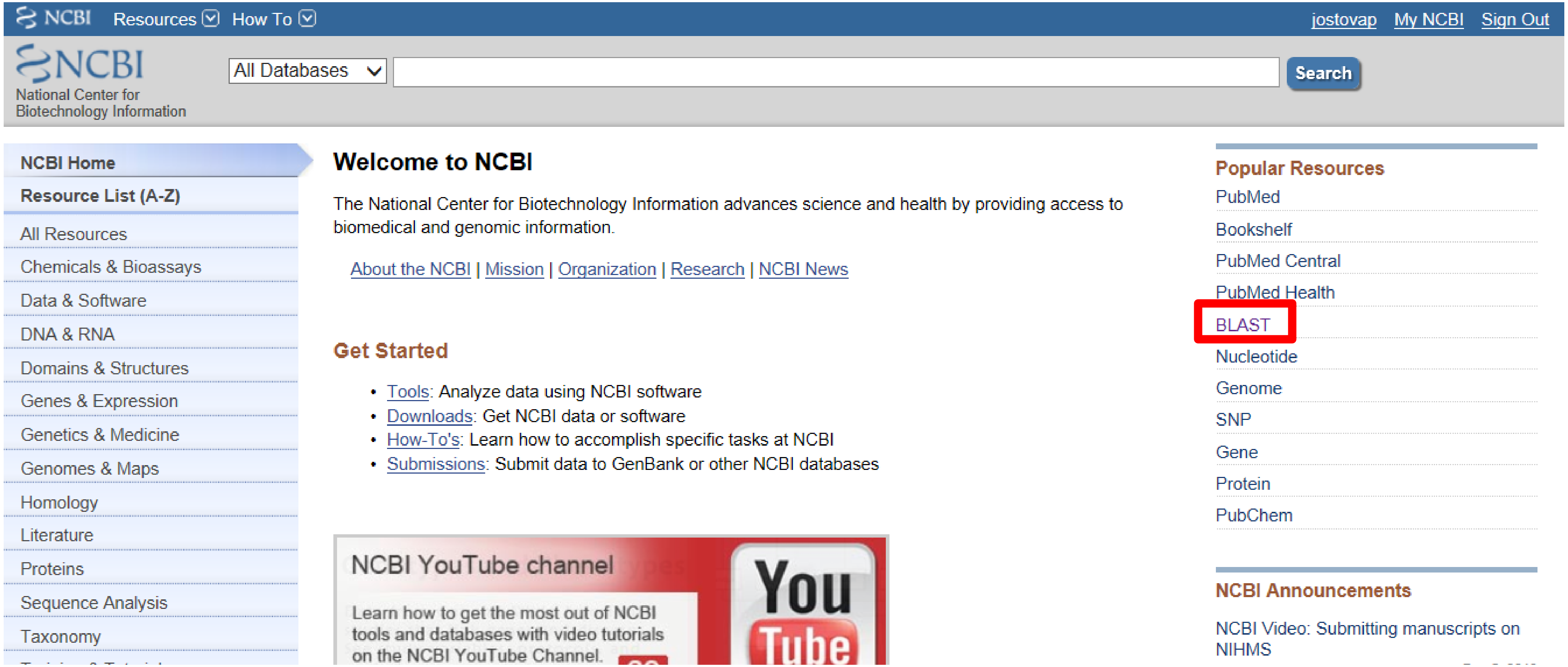
More BLAST news...

Web BLAST

BLAST Genomes

NCBI/BLAST

http://blast.ncbi.nlm.nih.gov/Blast.cgi



Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | Research | NCBI News

Get Started

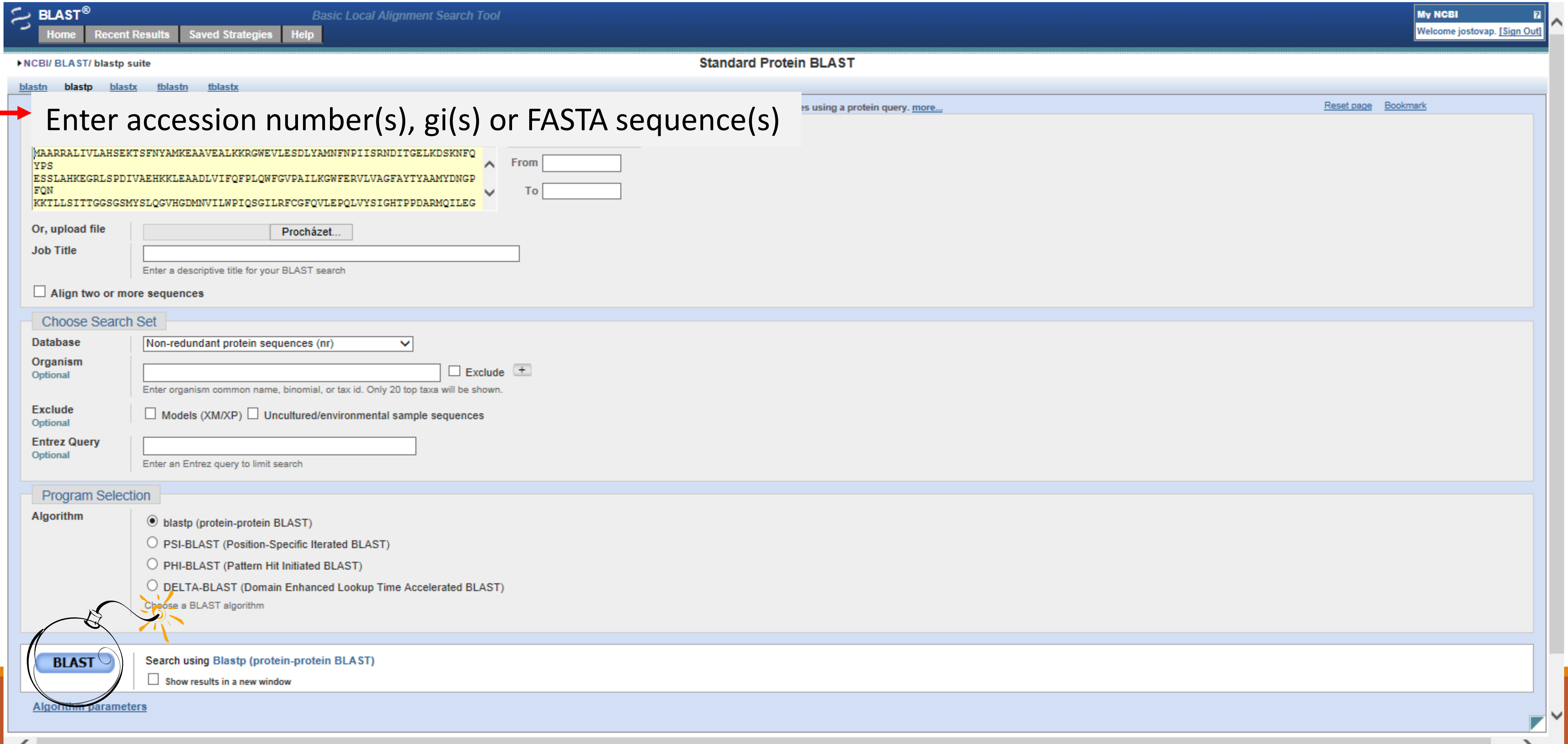
- Tools: Analyze data using NCBI software
- Downloads: Get NCBI data or software
- How-To's: Learn how to accomplish specific tasks at NCBI
- Submissions: Submit data to GenBank or other NCBI databases

NCBI YouTube channel

Learn how to get the most out of NCBI tools and databases with video tutorials on the NCBI YouTube Channel.

http://www.youtube.com/watchBLAST

NCBI/BLAST



Enter accession number(s), gi(s) or FASTA sequence(s)

Standard Protein BLAST

…es using a protein query. more...

```
MAARRALIVLAHSEKTSFNYAMKEAAVEALKKRGWEVLESDLYAMNFNPIISRNDITGELKDSKNFQ
YPS
ESSLAHKEGRLSPDIVAEHKKLEAADLVIFQFPLQWFGVPAILKGWFERVLVAGFAYTYAAMYDNGP
FQN
KKTLLSITTGGSGSMYSLQGVHGDMNVILWPIQSGILRFCGFQVLEPQLVYSIGHTPPDARMQILEG
```

From

To

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional ☐ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query Optional

Enter an Entrez query to limit search

Program Selection

Algorithm

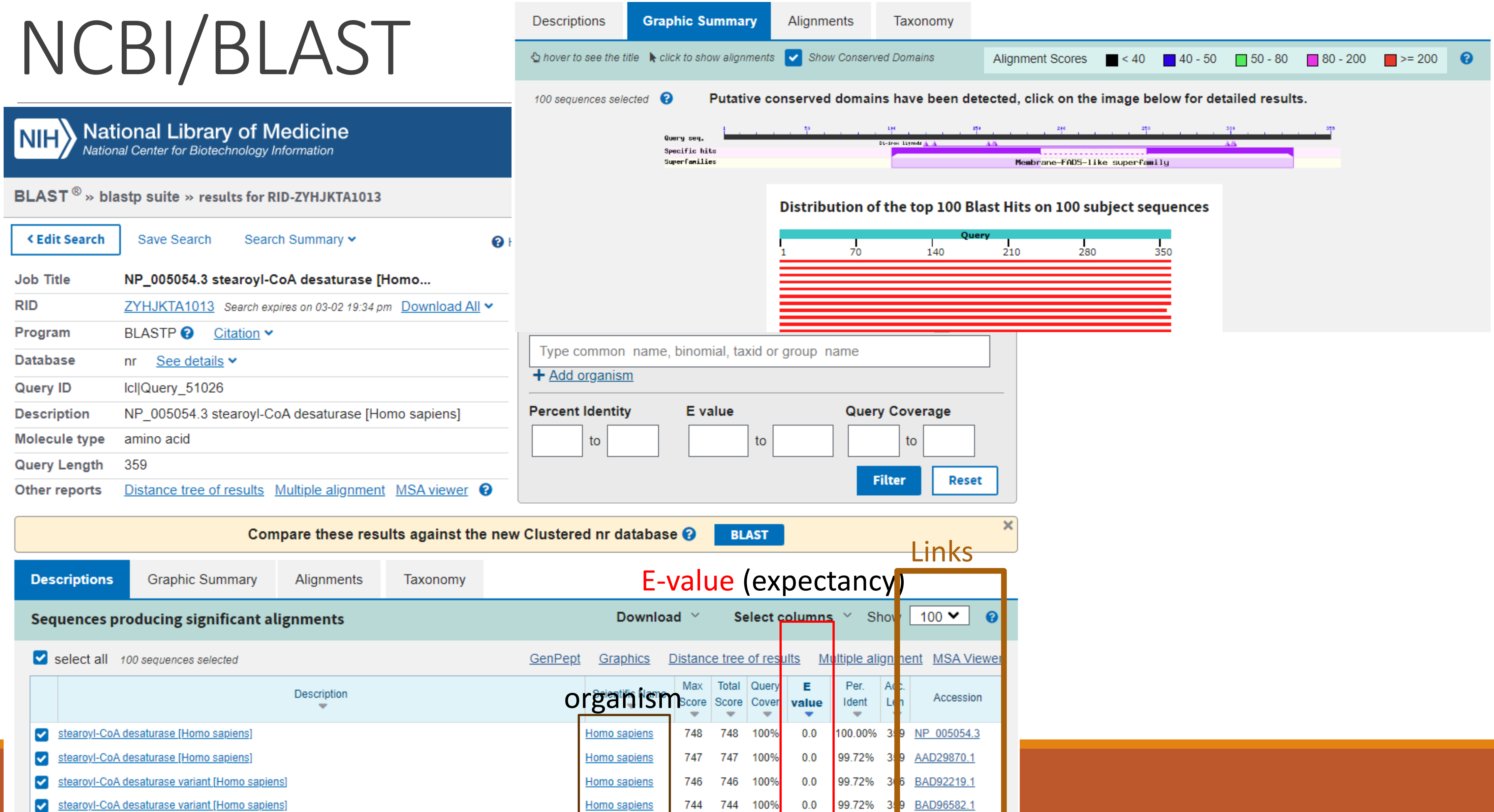
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search using Blastp (protein-protein BLAST)

☐ Show results in a new window

NCBI/BLAST



BLAST ® » **blastp suite** » **results for RID-ZYHJKTA1013**

Job Title	**NP_005054.3 stearoyl-CoA desaturase [Homo...**
RID	ZYHJKTA1013 *Search expires on 03-02 19:34 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_51026
Description	NP_005054.3 stearoyl-CoA desaturase [Homo sapiens]
Molecule type	amino acid
Query Length	359
Other reports	Distance tree of results Multiple alignment MSA viewer

100 sequences selected

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 100 Blast Hits on 100 subject sequences

Type common name, binomial, taxid or group name

Percent Identity ☐ to ☐ **E value** ☐ to ☐ **Query Coverage** ☐ to ☐

Compare these results against the new Clustered nr database **BLAST**

Sequences producing significant alignments

☑ select all *100 sequences selected*

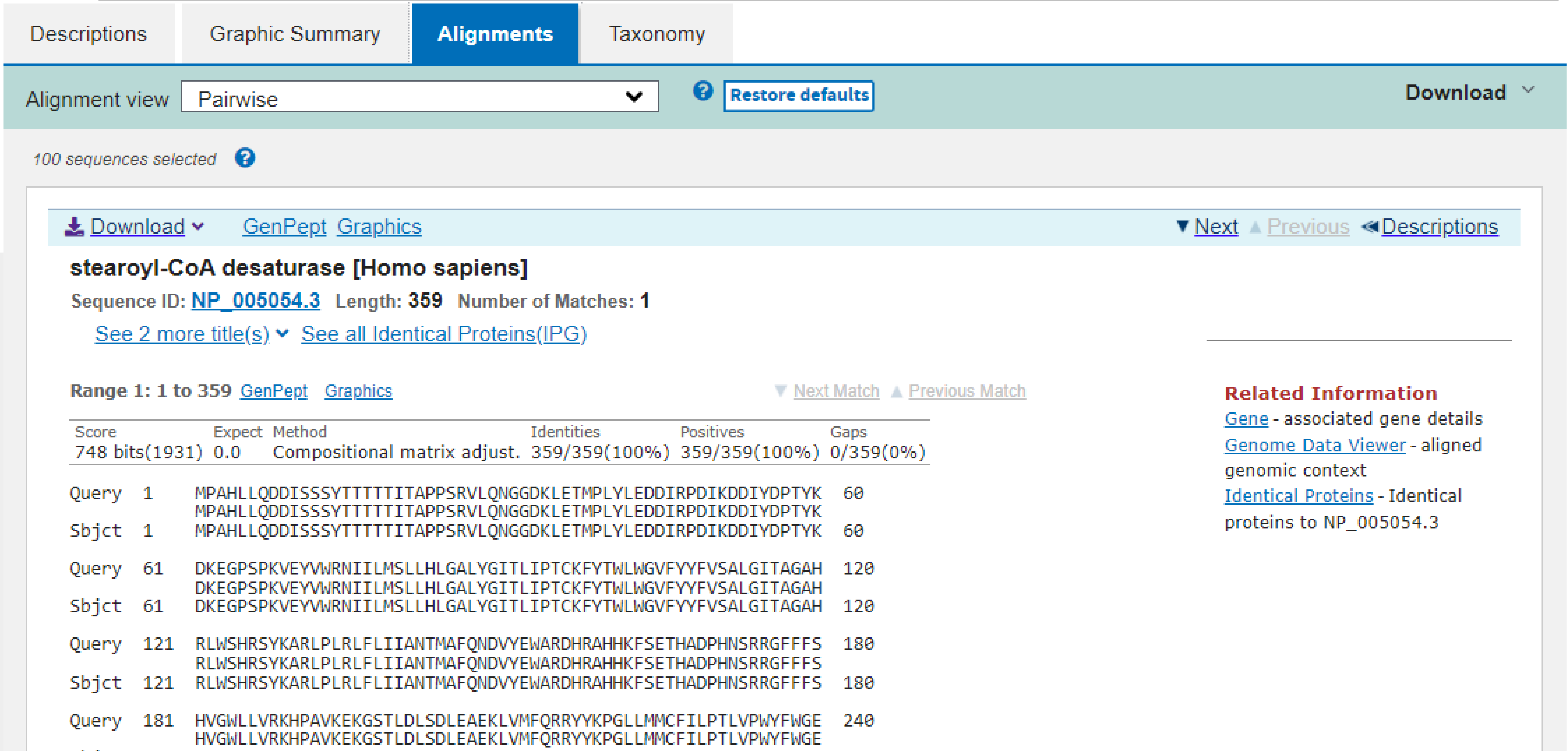
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	stearoyl-CoA desaturase [Homo sapiens]	Homo sapiens	748	748	100%	0.0	100.00%	3[illegible]9	NP_005054.3
☑	stearoyl-CoA desaturase [Homo sapiens]	Homo sapiens	747	747	100%	0.0	99.72%	3[illegible]9	AAD29870.1
☑	stearoyl-CoA desaturase variant [Homo sapiens]	Homo sapiens	746	746	100%	0.0	99.72%	3[illegible]6	BAD92219.1
☑	stearoyl-CoA desaturase variant [Homo sapiens]	Homo sapiens	744	744	100%	0.0	99.72%	3[illegible]9	BAD96582.1

Links

E-value (expectancy)

organism

NCBI/BLAST



Alignment view: Pairwise

100 sequences selected

stearoyl-CoA desaturase [Homo sapiens]

Sequence ID: NP_005054.3 **Length:** 359 **Number of Matches:** 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 359 GenPept Graphics

Next Match Previous Match

```
Score          Expect  Method                         Identities      Positives       Gaps
748 bits(1931) 0.0     Compositional matrix adjust.  359/359(100%)  359/359(100%)  0/359(0%)

Query  1    MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYK  60
            MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYK
Sbjct  1    MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYK  60

Query  61   DKEGPSPKVEYVWRNIILMSLLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAH  120
            DKEGPSPKVEYVWRNIILMSLLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAH
Sbjct  61   DKEGPSPKVEYVWRNIILMSLLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAH  120

Query  121  RLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFS  180
            RLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFS
Sbjct  121  RLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFS  180

Query  181  HVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGE  240
            HVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGE
```

Related Information

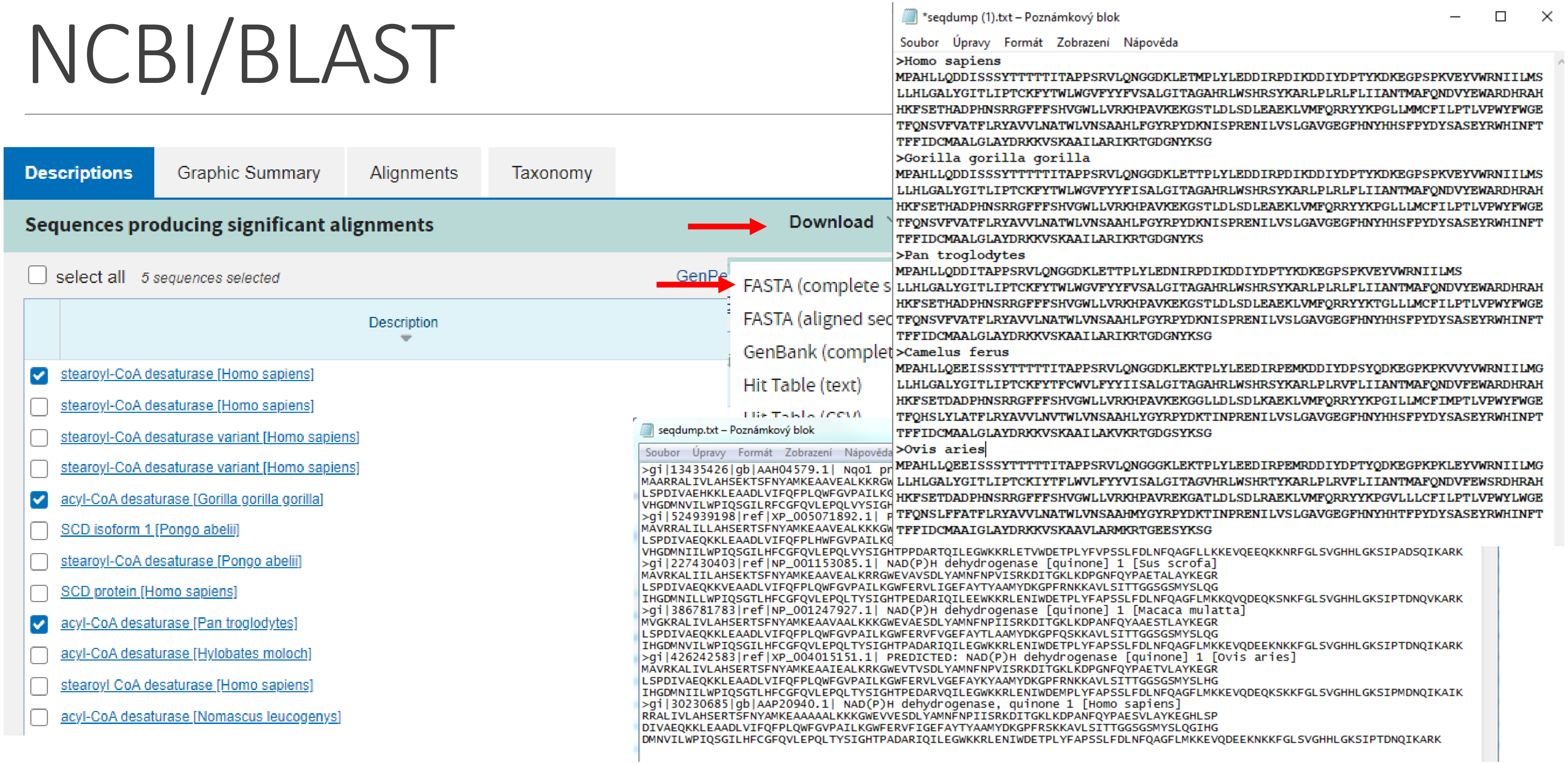
Gene - associated gene details

Genome Data Viewer - aligned genomic context

Identical Proteins - Identical proteins to NP_005054.3

→change sequences (FASTA) names into organism only

NCBI/BLAST



☐ select all *5 sequences selected*

√ ☑ stearoyl-CoA desaturase [Homo sapiens]
☐ stearoyl-CoA desaturase [Homo sapiens]
☐ stearoyl-CoA desaturase variant [Homo sapiens]
☐ stearoyl-CoA desaturase variant [Homo sapiens]
√ ☑ acyl-CoA desaturase [Gorilla gorilla gorilla]
☐ SCD isoform 1 [Pongo abelii]
☐ stearoyl-CoA desaturase [Pongo abelii]
☐ SCD protein [Homo sapiens]
√ ☑ acyl-CoA desaturase [Pan troglodytes]
☐ acyl-CoA desaturase [Hylobates moloch]
☐ stearoyl CoA desaturase [Homo sapiens]
☐ acyl-CoA desaturase [Nomascus leucogenys]

```
>Homo sapiens
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVEYVWRNIILMS
LLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
HKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGE
TFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFT
TFFIDCMAALGLAYDRKKVSKAAILARIKRTGDGNYKSG
>Gorilla gorilla gorilla
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETTPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVEYVWRNIILMS
LLHLGALYGITLIPTCKFYTWLWGVFYYFISALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
HKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLLMCFILPTLVPWYFWGE
TFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFT
TFFIDCMAALGLAYDRKKVSKAAILARIKRTGDGNYKS
>Pan troglodytes
MPAHLLQDDITAPPSRVLQNGGDKLETTPLYLEDNIRPDIKDDIYDPTYKDKEGPSPKVEYVWRNIILMS
LLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
HKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKTGLLLMCFILPTLVPWYFWGE
TFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFT
TFFIDCMAALGLAYDRKKVSKAAILARIKRTGDGNYKSG
>Camelus ferus
MPAHLLQEEISSSYTTTTTITAPPSRVLQNGGDKLEKTPLYLEEDIRPEMKDDIYDPSYQDKEGPKPKVVYVWRNIILMG
LLHLGALYGITLIPTCKFYTFCWVLFYYIISALGITAGAHRLWSHRSYKARLPLRVFLIIANTMAFQNDVFEWARDHRAH
HKFSETDADPHNSRRGFFFSHVGWLLVRKHPAVKEKGGLLDLSDLKAEKLVMFQRRYYKPGILLMCFIMPTLVPWYFWGE
TFQHSLYLATFLRYAVVLNVTWLVNSAAHLYGYRPYDKTINPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINPT
TFFIDCMAALGLAYDRKKVSKAAILAKVKRTGDGSYKSG
>Ovis aries
MPAHLLQEEISSSYTTTTTITAPPSRVLQNGGGKLEKTPLYLEEDIRPEMRDDIYDPTYQDKEGPKPKLEYVWRNIILMG
LLHLGALYGITLIPTCKIYTFLWVLFYYVISALGITAGVHRLWSHRTYKARLPLRVFLIIANTMAFQNDVFEWSRDHRAH
HKFSETDADPHNSRRGFFFSHVGWLLVRKHPAVREKGATLDLSDLRAEKLVMFQRRYYKPGVLLLCFILPTLVPWYLWGE
TFQNSLFFATFLRYAVVLNATWLVNSAAHMYGYRPYDKTINPRENILVSLGAVGEGFHNYHHTFPYDYSASEYRWHINFT
TFFIDCMAAIGLAYDRKKVSKAAVLARMKRTGEESYKSG
```

```
>gi|13435426|gb|AAH04579.1| Nqo1 pr
MAARRALIVLAHSEKTSFNYAMKEAAVEALKKRGW
LSPDIVAEHKKLEAADLVIFQFPLQWFGVPAILKG
VHGDMNVILWPIQSGILRFCGFQVLEPQLVYSIGH
>gi|524939198|ref|XP_005071892.1| P
MAVRRALILLAHSERTSFNYAMKEAAVEALKKKGW
LSPDIVAEQKKLEAADLVIFQFPLHWFGVPAILKG
VHGDMNIILWPIQSGILHFCGFQVLEPQLVYSIGHTPPDARTQILEGWKKRLETVWDETPLYFVPSSLFDLNFQAGFLLKKEVQEEQKKNRFGLSVGHHLGKSIPADSQIKARK
>gi|227430403|ref|NP_001153085.1| NAD(P)H dehydrogenase [quinone] 1 [Sus scrofa]
MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGR
LSPDIVAEQKKVEAADLVIFQFPLQWFGVPAILKGWFERVLIGEFAYTYAAMYDKGPFRNKKAVLSITTGGSGSMYSLQG
IHGDMNILLWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKQVQDEQKSNKFGLSVGHHLGKSIPTDNQVKARK
>gi|386781783|ref|NP_001247927.1| NAD(P)H dehydrogenase [quinone] 1 [Macaca mulatta]
MVGKRALIVLAHSERTSFNYAMKEAAVAALKKKGWEVAESDLYAMNFNPIISRKDITGKLKDPANFQYAAESTLAYKEGR
LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFVGEFAYTLAAMYDKGPFQSKKAVLSITTGGSGSMYSLQG
IHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK
>gi|426242583|ref|XP_004015151.1| PREDICTED: NAD(P)H dehydrogenase [quinone] 1 [Ovis aries]
MAVRKALIVLAHSERTSFNYAMKEAAIEALKRKGWEVTVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETVLAYKEGR
LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVLVGEFAYKYAAMYDKGPFRNKKAVLSITTGGSGSMYSLHG
IHGDMNIILWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARVQILEGWKKRLENIWDEMPLYFAPSSLFDLNFQAGFLMKKEVQDEQKSKKFGLSVGHHLGKSIPMDNQIKAIK
>gi|30230685|gb|AAP20940.1| NAD(P)H dehydrogenase, quinone 1 [Homo sapiens]
RRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSP
DIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHG
DMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK
```

Practical part

Try BLAST.

Download five similar sequences from different organisms.

„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**

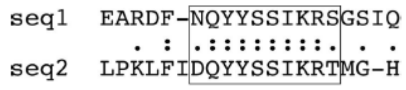
…

Pairwise alignment

Global alignment – aligns full length sequence

Local alignment – aligns part of the sequences that fit best

(eg similar domains comparison, repetitive sequences…)

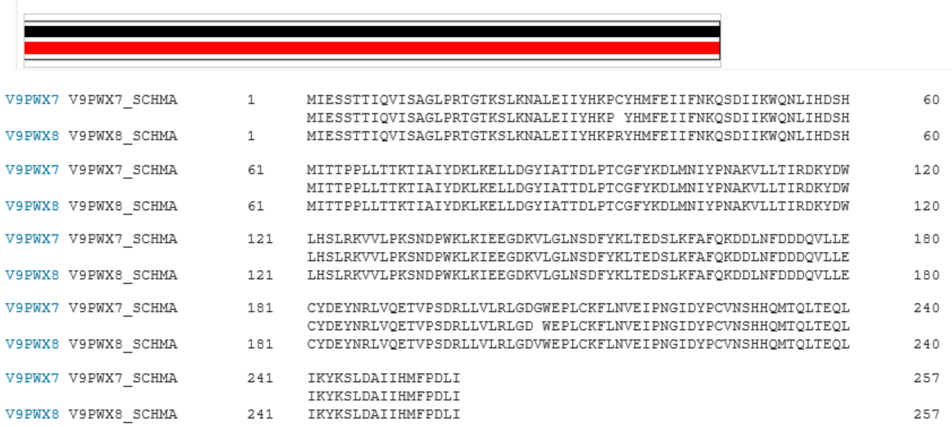


```
seq1  EARDF-NQYYSSIKRSGSIQ
          . : .:::::::::.  .  .
seq2  LPKLFIDQYYSSIKRTMG-H
```

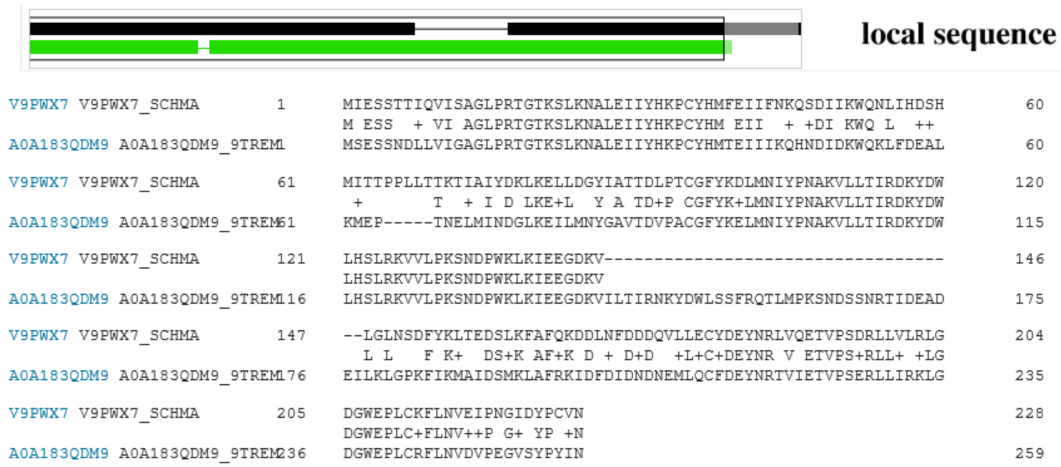
global sequence alignment

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

local sequence alignment



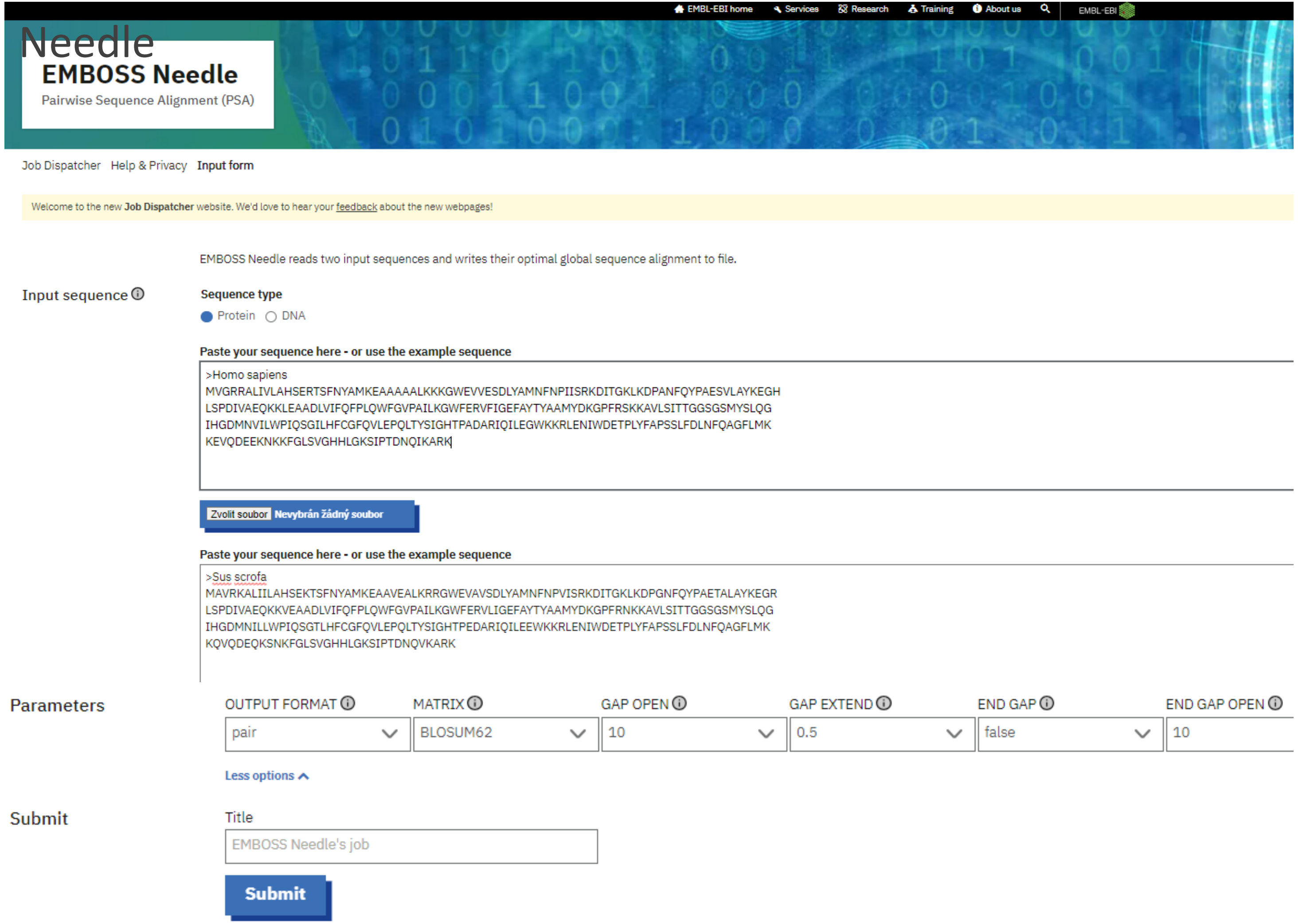
```
V9PWX7 V9PWX7_SCHMA       1     MIESSTTIQVISAGLPRTGTKSLKNALEIIYHKPCYHMFEIIFNKQSDIIKWQNLIHDSH      60
                                MIESSTTIQVISAGLPRTGTKSLKNALEIIYHKP YHMFEIIFNKQSDIIKWQNLIHDSH
V9PWX8 V9PWX8_SCHMA       1     MIESSTTIQVISAGLPRTGTKSLKNALEIIYHKPRYHMFEIIFNKQSDIIKWQNLIHDSH      60

V9PWX7 V9PWX7_SCHMA       61    MITTPPLLTTKTIAIYDKLKELLDGYIATTDLPTCGFYKDLMNIYPNAKVLLTIRDKYDW      120
                                MITTPPLLTTKTIAIYDKLKELLDGYIATTDLPTCGFYKDLMNIYPNAKVLLTIRDKYDW
V9PWX8 V9PWX8_SCHMA       61    MITTPPLLTTKTIAIYDKLKELLDGYIATTDLPTCGFYKDLMNIYPNAKVLLTIRDKYDW      120

V9PWX7 V9PWX7_SCHMA       121   LHSLRKVVLPKSNDPWKLKIEEGDKVLGLNSDFYKLTEDSLKFAFQKDDLNFDDDQVLLE      180
                                LHSLRKVVLPKSNDPWKLKIEEGDKVLGLNSDFYKLTEDSLKFAFQKDDLNFDDDQVLLE
V9PWX8 V9PWX8_SCHMA       121   LHSLRKVVLPKSNDPWKLKIEEGDKVLGLNSDFYKLTEDSLKFAFQKDDLNFDDDQVLLE      180

V9PWX7 V9PWX7_SCHMA       181   CYDEYNRLVQETVPSDRLLVLRLGDGWEPLCKFLNVEIPNGIDYPCVNSHHQMTQLTEQL      240
                                CYDEYNRLVQETVPSDRLLVLRLGD WEPLCKFLNVEIPNGIDYPCVNSHHQMTQLTEQL
V9PWX8 V9PWX8_SCHMA       181   CYDEYNRLVQETVPSDRLLVLRLGDVWEPLCKFLNVEIPNGIDYPCVNSHHQMTQLTEQL      240

V9PWX7 V9PWX7_SCHMA       241   IKYKSLDAIIHMFPDLI                                                 257
                                IKYKSLDAIIHMFPDLI
V9PWX8 V9PWX8_SCHMA       241   IKYKSLDAIIHMFPDLI                                                 257
```



```
V9PWX7 V9PWX7_SCHMA          1      MIESSTTIQVISAGLPRTGTKSLKNALEIIYHKPCYHMFEIIFNKQSDIIKWQNLIHDSH      60
                                    M ESS   + VI AGLPRTGTKSLKNALEIIYHKPCYHM EII   + +DI KWQ L   ++
A0A183QDM9 A0A183QDM9_9TREM1        MSESSNDLLVIGAGLPRTGTKSLKNALEIIYHKPCYHMTEIIIKQHNDIDKWQKLFDEAL      60

V9PWX7 V9PWX7_SCHMA          61     MITTPPLLTTKTIAIYDKLKELLDGYIATTDLPTCGFYKDLMNIYPNAKVLLTIRDKYDW      120
                                     +        T   + I  D LKE+L   Y A TD+P CGFYK+LMNIYPNAKVLLTIRDKYDW
A0A183QDM9 A0A183QDM9_9TREM61       KMEP-----TNELMINDGLKEILMNYGAVTDVPACGFYKELMNIYPNAKVLLTIRDKYDW      115

V9PWX7 V9PWX7_SCHMA          121    LHSLRKVVLPKSNDPWKLKIEEGDKV----------------------------------      146
                                    LHSLRKVVLPKSNDPWKLKIEEGDKV
A0A183QDM9 A0A183QDM9_9TREM116      LHSLRKVVLPKSNDPWKLKIEEGDKVILTIRNKYDWLSSFRQTLMPKSNDSSNRTIDEAD      175

V9PWX7 V9PWX7_SCHMA          147    --LGLNSDFYKLTEDSLKFAFQKDDLNFDDDQVLLECYDEYNRLVQETVPSDRLLVLRLG      204
                                      L L    F K+   DS+K AF+K D + D+D   +L+C+DEYNR V ETVPS+RLL+ +LG
A0A183QDM9 A0A183QDM9_9TREM176      EILKLGPKFIKMAIDSMKLAFRKIDFDIDNDNEMLQCFDEYNRTVIETVPSERLLIRKLG      235

V9PWX7 V9PWX7_SCHMA          205    DGWEPLCKFLNVEIPNGIDYPCVN                                          228
                                    DGWEPLC+FLNV++P G+ YP +N
A0A183QDM9 A0A183QDM9_9TREM236      DGWEPLCRFLNVDVPEGVSYPYIN                                          259
```

Pairwise alignment- Global

Needle



EMBOSS Needle

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!

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
MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGH
LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQG
IHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK
KEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK
```

Paste your sequence here - or use the example sequence

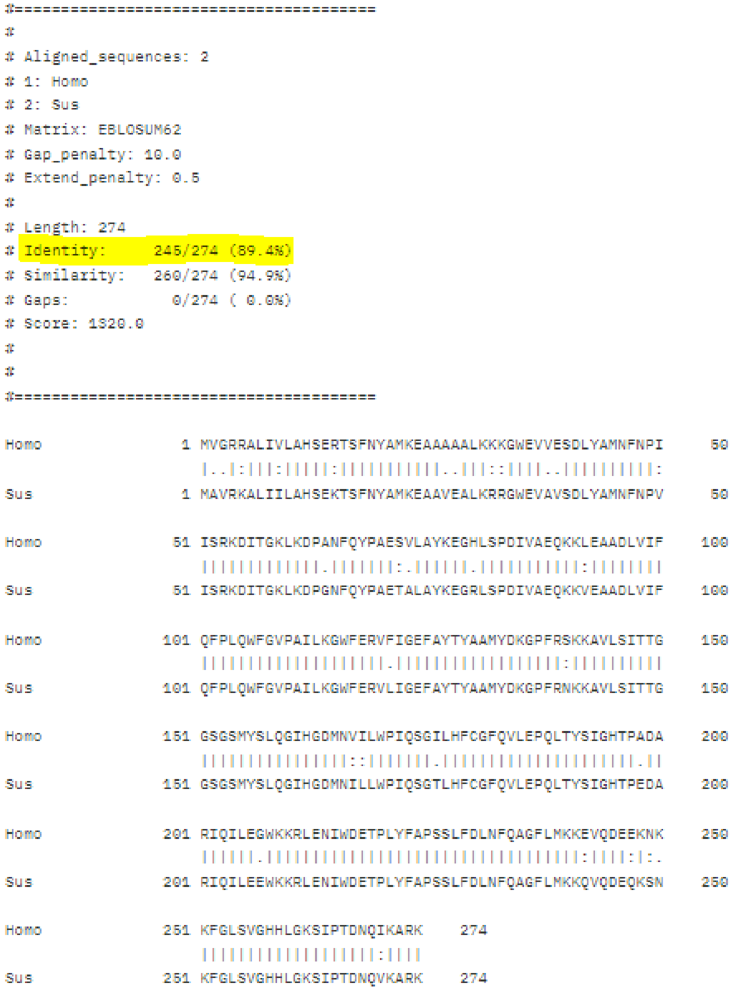
```
>Sus scrofa
MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGR
LSPDIVAEQKKVEAADLVIFQFPLQWFGVPAILKGWFERVLIGEFAYTYAAMYDKGPFRNKKAVLSITTGGSGSMYSLQG
IHGDMNILLWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK
KQVQDEQKSNKFGLSVGHHLGKSIPTDNQVKARK
```

Parameters

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

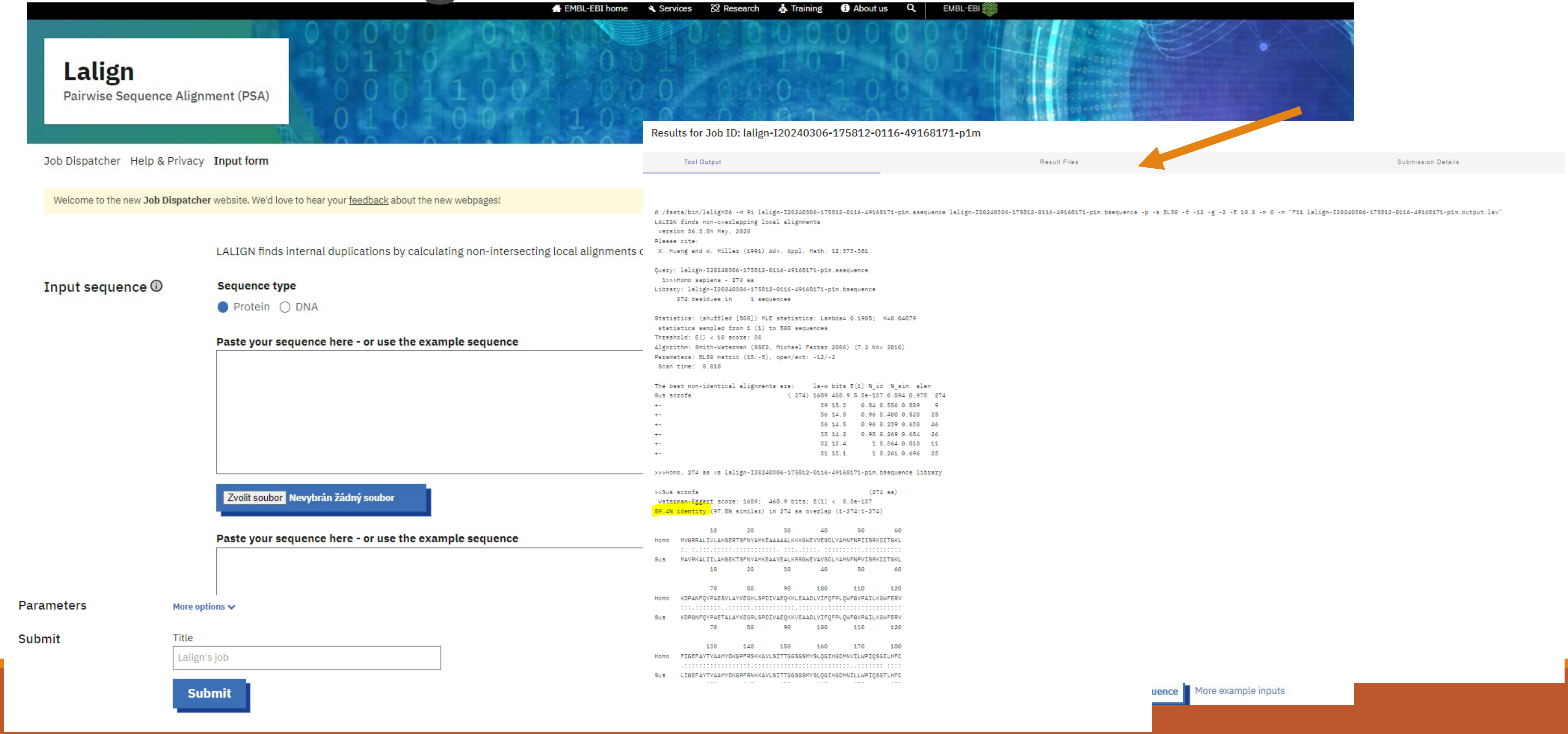
Submit

Title: EMBOSS Needle's job



```
#=======================================
#
# 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 MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPI     50
                     |..|:|||:|||||:|||||||||||..|||::||||..||||||||||:
Sus                1 MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPV     50

Homo              51 ISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIF    100
                     |||||||||||||.|||||||:.||||||.|||||||||||:||||||||
Sus               51 ISRKDITGKLKDPGNFQYPAETALAYKEGRLSPDIVAEQKKVEAADLVIF    100

Homo             101 QFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRSKKAVLSITTG    150
                     ||||||||||||||||||||.||||||||||||||||||:||||||||||
Sus              101 QFPLQWFGVPAILKGWFERVLIGEFAYTYAAMYDKGPFRNKKAVLSITTG    150

Homo             151 GSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADA    200
                     ||||||||||||||||::|||||||.|||||||||||||||||||||.||
Sus              151 GSGSMYSLQGIHGDMNILLWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDA    200

Homo             201 RIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNK    250
                     ||||||.||||||||||||||||||||||||||||||||||:||||:|:.
Sus              201 RIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKQVQDEQKSN    250

Homo             251 KFGLSVGHHLGKSIPTDNQIKARK    274
                     |||||||||||||||||||:||||
Sus              251 KFGLSVGHHLGKSIPTDNQVKARK    274
```

Pairwise alignment- Local



Lalign

Pairwise Sequence Alignment (PSA)

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 c

Input sequence

Sequence type

● Protein ○ DNA

Paste your sequence here - or use the example sequence

Paste your sequence here - or use the example sequence

Parameters

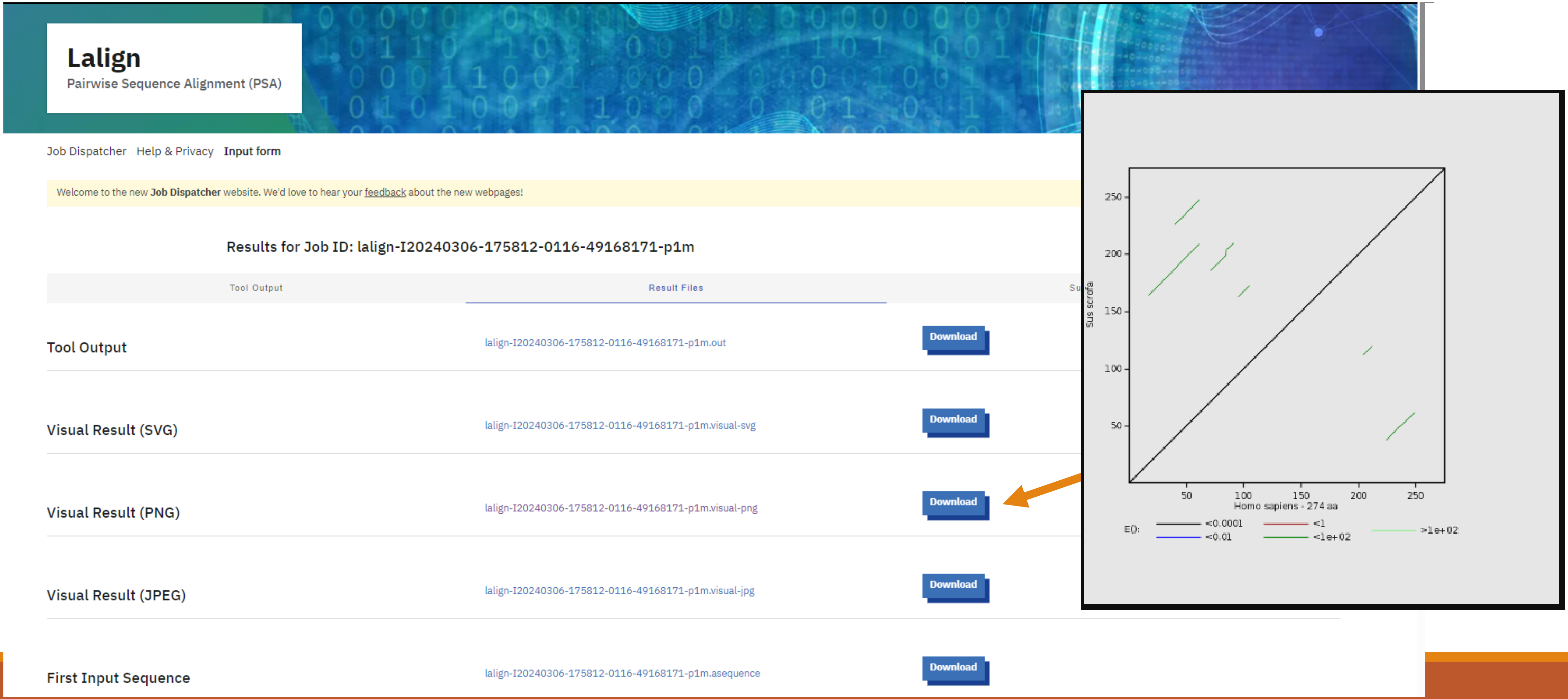
Submit

Title: Lalign's job

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

```
# /fasta/bin/lalign36 -m 9i 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 "F11 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: 30
Algorithm: Smith-Waterman (SSE2, Michael Farrar 2006) (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) %_id  %_sim  alen
Sus scrofa                             ( 274) 1689 468.9 5.3e-137 0.894 0.978  274
+-                                        39 15.3    0.84 0.556 0.889    9
+-                                        36 14.5    0.96 0.400 0.520   25
+-                                        36 14.5    0.96 0.239 0.630   46
+-                                        35 14.2    0.98 0.269 0.654   26
+-                                        32 13.4       1 0.364 0.818   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)
 Waterman-Eggert score: 1689;  468.9 bits; E(1) <  5.3e-137
89.4% identity (97.8% similar) in 274 aa overlap (1-274:1-274)

               10        20        30        40        50        60
Homo    MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKL
        :. :.:::.:::::.:::::::::::. :::..::::. ::::::::::.::::::::::
Sus     MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKL
               10        20        30        40        50        60

               70        80        90       100       110       120
Homo    KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV
        :::.:::::::..::::::.:::::::::::.::::::::::::::::::::::::::::
Sus     KDPGNFQYPAETALAYKEGRLSPDIVAEQKKVEAADLVIFQFPLQWFGVPAILKGWFERV
               70        80        90       100       110       120

              130       140       150       160       170       180
Homo    FIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFC
        .::::::::::::::::::.::::::::::::::::::::::::::..::::::: ::::
Sus     LIGEFAYTYAAMYDKGPFRNKKAVLSITTGGSGSMYSLQGIHGDMNILLWPIQSGTLHFC
```

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	Su
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

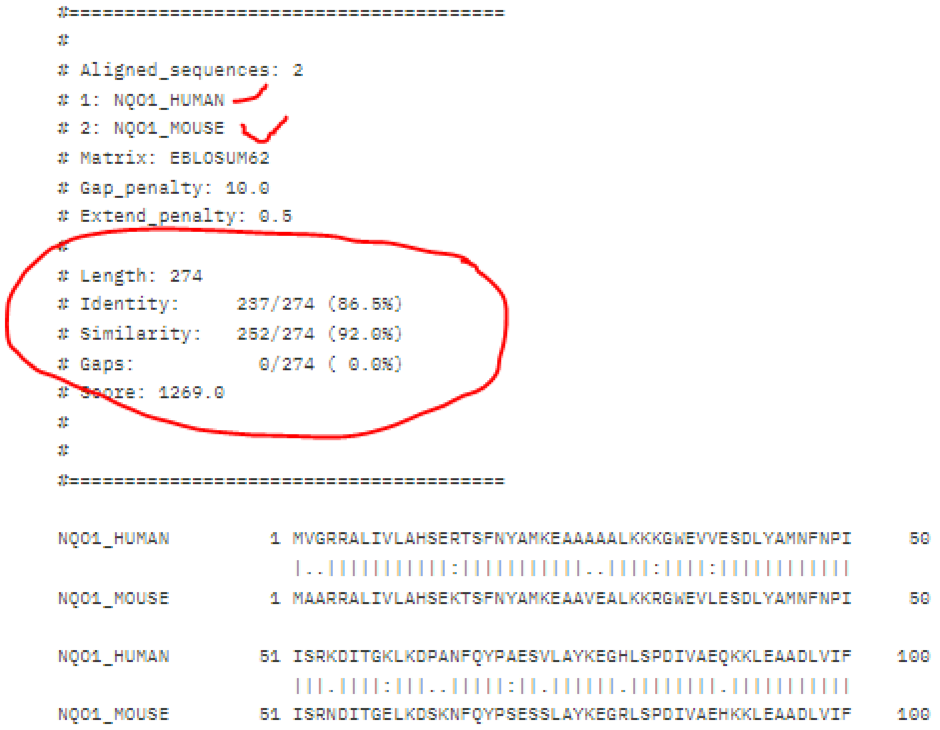
Try pairwise alignment.
(global and local)

Hw:Compare „your“ sequence (human) with sequence of the same protein from mouse (Mus musculus).

How similar are these proteins?

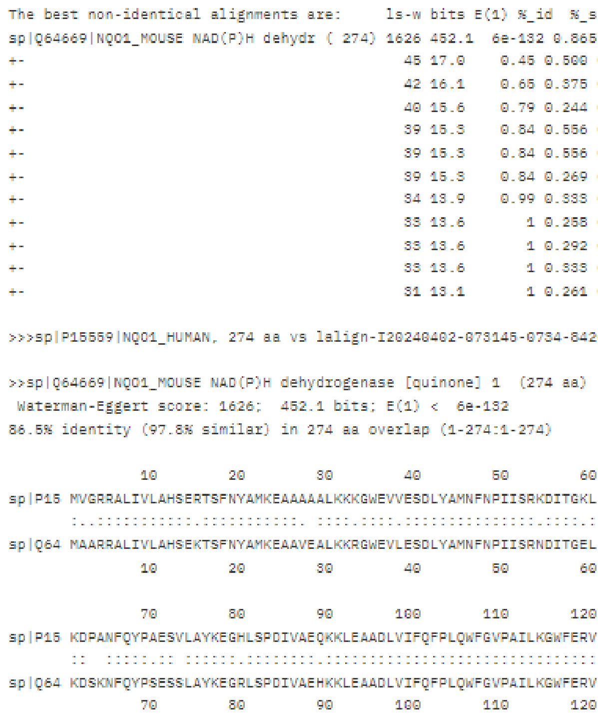
Practical part - results

Global a. (Needle)

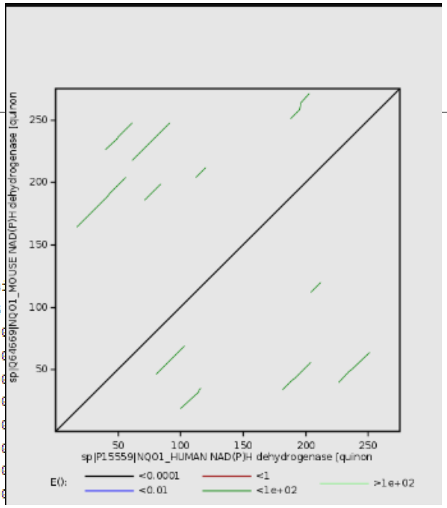


```
#=======================================
#
# Aligned_sequences: 2
# 1: NQO1_HUMAN
# 2: NQO1_MOUSE
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 274
# Identity:     237/274 (86.5%)
# Similarity:   252/274 (92.0%)
# Gaps:           0/274 ( 0.0%)
# Score: 1269.0
#
#
#=======================================

NQO1_HUMAN         1 MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPI     50
                     |..|||||||||||:|||||||||||..||||:||||:|||||||||||||
NQO1_MOUSE         1 MAARRALIVLAHSEKTSFNYAMKEAAVEALKKRGWEVLESDLYAMNFNPI     50

NQO1_HUMAN        51 ISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIF    100
                     |||.||||:|||..|||||:||.||||||.||||||||.||||||||||||
NQO1_MOUSE        51 ISRNDITGELKDSKNFQYPSESSLAYKEGRLSPDIVAEHKKLEAADLVIF    100
```

Local a. (Laling)



```
The best non-identical alignments are:     ls-w bits E(1) %_id  %_si
sp|Q64669|NQO1_MOUSE NAD(P)H dehydr ( 274) 1626 452.1  6e-132 0.865
+-                                           45 17.0    0.45 0.500
+-                                           42 16.1    0.65 0.375
+-                                           40 15.6    0.79 0.244
+-                                           39 15.3    0.84 0.556
+-                                           39 15.3    0.84 0.556
+-                                           39 15.3    0.84 0.269
+-                                           34 13.9    0.99 0.333
+-                                           33 13.6       1 0.258
+-                                           33 13.6       1 0.292 0.542   24
+-                                           33 13.6       1 0.333 0.667   18
+-                                           31 13.1       1 0.261 0.696   23

>>>sp|P15559|NQO1_HUMAN, 274 aa vs lalign-I20240402-073145-0734-84207820-p1m.bsequence library

>>sp|Q64669|NQO1_MOUSE NAD(P)H dehydrogenase [quinone] 1  (274 aa)
 Waterman-Eggert score: 1626;  452.1 bits; E(1) <  6e-132
86.5% identity (97.8% similar) in 274 aa overlap (1-274:1-274)

               10        20        30        40        50        60
sp|P15 MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKL
       :..:::::::::::.:::::::::::. ::::.::::.:::::::::::::::.:::::.:
sp|Q64 MAARRALIVLAHSEKTSFNYAMKEAAVEALKKRGWEVLESDLYAMNFNPIISRNDITGEL
               10        20        30        40        50        60

               70        80        90       100       110       120
sp|P15 KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV
       ::  ::::::.:: :::::::.::::::::.:::::::::::::::::::::::::::::::
sp|Q64 KDSKNFQYPSESSLAYKEGRLSPDIVAEHKKLEAADLVIFQFPLQWFGVPAILKGWFERV
               70        80        90       100       110       120
```



Multiple sequence alignment (MSA)

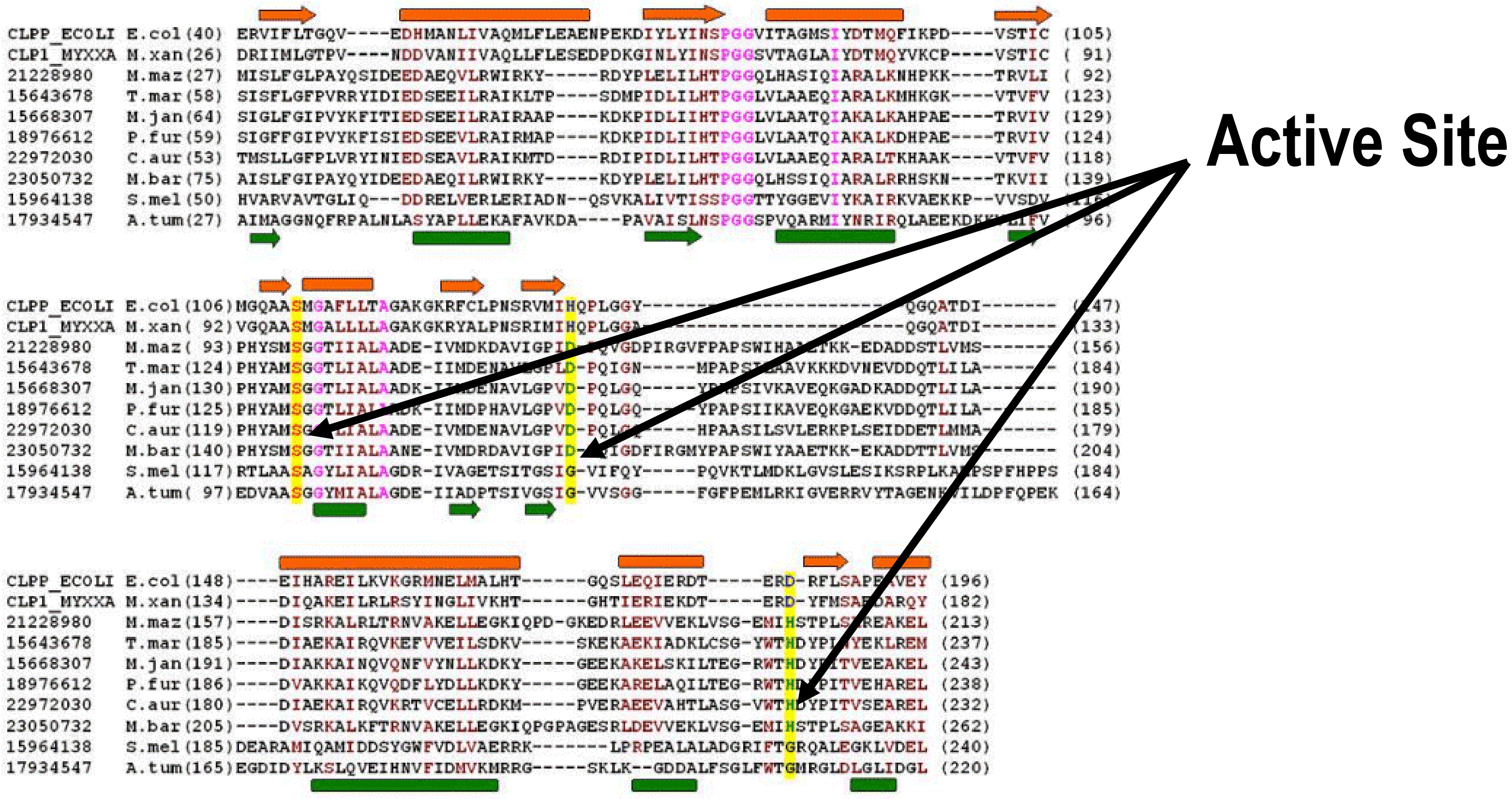
=The alignment of more than two sequences

The goal of MSA is twofold:

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

An Example of Conserved Positions:

(The Serine Proteases Active Site)



```
CLPP_ECOLI E.col(40) ERVIFLTGQV----EDHMANLIVAQMLFLEAENPEKDIYLYINSPGGVITAGMSIYDTMQFIKPD----VSTIC (105)
CLP1_MYXXA M.xan(26) DRIIMLGTPV----NDDVANIIVAQLLFLESEDPDKGINLYINSPGGSVTAGLAIYDTMQYVKCP----VSTIC ( 91)
21228980   M.maz(27) MISLFGLPAYQSIDEEDAEQVLRWIRKY-----RDYPLELILHTPGGQLHASIQIARALKNHPKK----TRVLI ( 92)
15643678   T.mar(58) SISFLGFPVRRYIDIEDSEEILRAIKLTP----SDMPIDLILHTPGGLVLAAEQIARALKMHKGK----VTVFV (123)
15668307   M.jan(64) SIGLFGIPVYKFITIEDSEEILRAIRAAP----KDKPIDLIIHTPGGLVLAATQIAKALKAHPAE----TRVIV (129)
18976612   P.fur(59) SIGFFGIPVYKFISIEDSEEVLRAIRMAP----KDKPIDLIIHTPGGLVLAATQIAKALKDHPAE----TRVIV (124)
22972030   C.aur(53) TMSLLGFPLVRYINIEDSEAVLRAIKMTD----RDIPIDLILHTPGGLVLAAEQIARALTKHAAK----VTVFV (118)
23050732   M.bar(75) AISLFGIPAYQYIDEEDAEQILRWIRKY-----KDYPLELILHTPGGQLHSSIQIARALRRHSKN----TKVII (139)
15964138   S.mel(50) HVARVAVTGLIQ---DDRELVERLERIADN--QSVKALIVTISSPGGTTYGGEVIYKAIRKVAEKKP--VVSDV (116)
17934547   A.tum(27) AIMAGGNQFRPALNLASYAPLLEKAFAVKDA----PAVAISLNSPGGSPVQARMIYNRIRQLAEEKDKKVLIFV ( 96)

CLPP_ECOLI E.col(106)MGQAASMGAFLLTAGAKGKRFCLPNSRVMIHQPLGGY----------------------QGQATDI-------- (147)
CLP1_MYXXA M.xan( 92)VGQAASMGALLLLAGAKGKRYALPNSRIMIHQPLGGA----------------------QGQATDI-------- (133)
21228980   M.maz( 93)PHYSMSGGTIIALAADE-IVMDKDAVIGPID-PQVGDPIRGVFPAPSWIHAAETKK-EDADDSTLVMS-------- (156)
15643678   T.mar(124)PHYAMSGGTLIALAADE-IIMDENAVLGPLD-PQIGN-----MPAPSIAAAVKKKDVNEVDDQTLILA-------- (184)
15668307   M.jan(130)PHYAMSGGTLIALAADK-IIMDENAVLGPVD-PQLGQ-----YPAPSIVKAVEQKGADKADDQTLILA-------- (190)
18976612   P.fur(125)PHYAMSGGTLIALAADK-IIMDPHAVLGPVD-PQLGQ-----YPAPSIIKAVEQKGAEKVDDQTLILA-------- (185)
22972030   C.aur(119)PHYAMSGGTLIALAADE-IVMDENAVLGPVD-PQLGQ-----HPAASILSVLERKPLSEIDDETLMMA-------- (179)
23050732   M.bar(140)PHYSMSGGTIIALAANE-IVMDRDAVIGPID-PQIGDFIRGMYPAPSWIYAAETKK-EKADDTTLVMS-------- (204)
15964138   S.mel(117)RTLAASAGYLIALAGDR-IVAGETSITGSIG-VIFQY-----PQVKTLMDKLGVSLESIKSRPLKAAPSPFHPPS (184)
17934547   A.tum( 97)EDVAASGGYMIALAGDE-IIADPTSIVGSIG-VVSGG-----FGFPEMLRKIGVERRVYTAGENKVILDPFQPEK (164)

CLPP_ECOLI E.col(148)----EIHAREILKVKGRMNELMALHT------GQSLEQIERDT-----ERD-RFLSAPEAVEY (196)
CLP1_MYXXA M.xan(134)----DIQAKEILRLRSYINGLIVKHT------GHTIERIEKDT-----ERD-YFMSAPDARQY (182)
21228980   M.maz(157)----DISRKALRLTRNVAKELLEGKIQPD-GKEDRLEEVVEKLVSG-EMIHSTPLSAREAKEL (213)
15643678   T.mar(185)----DIAEKAIRQVKEFVVEILSDKV-----SKEKAEKIADKLCSG-YWTHDYPINYEKLREM (237)
15668307   M.jan(191)----DIAKKAINQVQNFVYNLLKDKY-----GEEKAKELSKILTEG-RWTHDYPITVEEAKEL (243)
18976612   P.fur(186)----DVAKKAIKQVQDFLYDLLKDKY-----GEEKARELAQILTEG-RWTHDYPITVEHAREL (238)
22972030   C.aur(180)----DIAEKAIRQVKRTVCELLRDKM-----PVERAEEVAHTLASG-VWTHDYPITVSEAREL (232)
23050732   M.bar(205)----DVSRKALKFTRNVAKELLEGKIQPGPAGESRLDEVVEKLVSG-EMIHSTPLSAGEAKKI (262)
15964138   S.mel(185)DEARAMIQAMIDDSYGWFVDLVAERRK-------LPRPEALALADGRIFTGRQALEGKLVDEL (240)
17934547   A.tum(165)EGDIDYLKSLQVEIHNVFIDMVKMRRG-----SKLK--GDDALFSGLFWTGMRGLDLGLIDGL (220)
```

„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

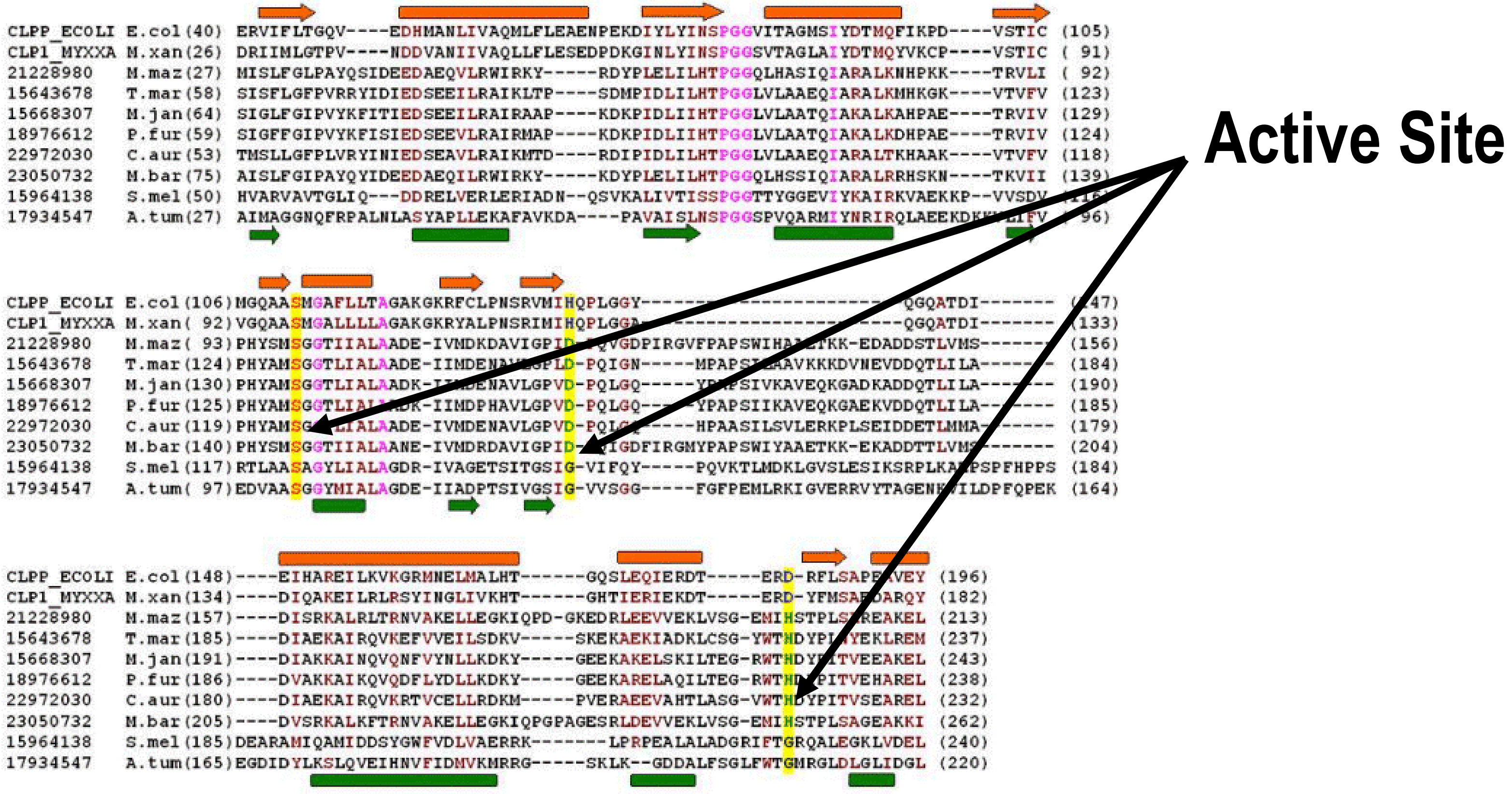
Mutations of important positions (such as active sites) are almost always harmful

You can recognize important positions because they never mutate!

MSAs reveal these ***conserved*** positions

An Example of Conserved Positions:

(The Serine Proteases Active Site)



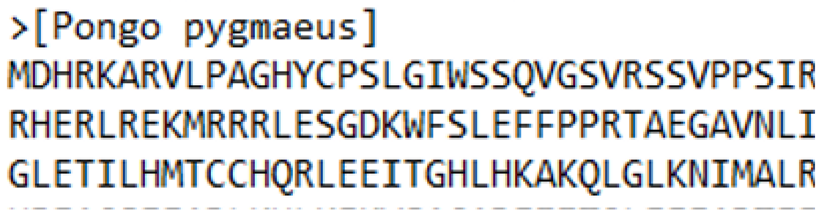
```
CLPP_ECOLI E.col(40) ERVIFLTGQV----EDHMANLIVAQMLFLEAENPEKDIYLYINSPGGVITAGMSIYDTMQFIKPD----VSTIC (105)
CLP1_MYXXA M.xan(26) DRIIMLGTPV----NDDVANIIVAQLLFLESEDPDKGINLYINSPGGSVTAGLAIYDTMQYVKCP----VSTIC ( 91)
21228980   M.maz(27) MISLFGLPAYQSIDEEDAEQVLRWIRKY-----RDYPLELILHTPGGQLHASIQIARALKNHPKK----TRVLI ( 92)
15643678   T.mar(58) SISFLGFPVRRYIDIEDSEEILRAIKLTP----SDMPIDLILHTPGGLVLAAEQIARALKMHKGK----VTVFV (123)
15668307   M.jan(64) SIGLFGIPVYKFITIEDSEEILRAIRAAP----KDKPIDLIIHTPGGLVLAATQIAKALKAHPAE----TRVIV (129)
18976612   P.fur(59) SIGFFGIPVYKFISIEDSEEVLRAIRMAP----KDKPIDLIIHTPGGLVLAATQIAKALKDHPAE----TRVIV (124)
22972030   C.aur(53) TMSLLGFPLVRYINIEDSEAVLRAIKMTD----RDIPIDLILHTPGGLVLAAEQIARALTKHAAK----VTVFV (118)
23050732   M.bar(75) AISLFGIPAYQYIDEEDAEQILRWIRKY-----KDYPLELILHTPGGQLHSSIQIARALRRHSKN----TKVII (139)
15964138   S.mel(50) HVARVAVTGLIQ---DDRELVERLERIADN--QSVKALIVTISSPGGTTYGGEVIYKAIRKVAEKKP--VVSDV (116)
17934547   A.tum(27) AIMAGGNQFRPALNLASYAPLLEKAFAVKDA----PAVAISLNSPGGSPVQARMIYNRIRQLAEEKDKKVLIFV ( 96)

CLPP_ECOLI E.col(106)MGQAASMGAFLLTAGAKGKRFCLPNSRVMIHQPLGGY------------------------QGQATDI------- (147)
CLP1_MYXXA M.xan( 92)VGQAASMGALLLLAGAKGKRYALPNSRIMIHQPLGGA------------------------QGQATDI------- (133)
21228980   M.maz( 93)PHYSMSGGTIIALAADE-IVMDKDAVIGPID-PQVGDPIRGVFPAPSWIHAAETKK-EDADDSTLVMS------- (156)
15643678   T.mar(124)PHYAMSGGTLIALAADE-IIMDENAVLGPLD-PQIGN-----MPAPSILAAVKKKDVNEVDDQTLILA------- (184)
15668307   M.jan(130)PHYAMSGGTLIALAADK-IIMDENAVLGPVD-PQLGQ-----YPAPSIVKAVEQKGADKADDQTLILA------- (190)
18976612   P.fur(125)PHYAMSGGTLIALAADK-IIMDPHAVLGPVD-PQLGQ-----YPAPSIIKAVEQKGAEKVDDQTLILA------- (185)
22972030   C.aur(119)PHYAMSGGTLIALAADE-IVMDENAVLGPVD-PQLGQ-----HPAASILSVLERKPLSEIDDETLMMA------- (179)
23050732   M.bar(140)PHYSMSGGTIIALAANE-IVMDRDAVIGPID-PQIGDFIRGMYPAPSWIYAAETKK-EKADDTTLVMS------- (204)
15964138   S.mel(117)RTLAASAGYLIALAGDR-IVAGETSITGSIG-VIFQY-----PQVKTLMDKLGVSLESIKSRPLKALPSPFHPPS (184)
17934547   A.tum( 97)EDVAASGGYMIALAGDE-IIADPTSIVGSIG-VVSGG-----FGFPEMLRKIGVERRVYTAGENKVILDPFQPEK (164)

CLPP_ECOLI E.col(148)----EIHAREILKVKGRMNELMALHT------GQSLEQIERDT-----ERD-RFLSAPEAVEY (196)
CLP1_MYXXA M.xan(134)----DIQAKEILRLRSYINGLIVKHT------GHTIERIEKDT-----ERD-YFMSAPDARQY (182)
21228980   M.maz(157)----DISRKALRLTRNVAKELLEGKIQPD-GKEDRLEEVVEKLVSG-EMIHSTPLSAREAKEL (213)
15643678   T.mar(185)----DIAEKAIRQVKEFVVEILSDKV-----SKEKAEKIADKLCSG-YWTHDYPINYEKLREM (237)
15668307   M.jan(191)----DIAKKAINQVQNFVYNLLKDKY-----GEEKAKELSKILTEG-RWTHDYPITVEEAKEL (243)
18976612   P.fur(186)----DVAKKAIKQVQDFLYDLLKDKY-----GEEKARELAQILTEG-RWTHDYPITVEHAREL (238)
22972030   C.aur(180)----DIAEKAIRQVKRTVCELLRDKM-----PVERAEEVAHTLASG-VWTHDYPITVSEAREL (232)
23050732   M.bar(205)----DVSRKALKFTRNVAKELLEGKIQPGPAGESRLDEVVEKLVSG-EMIHSTPLSAGEAKKI (262)
15964138   S.mel(185)DEARAMIQAMIDDSYGWFVDLVAERRK-------LPRPEALALADGRIFTGRQALEGKLVDEL (240)
17934547   A.tum(165)EGDIDYLKSLQVEIHNVFIDMVKMRRG-----SKLK--GDDALFSGLFWTGMRGLDLGLIDGL (220)
```

Active Site

Gathering Sequences with BLAST

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

- Homework 4 : 1) Find and download five similar sequences.

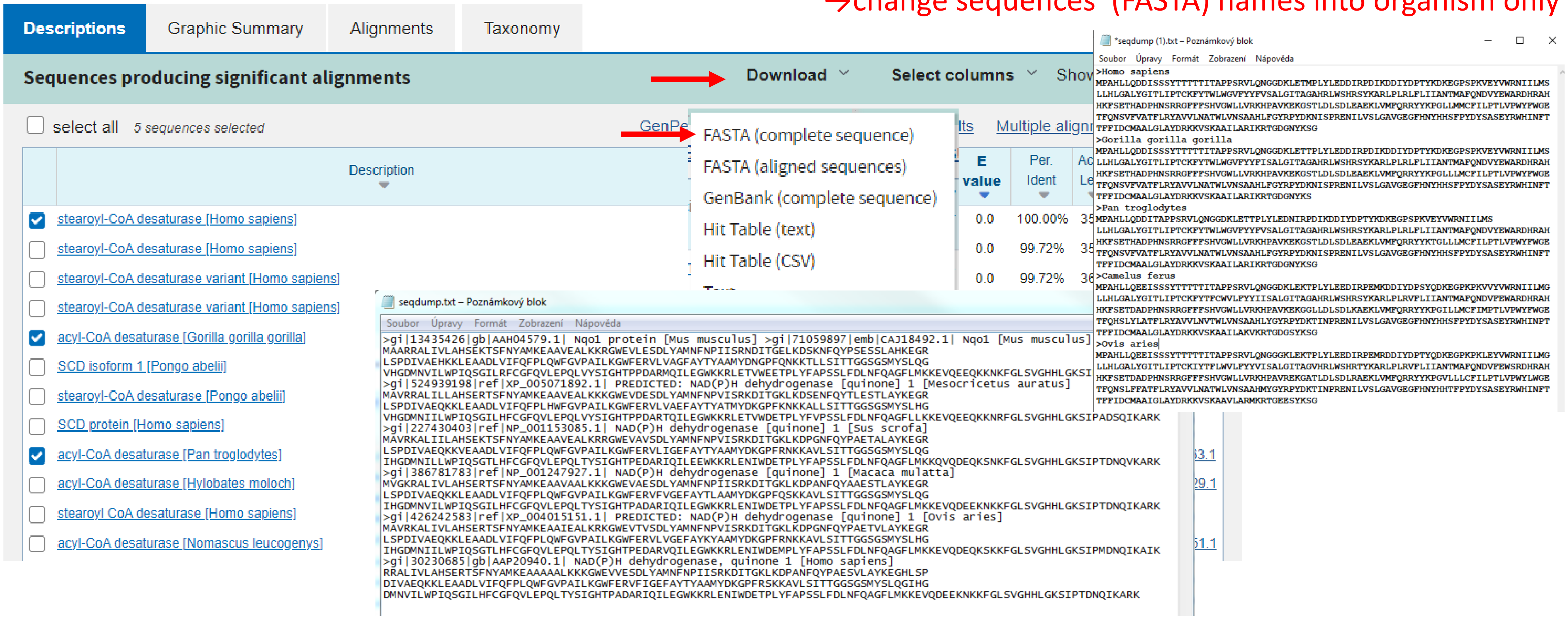


```
>[Pongo pygmaeus]
MDHRKARVLPAGHYCPSLGIWSSQVGSVRSSVPPSIR
RHERLREKMRRRLESGDKWFSLEFFPPRTAEGAVNLI
GLETILHMTCCHQRLEEITGHLHKAKQLGLKNIMALR
```

remove brackets [ ]

Gathering Sequences with BLAST

→change sequences (FASTA) names into organism only



☐ select all *5 sequences selected*

- √ ☑ stearoyl-CoA desaturase [Homo sapiens] 0.0 100.00% 35
- ☐ stearoyl-CoA desaturase [Homo sapiens] 0.0 99.72% 35
- ☐ stearoyl-CoA desaturase variant [Homo sapiens] 0.0 99.72% 36
- ☐ stearoyl-CoA desaturase variant [Homo sapiens]
- √ ☑ acyl-CoA desaturase [Gorilla gorilla gorilla]
- ☐ SCD isoform 1 [Pongo abelii]
- ☐ stearoyl-CoA desaturase [Pongo abelii]
- ☐ SCD protein [Homo sapiens]
- √ ☑ acyl-CoA desaturase [Pan troglodytes]
- ☐ acyl-CoA desaturase [Hylobates moloch]
- ☐ stearoyl CoA desaturase [Homo sapiens]
- ☐ acyl-CoA desaturase [Nomascus leucogenys]

seqdump.txt – Poznámkový blok

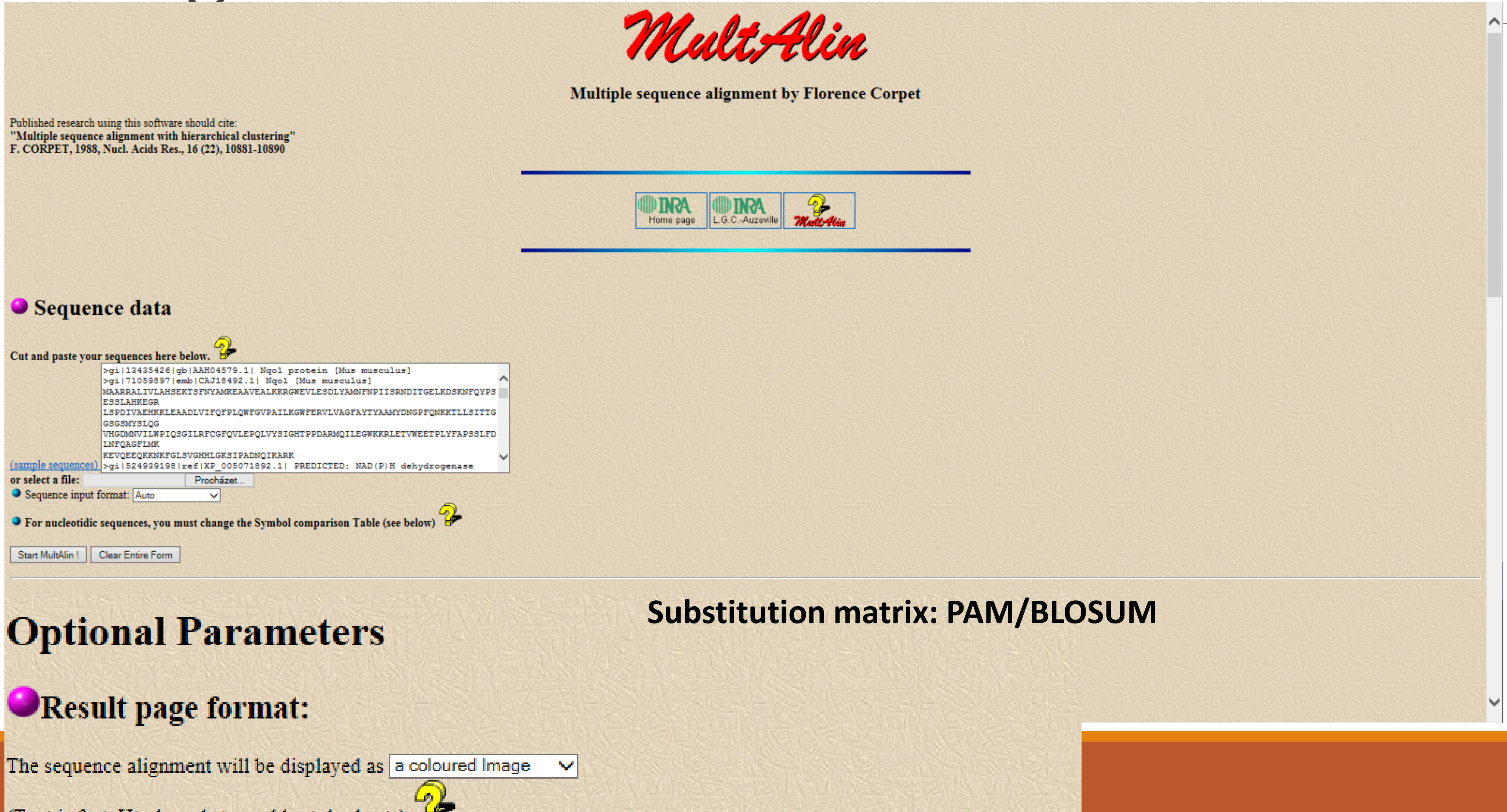
```
>gi|13435426|gb|AAH04579.1| Nqo1 protein [Mus musculus] >gi|71059897|emb|CAJ18492.1| Nqo1 [Mus musculus]
MAARRALIVLAHSEKTSFNYAMKEAAVEALKKRGWEVLESDLYAMNFNPIISRNDITGELKDSKNFQYPSESSLAHKEGR
LSPDIVAEHKKLEAADLVIFQFPLQWFGVPAILKGWFERVLVAGFAYTYAAMYDNGPFQNKKTLLSITTGGSGSMYSLQG
VHGDMNVILWPIQSGILRFCGFQVLEPQLVYSIGHTPPDARMQILEGWKKRLETVWEETPLYFAPSSLFDLNFQAGFLMKKEVQEEQKKNKFGLSVGHHLGKSI
>gi|524939198|ref|XP_005071892.1| PREDICTED: NAD(P)H dehydrogenase [quinone] 1 [Mesocricetus auratus]
MAVRRALILLAHSERTSFNYAMKEAAVEALKKKGWEVDESDLYAMNFNPVISRKDITGKLKDSENFQYTLESTLAYKEGR
LSPDIVAEQKKLEAADLVIFQFPLHWFGVPAILKGWFERVLVAEFAYTYATMYDKGPFKNKKALLSITTGGSGSMYSLHG
VHGDMNIILWPIQSGILHFCGFQVLEPQLVYSIGHTPPDARTQILEGWKKRLETVWDETPLYFVPSSLFDLNFQAGFLLKKEVQEEQKKNRFGLSVGHHLGKSIPADSQIKARK
>gi|227430403|ref|NP_001153085.1| NAD(P)H dehydrogenase [quinone] 1 [Sus scrofa]
MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGR
LSPDIVAEQKKVEAADLVIFQFPLQWFGVPAILKGWFERVLIGEFAYTYAAMYDKGPFRNKKAVLSITTGGSGSMYSLQG
IHGDMNILLWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKQVQDEQKSNKFGLSVGHHLGKSIPTDNQVKARK
>gi|386781783|ref|NP_001247927.1| NAD(P)H dehydrogenase [quinone] 1 [Macaca mulatta]
MVGKRALIVLAHSERTSFNYAMKEAAVAALKKKGWEVAESDLYAMNFNPIISRKDITGKLKDPANFQYAAESTLAYKEGR
LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFVGEFAYTLAAMYDKGPFQSKKAVLSITTGGSGSMYSLQG
IHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK
>gi|426242583|ref|XP_004015151.1| PREDICTED: NAD(P)H dehydrogenase [quinone] 1 [Ovis aries]
MAVRKALIVLAHSERTSFNYAMKEAAIEALKRKGWEVTVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETVLAYKEGR
LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVLVGEFAYKYAAMYDKGPFRNKKAVLSITTGGSGSMYSLHG
IHGDMNIILWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARVQILEGWKKRLENIWDEMPLYFAPSSLFDLNFQAGFLMKKEVQDEQKSKKFGLSVGHHLGKSIPMDNQIKAIK
>gi|30230685|gb|AAP20940.1| NAD(P)H dehydrogenase, quinone 1 [Homo sapiens]
RRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSP
DIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHG
DMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK
```

*seqdump (1).txt – Poznámkový blok

```
>Homo sapiens
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVEYVWRNIILMS
LLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
HKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGE
TFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFT
TFFIDCMAALGLAYDRKKVSKAAILARIKRTGDGNYKSG
>Gorilla gorilla gorilla
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETTPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVEYVWRNIILMS
LLHLGALYGITLIPTCKFYTWLWGVFYYFISALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
HKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLLMCFILPTLVPWYFWGE
TFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFT
TFFIDCMAALGLAYDRKKVSKAAILARIKRTGDGNYKS
>Pan troglodytes
MPAHLLQDDITAPPSRVLQNGGDKLETTPLYLEDNIRPDIKDDIYDPTYKDKEGPSPKVEYVWRNIILMS
LLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
HKFSETHADPHNSRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKTGLLLMCFILPTLVPWYFWGE
TFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFT
TFFIDCMAALGLAYDRKKVSKAAILARIKRTGDGNYKSG
>Camelus ferus
MPAHLLQEEISSSYTTTTTITAPPSRVLQNGGDKLEKTPLYLEEDIRPEMKDDIYDPSYQDKEGPKPKVVYVWRNIILMG
LLHLGALYGITLIPTCKFYTFCWVLFYYIISALGITAGAHRLWSHRSYKARLPLRVFLIIANTMAFQNDVFEWARDHRAH
HKFSETDADPHNSRRGFFFSHVGWLLVRKHPAVKEKGGLLDLSDLKAEKLVMFQRRYYKPGILLMCFIMPTLVPWYFWGE
TFQHSLYLATFLRYAVVLNVTWLVNSAAHLYGYRPYDKTINPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINPT
TFFIDCMAALGLAYDRKKVSKAAILAKVKRTGDGSYKSG
>Ovis aries
MPAHLLQEEISSSYTTTTTITAPPSRVLQNGGGKLEKTPLYLEEDIRPEMRDDIYDPTYQDKEGPKPKLEYVWRNIILMG
LLHLGALYGITLIPTCKIYTFLWVLFYYVISALGITAGVHRLWSHRTYKARLPLRVFLIIANTMAFQNDVFEWSRDHRAH
HKFSETDADPHNSRRGFFFSHVGWLLVRKHPAVREKGATLDLSDLRAEKLVMFQRRYYKPGVLLLCFILPTLVPWYLWGE
TFQNSLFFATFLRYAVVLNATWLVNSAAHMYGYRPYDKTINPRENILVSLGAVGEGFHNYHHTFPYDYSASEYRWHINFT
TFFIDCMAAIGLAYDRKKVSKAAVLARMKRTGEESYKSG
```

Alignment: MultAlin

http://multalin.toulouse.inra.fr/multalin/



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]
MAARRALIVLAHSEKTSFNYAMKEAAVEALKKRGWEVLESDLYAMNFNPIISRNDITGELKDSKNFQYPS
ESSLAHKEGR
LSPDIVAEHKKLEAADLVIFQFPLQWFGVPAILKGWFERVLVAGFAYTYAAMYDNGPFQNKKTLLSITTG
GSGSMYSLQG
VHGDMNVILWPIQSGILRFCGFQVLEPQLVYSIGHTPPDARMQILEGWKKRLETVWEETPLYFAPSSLFD
LNFQAGFLMK
KEVQEEQKKNKFGLSVGHHLGKSIPADNQIKARK
>gi|524939198|ref|XP_005071892.1| PREDICTED: NAD(P)H dehydrogenase
```

(sample sequences)

or select a file:

Sequence input format: Auto

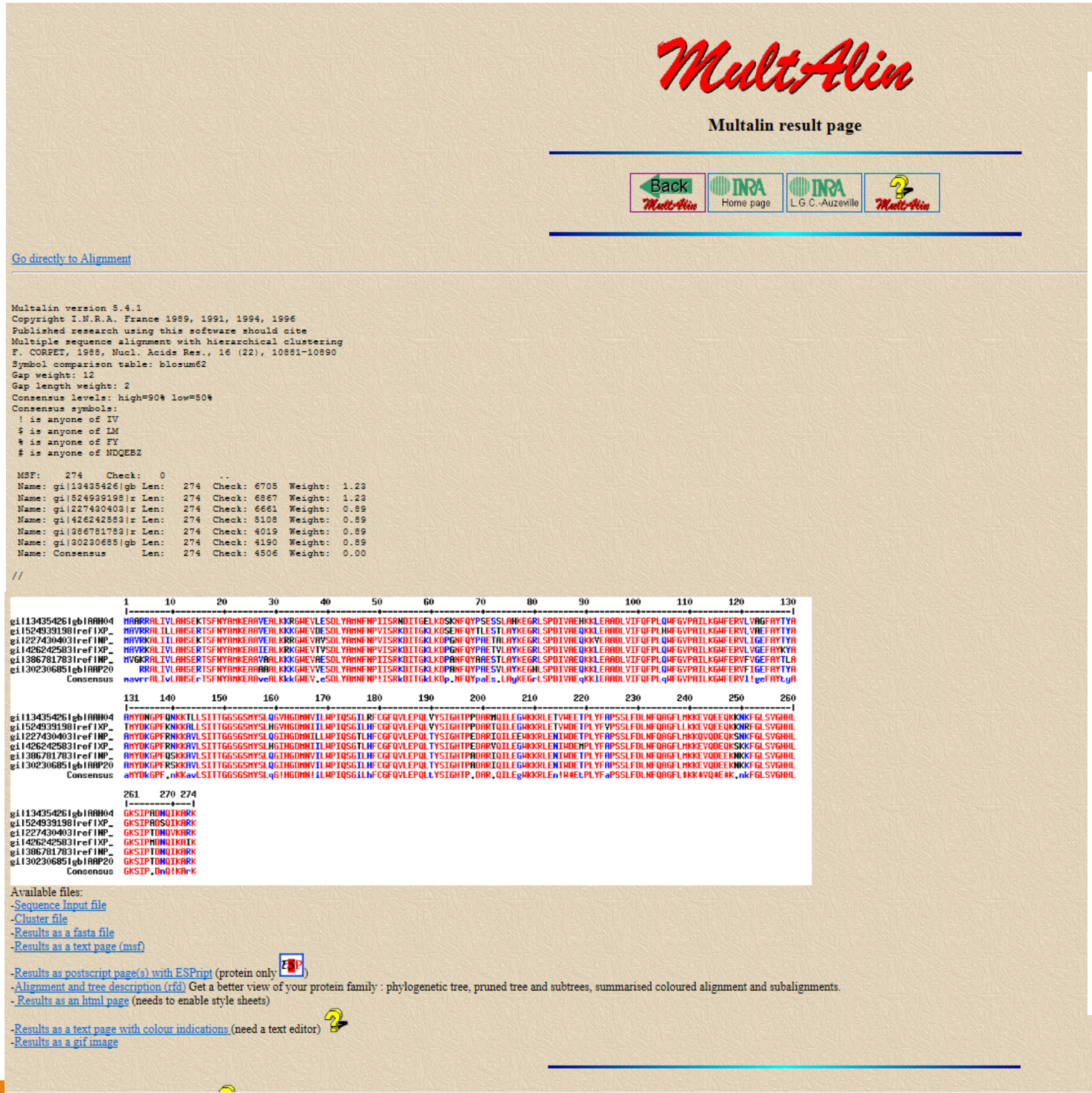
For nucleotidic sequences, you must change the Symbol comparison Table (see below)

Substitution matrix: PAM/BLOSUM

Optional Parameters

Result page format:

The sequence alignment will be displayed as a coloured Image



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. CORPET, 1988, Nucl. Acids Res., 16 (22), 10881-10890
Symbol comparison table: blosum62
Gap weight: 12
Gap length weight: 2
Consensus levels: high=90% low=50%
Consensus symbols:
 ! is anyone of IV
 $ is anyone of LM
 % is anyone of FY
 # is anyone of NDQEBZ

 MSF:    274    Check:   0        ..
 Name: gi|13435426|gb Len:   274  Check: 6705  Weight:  1.23
 Name: gi|524939198|r Len:   274  Check: 6867  Weight:  1.23
 Name: gi|227430403|r Len:   274  Check: 6661  Weight:  0.89
 Name: gi|426242583|r Len:   274  Check: 5108  Weight:  0.89
 Name: gi|386781783|r Len:   274  Check: 4019  Weight:  0.89
 Name: gi|30230685|gb 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
                      |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
gi|13435426|gb|AAH04  MAARRALIVLAHSEKTSFNYAMKEAAVEALKKRGWEVLESDLYAMNFNPIISRNDITGELKDSKNFQYPSESSLAHKEGRLSPDIVAEHKKLEAADLVIFQFPLQWFGVPAILKGWFERVLVAGFAYTYA
gi|524939198|ref|XP_  MAVRRALILLAHSERTSFNYAMKEAAVEALKKKGWEVDESDLYAMNFNPVISRKDITGKLKDSENFQYTLESTLAYKEGRLSPDIVAEQKKLEAADLVIFQFPLHWFGVPAILKGWFERVLVAEFAYTYA
gi|227430403|ref|NP_  MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGRLSPDIVAEQKKVEAADLVIFQFPLQWFGVPAILKGWFERVLIGEFAYTYA
gi|426242583|ref|XP_  MAVRKALIVLAHSERTSFNYAMKEAAIEALKRKGWEVTVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETVLAYKEGRLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVLVGEFAYKYA
gi|386781783|ref|NP_  MVGKRALIVLAHSERTSFNYAMKEAAVAALKKKGWEVAESDLYAMNFNPIISRKDITGKLKDPANFQYAAESTLAYKEGRLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFVGEFAYTLA
gi|30230685|gb|AAP20     RRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYA
           Consensus  mavrrALIvLAHSErTSFNYAMKEAAveALKkkGWEV.eSDLYAMNFNP!ISRkDITGkLKDp.NFQYpaEs.LAyKEGrLSPDIVAEqKKlEAADLVIFQFPLqWFGVPAILKGWFERVl!geFAYtyA

                      131    140       150       160       170       180       190       200       210       220       230       240       250       260
                      |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
gi|13435426|gb|AAH04  AMYDNGPFQNKKTLLSITTGGSGSMYSLQGVHGDMNVILWPIQSGILRFCGFQVLEPQLVYSIGHTPPDARMQILEGWKKRLETVWEETPLYFAPSSLFDLNFQAGFLMKKEVQEEQKKNKFGLSVGHHL
gi|524939198|ref|XP_  TMYDKGPFKNKKALLSITTGGSGSMYSLHGVHGDMNIILWPIQSGILHFCGFQVLEPQLVYSIGHTPPDARTQILEGWKKRLETVWDETPLYFVPSSLFDLNFQAGFLLKKEVQEEQKKNRFGLSVGHHL
gi|227430403|ref|NP_  AMYDKGPFRNKKAVLSITTGGSGSMYSLQGIHGDMNILLWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKQVQDEQKSNKFGLSVGHHL
gi|426242583|ref|XP_  AMYDKGPFRNKKAVLSITTGGSGSMYSLHGIHGDMNIILWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARVQILEGWKKRLENIWDEMPLYFAPSSLFDLNFQAGFLMKKEVQDEQKSKKFGLSVGHHL
gi|386781783|ref|NP_  AMYDKGPFQSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHL
gi|30230685|gb|AAP20  AMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHL
           Consensus  aMYDkGPF.nKKavLSITTGGSGSMYSLqG!HGDMN!iLWPIQSGiLhFCGFQVLEPQLtYSIGHTP.DAR.QILEgWKKRLEn!W#EtPLYFaPSSLFDLNFQAGFL$KK#VQ#E#K.nkFGLSVGHHL

                      261    270 274
                      |--------+---|
gi|13435426|gb|AAH04  GKSIPADNQIKARK
gi|524939198|ref|XP_  GKSIPADSQIKARK
gi|227430403|ref|NP_  GKSIPTDNQVKARK
gi|426242583|ref|XP_  GKSIPMDNQIKAIK
gi|386781783|ref|NP_  GKSIPTDNQIKARK
gi|30230685|gb|AAP20  GKSIPTDNQIKARK
           Consensus  GKSIP.DnQ!KArK
```

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 ESP)
-Alignment and tree description (rfd) 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

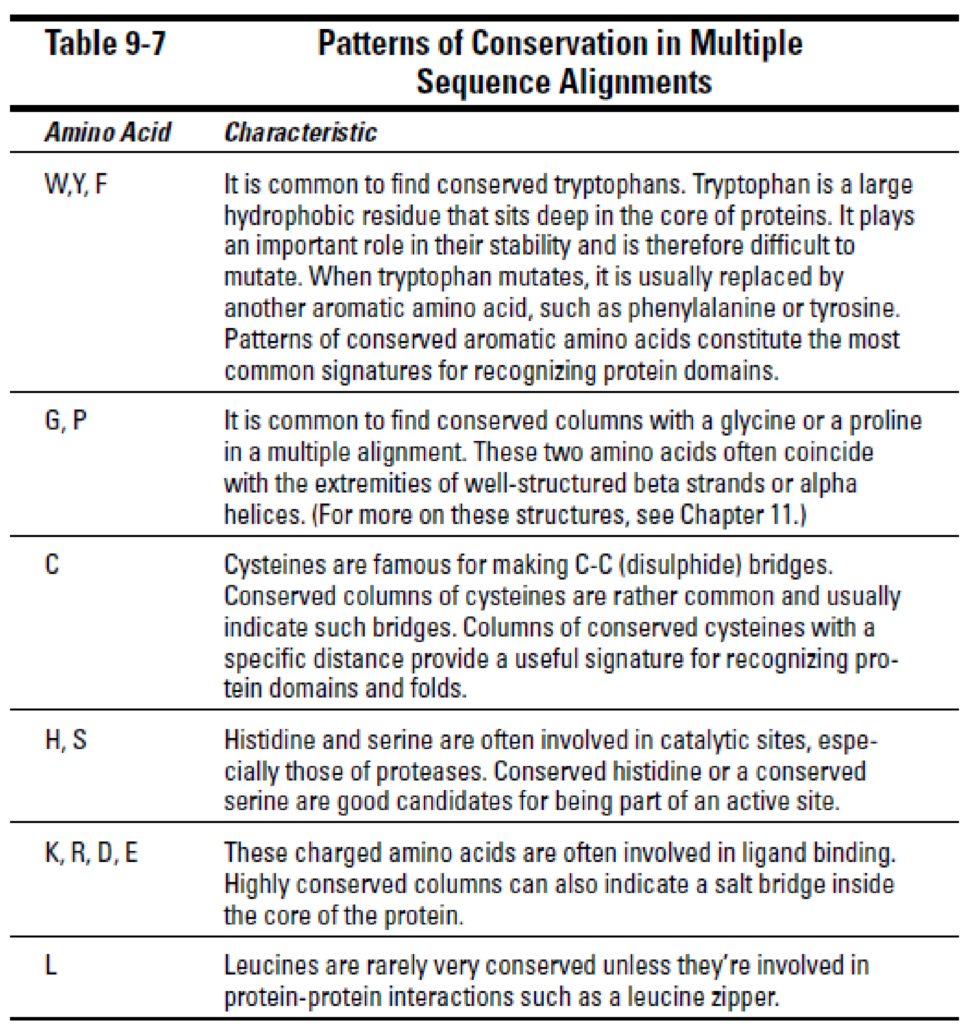
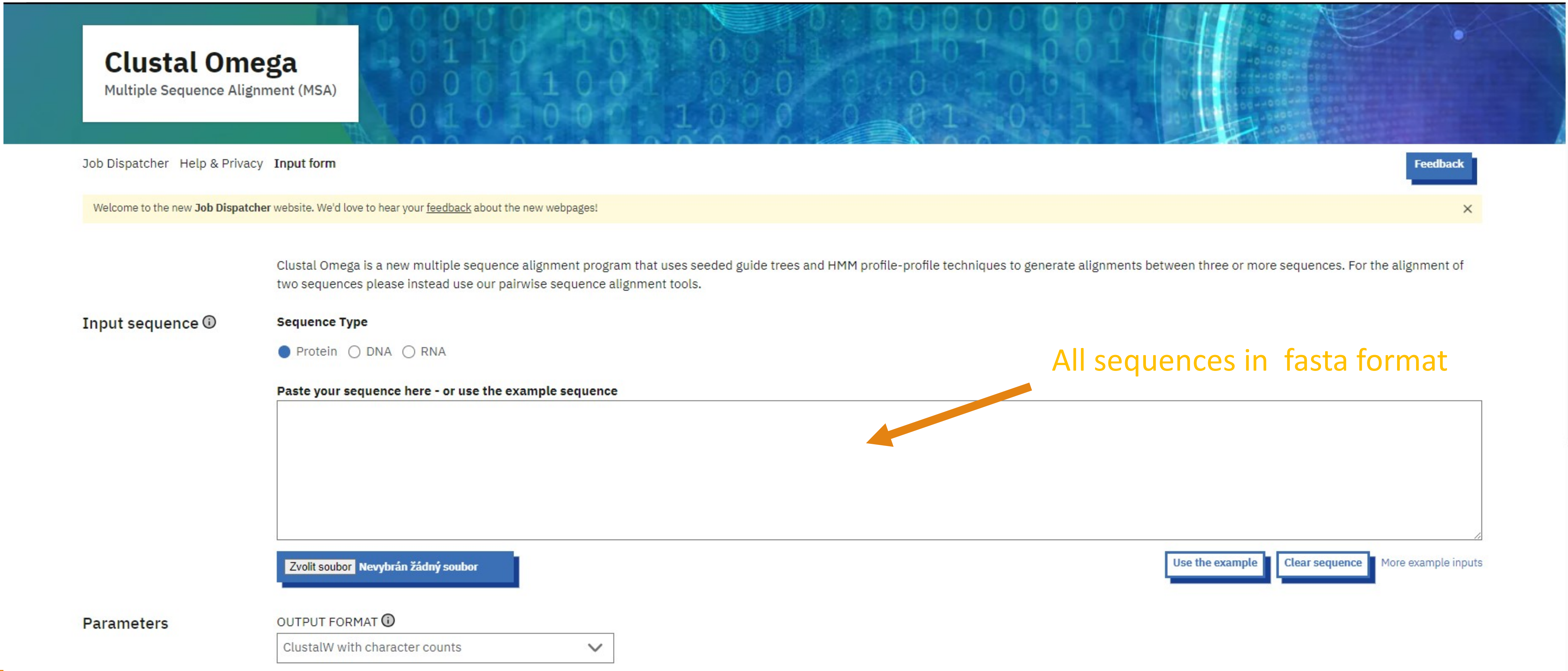


Table 9-7 Patterns of Conservation in Multiple Sequence Alignments

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: Clustal Omega



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

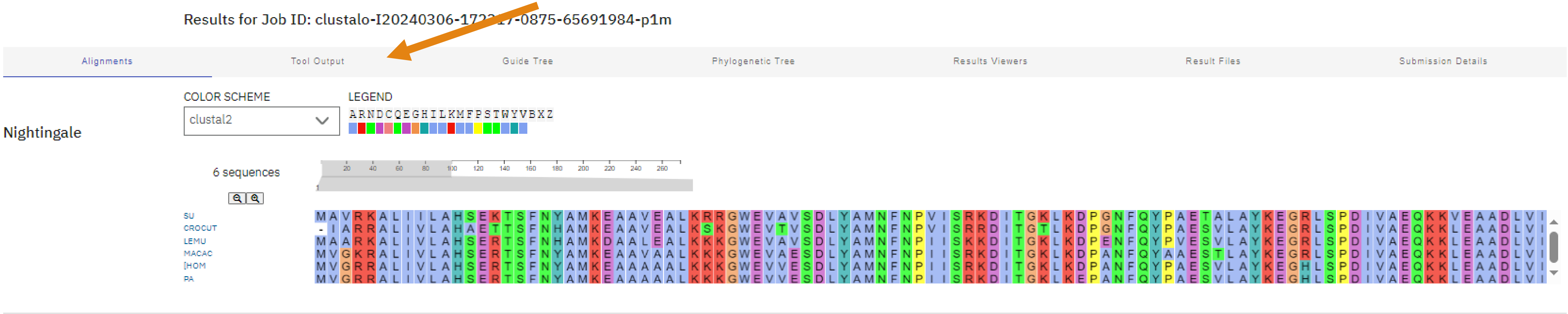
Parameters

OUTPUT FORMAT ⓘ

ClustalW with character counts

All sequences in fasta format

Alignment: Clustal Omega



Results for Job ID: clustalo-I20240306-172[illegible]7-0875-65691984-p1m

Nightingale

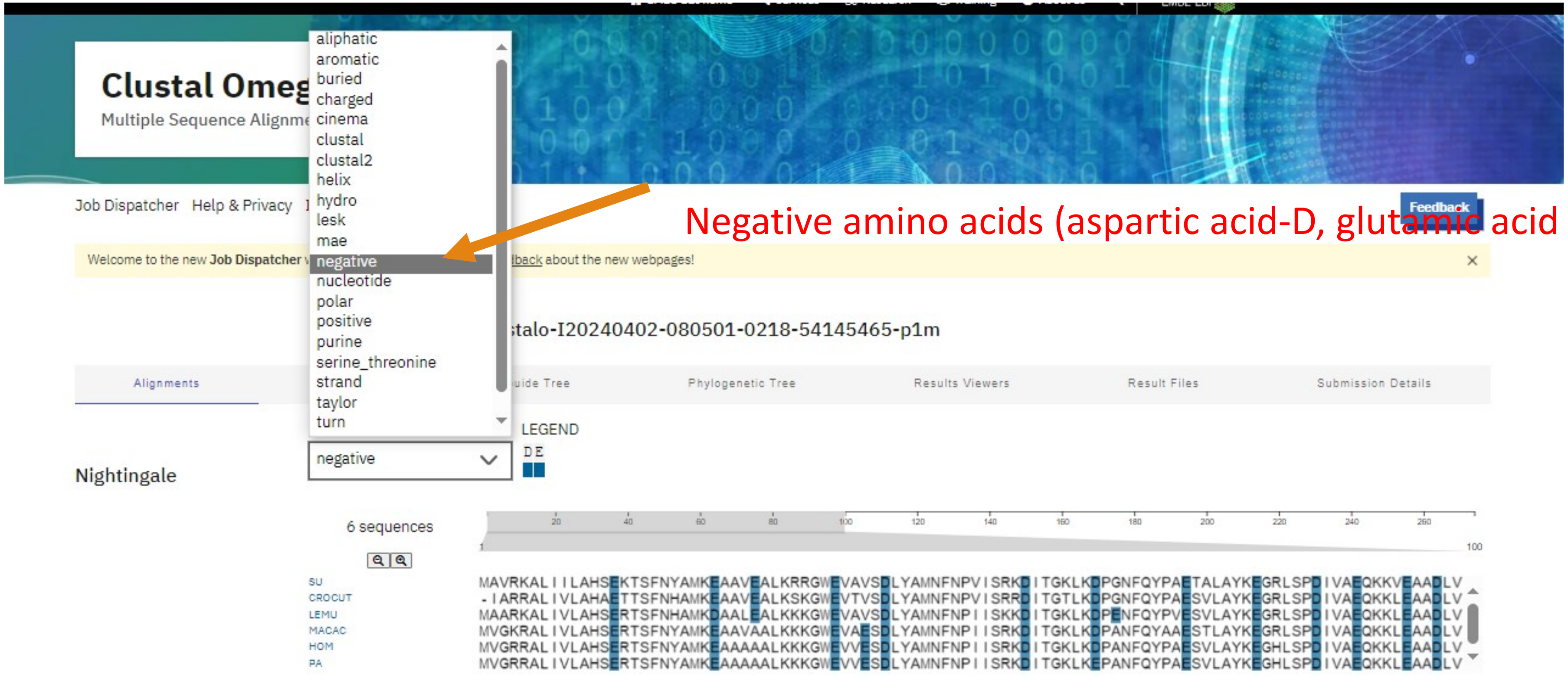
COLOR SCHEME: clustal2

LEGEND: ARNDCQEGHILKMFPSTWYVBXZ

6 sequences

```
SU      MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGRLSPDIVAEQKKVEAADLVI
CROCUT  -IARRALIVLAHAETTSFNHAMKEAAVEALKSKGWEVTVSDLYAMNFNPVISRRDITGTLKDPGNFQYPAESVLAYKEGRLSPDIVAEQKKLEAADLVI
LEMU    MAARKALIVLAHSERTSFNHAMKDAALEALKKKGWEVAVSDLYAMNFNPIISKKDITGKLKDPENFQYPVESVLAYKEGRLSPDIVAEQKKLEAADLVI
MACAC   MVGKRALIVLAHSERTSFNYAMKEAAVAALKKKGWEVAESDLYAMNFNPIISRKDITGKLKDPANFQYAAESTLAYKEGRLSPDIVAEQKKLEAADLVI
[HOM    MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVI
PA      MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKEPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVI
```

Alignment: Clustal Omega



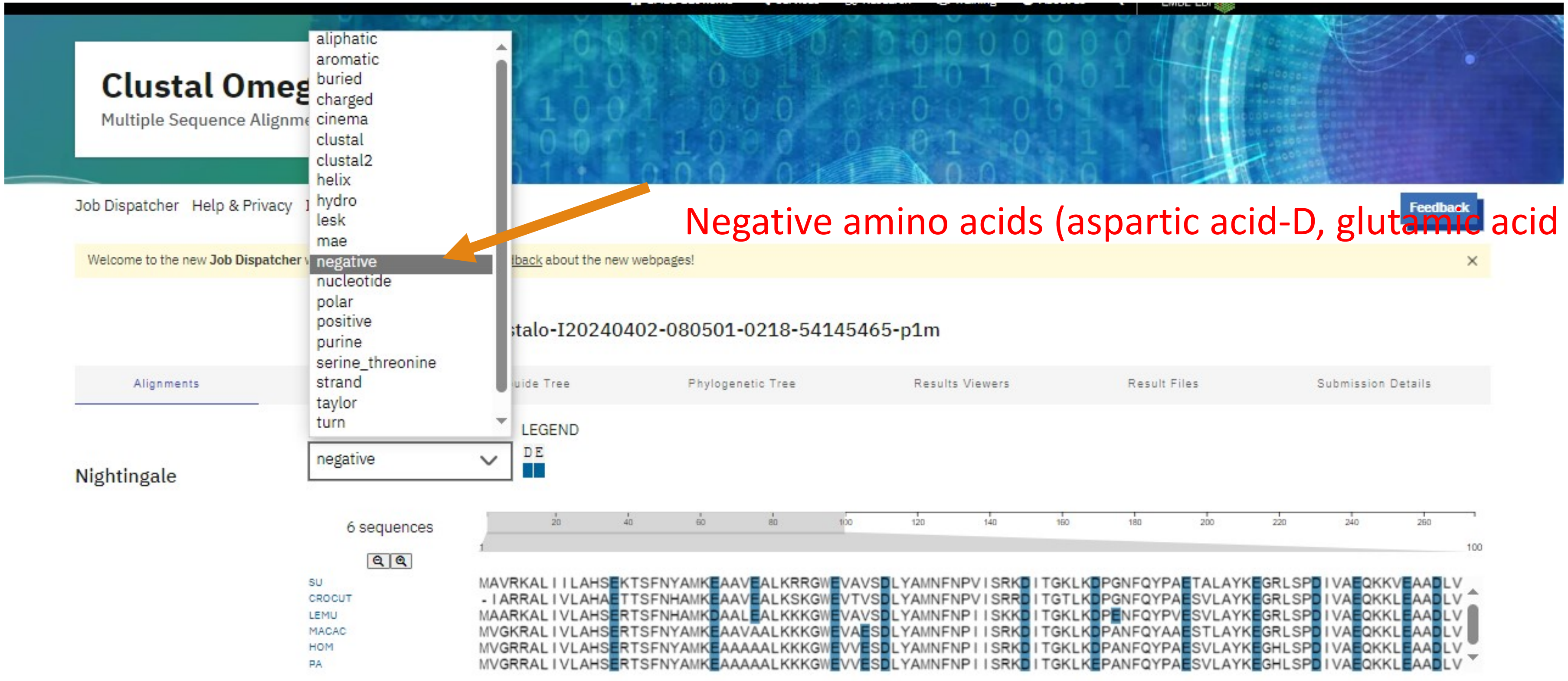
Welcome to the new **Job Dispatcher** … back about the new webpages!

talo-I20240402-080501-0218-54145465-p1m

```
SU      MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGRLSPDIVAEQKKVEAADLV
CROCUT  -IARRALIVLAHAETTSFNHAMKEAAVEALKSKGWEVTVSDLYAMNFNPVISRRDITGTLKDPGNFQYPAESVLAYKEGRLSPDIVAEQKKLEAADLV
LEMU    MAARKALIVLAHSERTSFNHAMKDAALEALKKKGWEVAVSDLYAMNFNPIISKKDITGKLKDPENFQYPVESVLAYKEGRLSPDIVAEQKKLEAADLV
MACAC   MVGKRALIVLAHSERTSFNYAMKEAAVAALKKKGWEVAESDLYAMNFNPIISRKDITGKLKDPANFQYAAESTLAYKEGRLSPDIVAEQKKLEAADLV
HOM     MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLV
PA      MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKEPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLV
```

Negative amino acids (aspartic acid-D, glutamic acid – E)

Alignment: Clustal Omega



Welcome to the new **Job Dispatcher** v … lback about the new webpages!

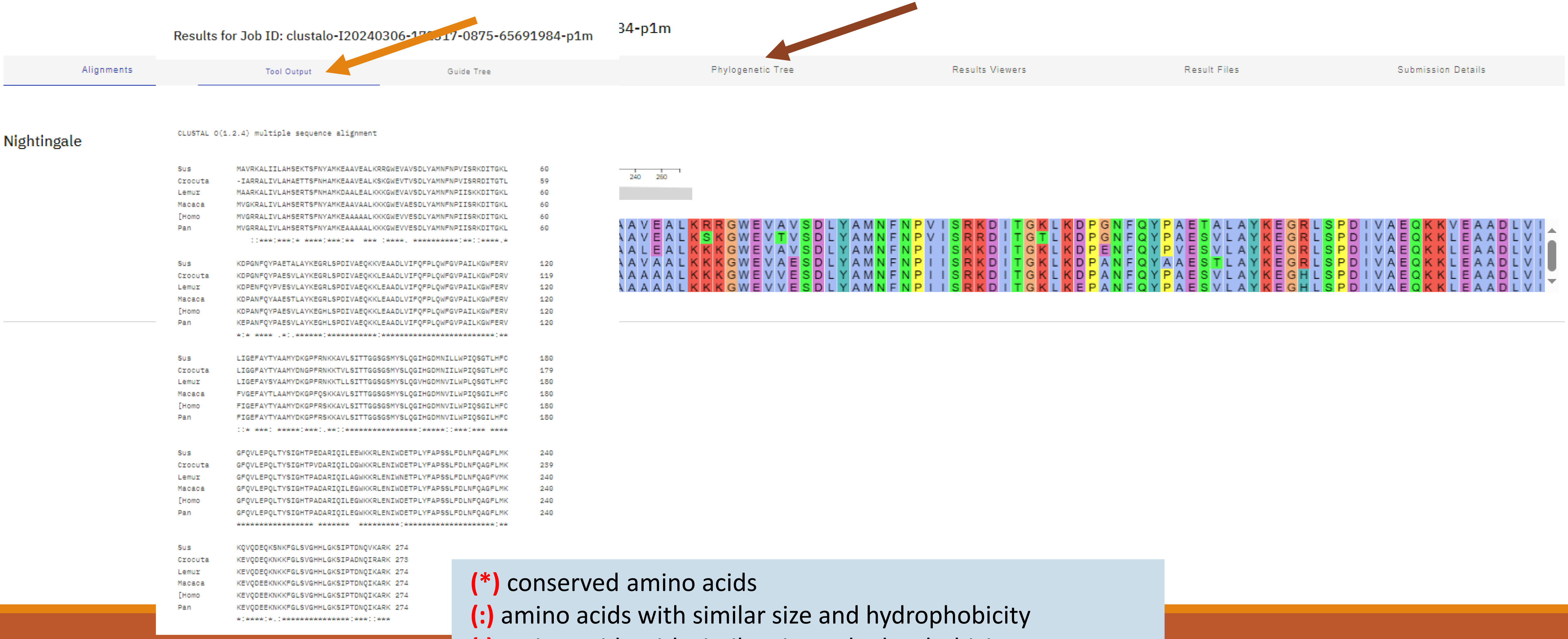
italo-I20240402-080501-0218-54145465-p1m

negative

```
SU      MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGRLSPDIVAEQKKVEAADLV
CROCUT  -IARRALIVLAHAETTSFNHAMKEAAVEALKSKGWEVTVSDLYAMNFNPVISRRDITGTLKDPGNFQYPAESVLAYKEGRLSPDIVAEQKKLEAADLV
LEMU    MAARKALIVLAHSERTSFNHAMKDAALEALKKKGWEVAVSDLYAMNFNPIISKKDITGKLKDPENFQYPVESVLAYKEGRLSPDIVAEQKKLEAADLV
MACAC   MVGKRALIVLAHSERTSFNYAMKEAAVAALKKKGWEVAESDLYAMNFNPIISRKDITGKLKDPANFQYAAESTLAYKEGRLSPDIVAEQKKLEAADLV
HOM     MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLV
PA      MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKEPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLV
```

Negative amino acids (aspartic acid-D, glutamic acid – E)

Alignment: Clustal Omega



```
CLUSTAL O(1.2.4) multiple sequence alignment


Sus          MAVRKALIILAHSEKTSFNYAMKEAAVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKL      60
Crocuta      -IARRALIVLAHAETTSFNHAMKEAAVEALKSKGWEVTVSDLYAMNFNPVISRRDITGTL      59
Lemur        MAARKALIVLAHSERTSFNHAMKDAALEALKKKGWEVAVSDLYAMNFNPIISKKDITGKL      60
Macaca       MVGKRALIVLAHSERTSFNYAMKEAAVAALKKKGWEVAESDLYAMNFNPIISRKDITGKL      60
[Homo        MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKL      60
Pan          MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKL      60
                ::***:***:* ****:***:**  *** :****. **********:**::****.*

Sus          KDPGNFQYPAETALAYKEGRLSPDIVAEQKKVEAADLVIFQFPLQWFGVPAILKGWFERV      120
Crocuta      KDPGNFQYPAESVLAYKEGRLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFDRV      119
Lemur        KDPENFQYPVESVLAYKEGRLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV      120
Macaca       KDPANFQYAAESTLAYKEGRLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV      120
[Homo        KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV      120
Pan          KEPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV      120
             *:* **** .*:.******:***********:*************************:**

Sus          LIGEFAYTYAAMYDKGPFRNKKAVLSITTGGSGSMYSLQGIHGDMNILLWPIQSGTLHFC      180
Crocuta      LIGGFAYTYAAMYDNGPFRNKKTVLSITTGGSGSMYSLQGIHGDMNIILWPIQSGTLHFC      179
Lemur        LIGEFAYSYAAMYDKGPFRNKKTLLSITTGGSGSMYSLQGVHGDMNVILWPLQSGTLHFC      180
Macaca       FVGEFAYTLAAMYDKGPFQSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFC      180
[Homo        FIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFC      180
Pan          FIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFC      180
             ::* ***: *****:***:.**::****************:*****::***:*** ****

Sus          GFQVLEPQLTYSIGHTPEDARIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK      240
Crocuta      GFQVLEPQLTYSIGHTPVDARIQILDGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK      239
Lemur        GFQVLEPQLTYSIGHTPADARIQILAGWKKRLENIWNETPLYFAPSSLFDLNFQAGFVMK      240
Macaca       GFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK      240
[Homo        GFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK      240
Pan          GFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK      240
             ***************** *******  *********:********************:**

Sus          KQVQDEQKSNKFGLSVGHHLGKSIPTDNQVKARK 274
Crocuta      KEVQDEQKNKKFGLSVGHHLGKSIPADNQIRARK 273
Lemur        KEVQDEQKNKKFGLSVGHHLGKSIPTDNQIKARK 274
Macaca       KEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK 274
[Homo        KEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK 274
Pan          KEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK 274
             *:****:*.:***************:***::***
```

```
AVEALKRRGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGRLSPDIVAEQKKVEAADLVI
AVEALKSKGWEVTVSDLYAMNFNPVISRRDITGTLKDPGNFQYPAESVLAYKEGRLSPDIVAEQKKLEAADLVI
ALEALKKKGWEVAVSDLYAMNFNPIISKKDITGKLKDPENFQYPVESVLAYKEGRLSPDIVAEQKKLEAADLVI
AVAALKKKGWEVAESDLYAMNFNPIISRKDITGKLKDPANFQYAAESTLAYKEGRLSPDIVAEQKKLEAADLVI
AAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVI
AAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKEPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVI
```

(*) 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:0.04835,
(
Macaca:0.02251,
(
Homo:0.00000,
Pan:0.00365)
:0.01033)
:0.03377);
```

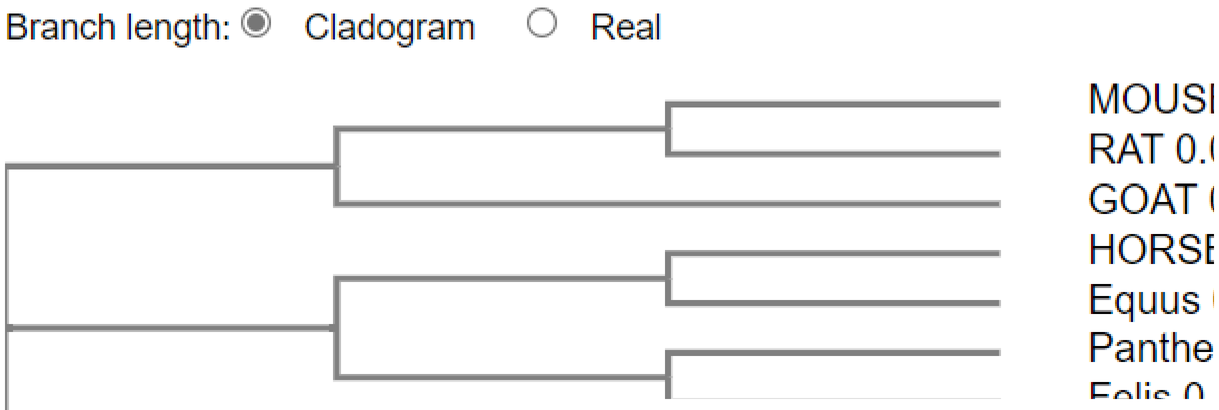
„Newick“ formate

Phylogram

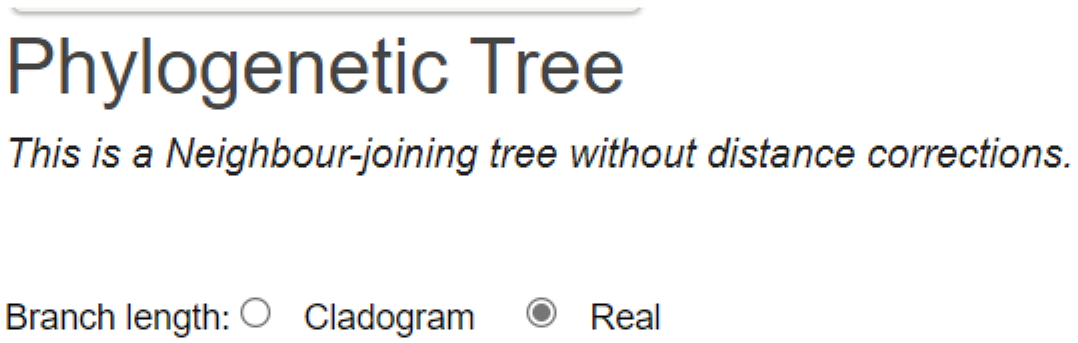
Phylogenetic tree

Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.



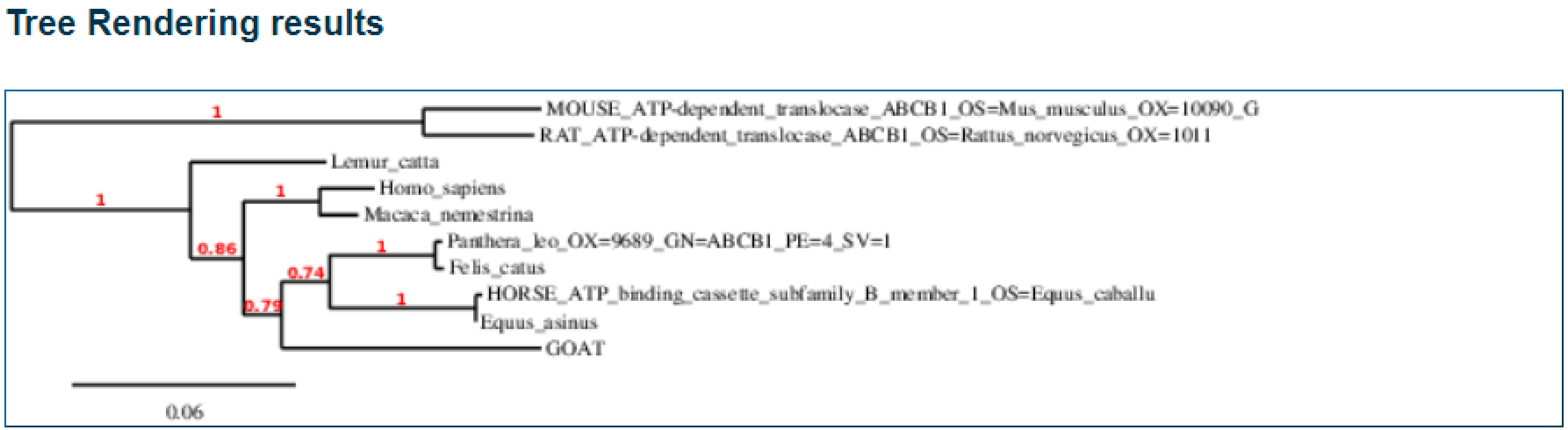
Branch length: ● Cladogram ○ Real



Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ○ Cladogram ● Real



Tree Rendering results

Practical part

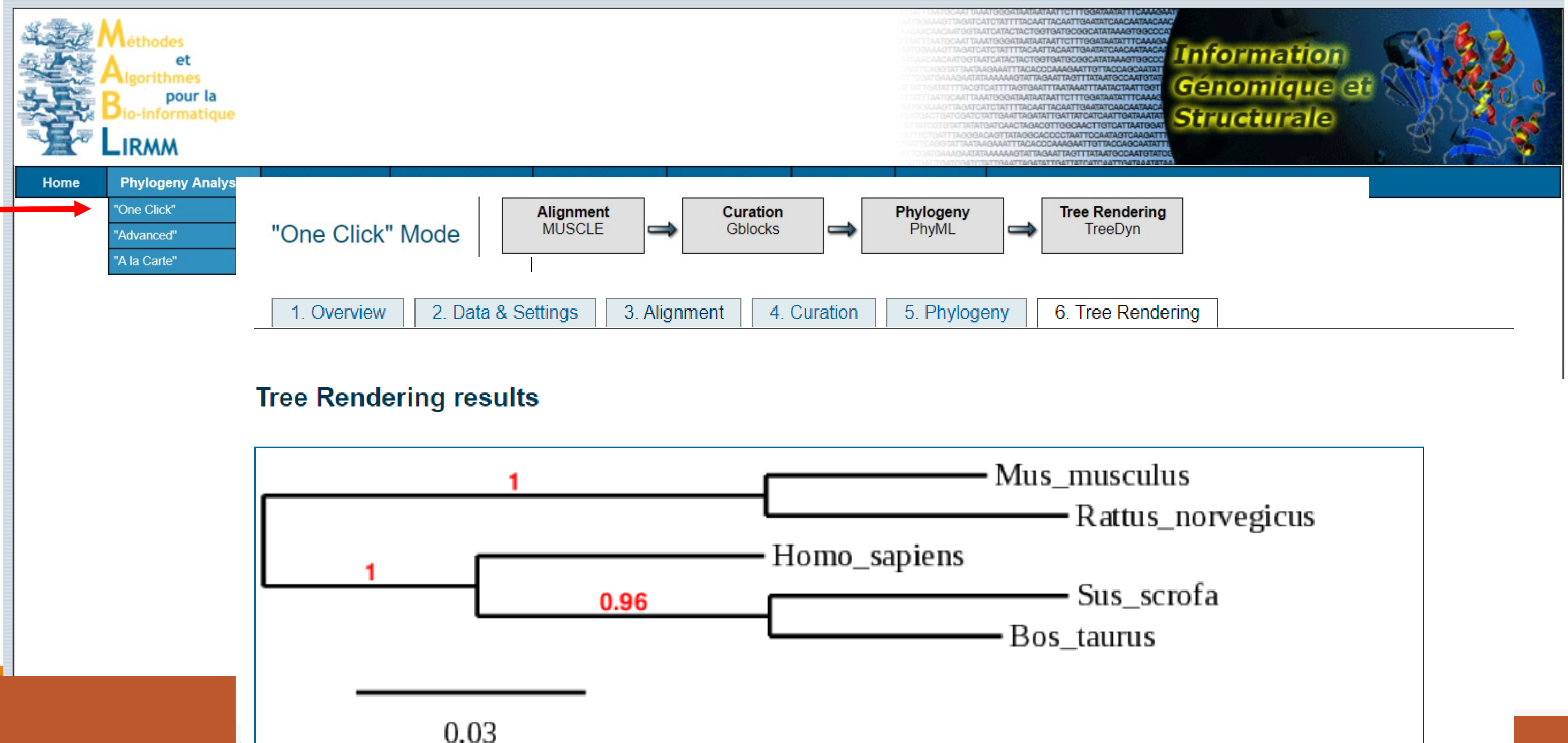
Try multiple alignment using five similar sequences from different organisms. (Hw4-1)

„advanced“ phylogeny analysis

Phylogeny.fr

Robust Phylogenetic Analysis For The Non-Specialist

„advanced“ phylogeny analysis

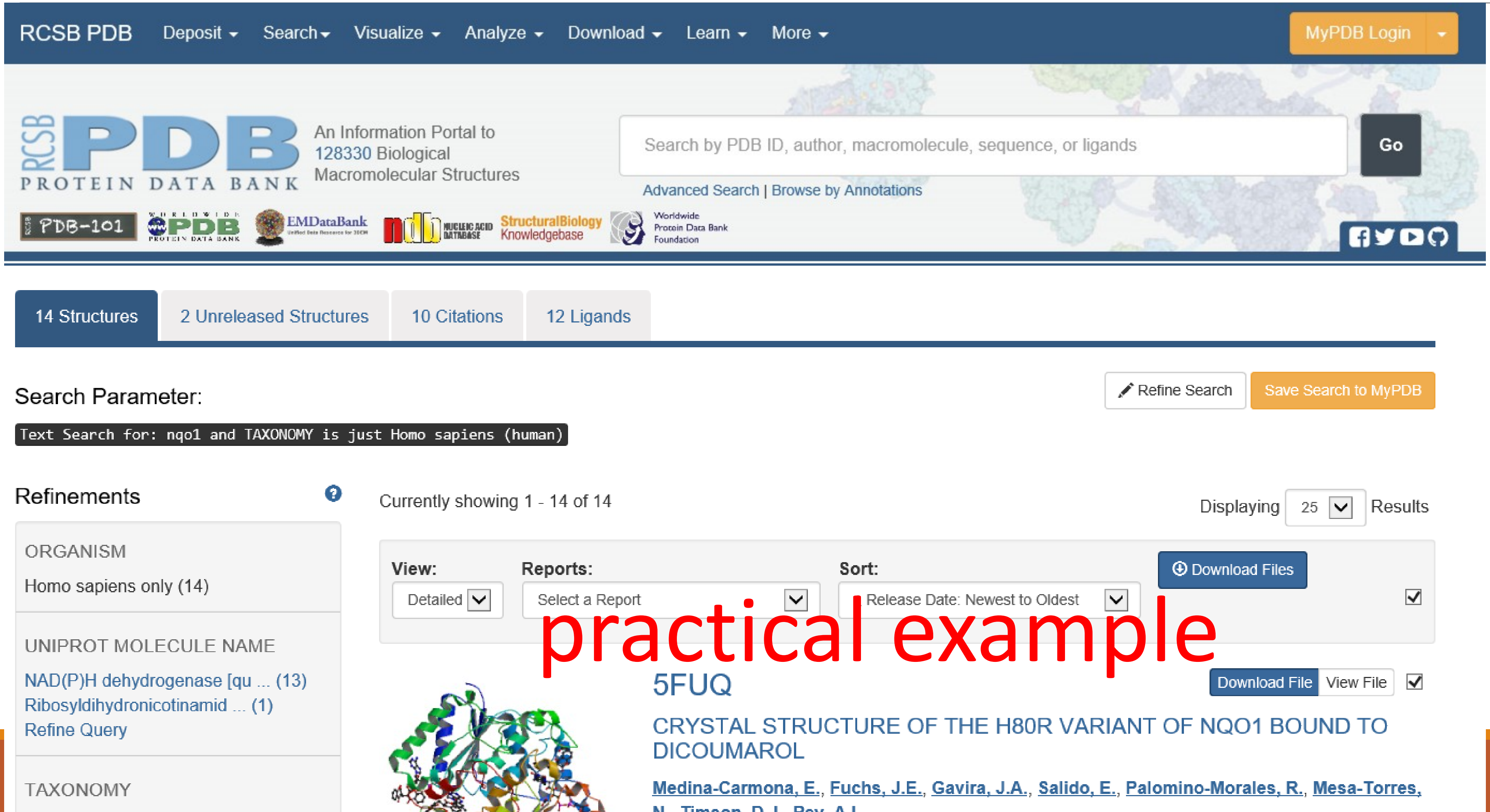


Practical part

Try building the phylogeny tree using phylogeny.org

Compare the trees

3-D protein structure: PDB



An Information Portal to 128330 Biological Macromolecular Structures

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

Advanced Search | Browse by Annotations

Search Parameter:

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

Refinements

ORGANISM

Homo sapiens only (14)

UNIPROT MOLECULE NAME

NAD(P)H dehydrogenase [qu ... (13)

Ribosyldihydronicotinamid ... (1)

Refine Query

TAXONOMY

Currently showing 1 - 14 of 14

Displaying 25 Results

View: Detailed

Reports: Select a Report

Sort: Release Date: Newest to Oldest

☑

practical example

5FUQ

CRYSTAL STRUCTURE OF THE H80R VARIANT OF NQO1 BOUND TO DICOUMAROL

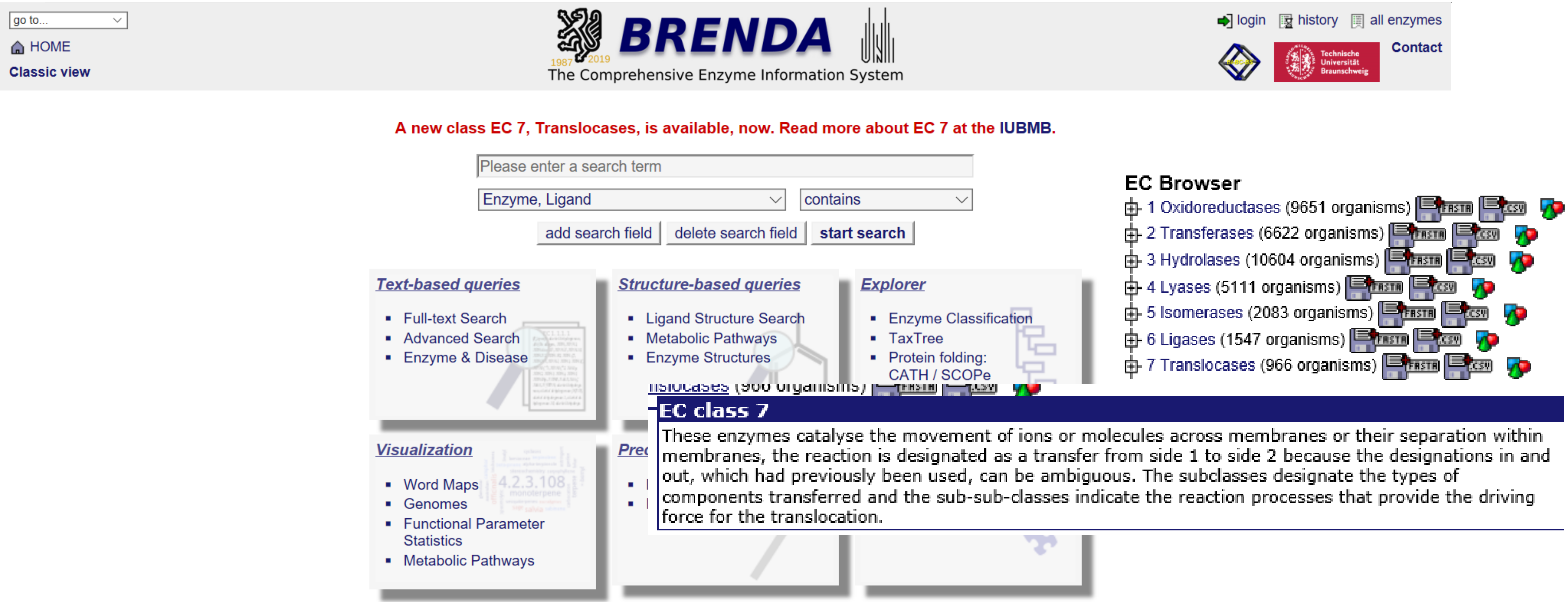
Medina-Carmona, E., Fuchs, J.E., Gavira, J.A., Salido, E., Palomino-Morales, R., Mesa-Torres,

Practical part

Try PDB.

Find out if your sequence has a 3D structure.

Enzyme database: Brenda



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

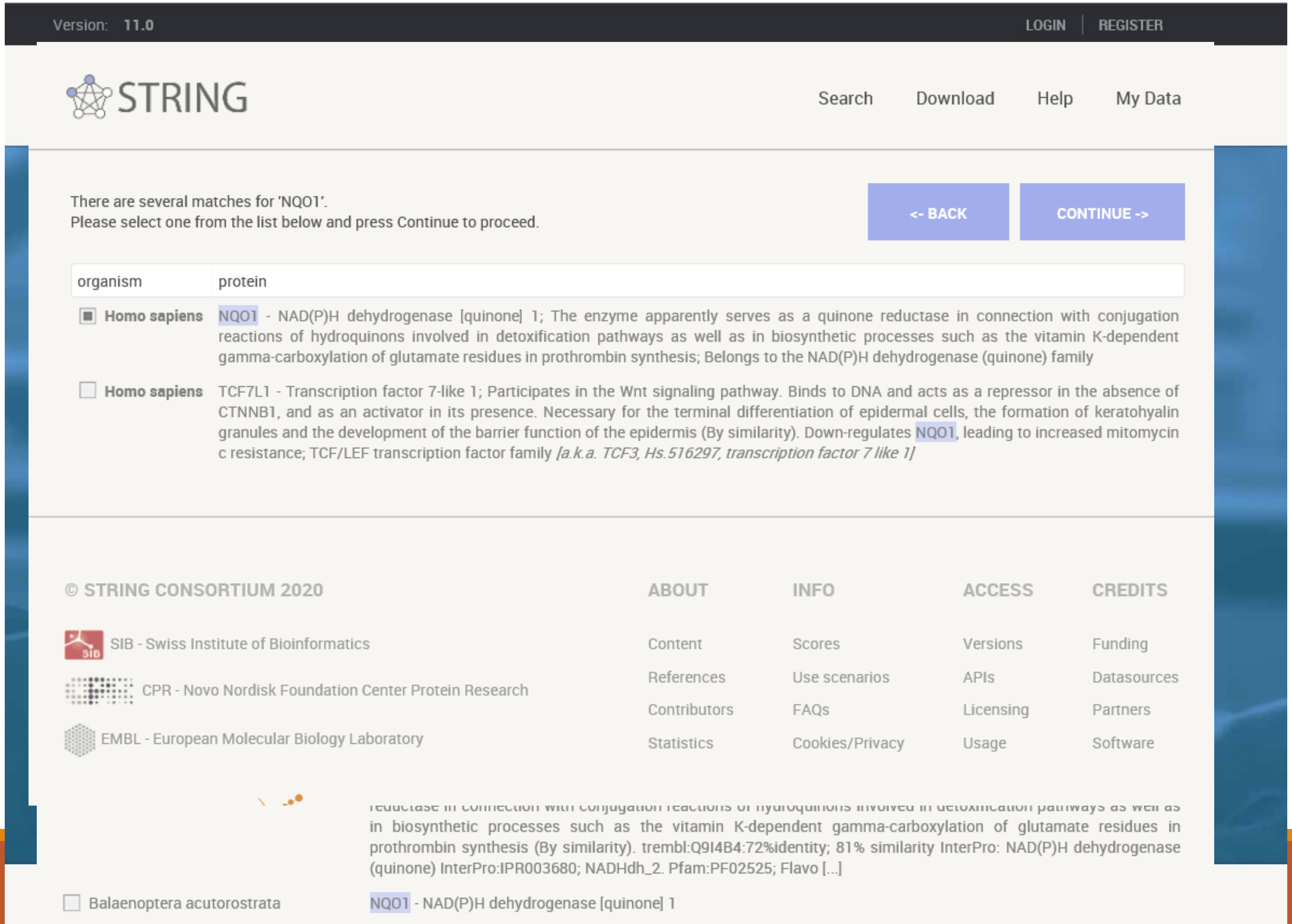
EC Browser

- 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)

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.

Protein interactions



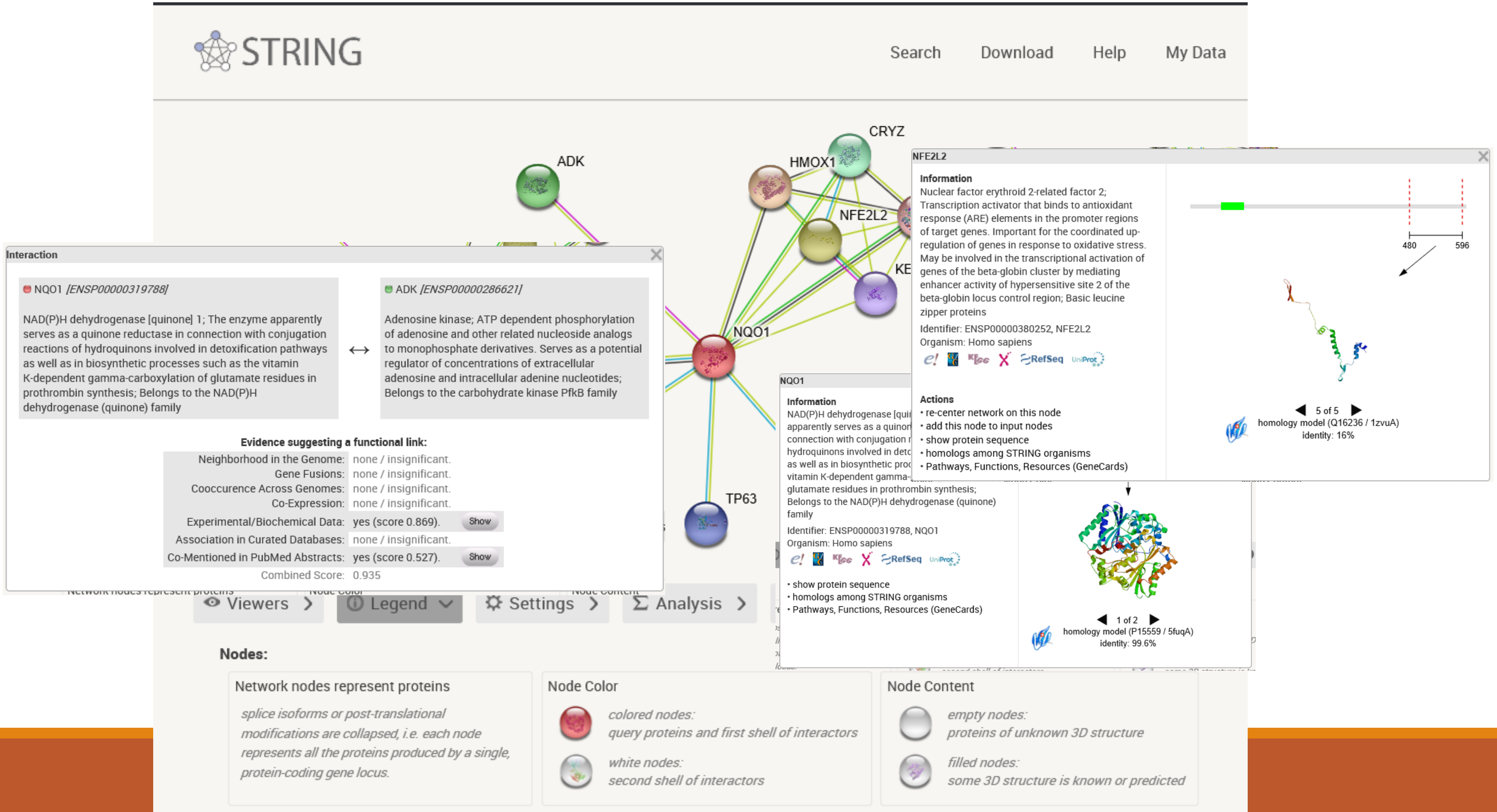
There are several matches for 'NQO1'.
Please select one from the list below and press Continue to proceed.

organism	protein
☒ **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
☐ **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. TCF3, Hs.516297, transcription factor 7 like 1]*

in biosynthetic processes such as the vitamin K-dependent gamma-carboxylation of glutamate residues in prothrombin synthesis (By similarity). trembl:Q9I4B4:72%identity; 81% similarity InterPro: NAD(P)H dehydrogenase (quinone) InterPro:IPR003680; NADHdh_2. Pfam:PF02525; Flavo [...]

☐ Balaenoptera acutorostrata NQO1 - NAD(P)H dehydrogenase [quinone] 1

Protein interactions



Interaction

NQO1 *[ENSP00000319788]*

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

↔

ADK *[ENSP00000286621]*

Adenosine kinase; ATP dependent phosphorylation of adenosine and other related nucleoside analogs to monophosphate derivatives. Serves as a potential regulator of concentrations of extracellular adenosine and intracellular adenine nucleotides; Belongs to the carbohydrate kinase PfkB family

Evidence suggesting a functional link:

Neighborhood in the Genome:	none / insignificant.	
Gene Fusions:	none / insignificant.	
Cooccurence Across Genomes:	none / insignificant.	
Co-Expression:	none / insignificant.	
Experimental/Biochemical Data:	yes (score 0.869).	Show
Association in Curated Databases:	none / insignificant.	
Co-Mentioned in PubMed Abstracts:	yes (score 0.527).	Show
Combined Score:	0.935	

NFE2L2

Information

Nuclear factor erythroid 2-related factor 2; Transcription activator that binds to antioxidant response (ARE) elements in the promoter regions of target genes. Important for the coordinated up-regulation of genes in response to oxidative stress. May be involved in the transcriptional activation of genes of the beta-globin cluster by mediating enhancer activity of hypersensitive site 2 of the beta-globin locus control region; Basic leucine zipper proteins

Identifier: ENSP00000380252, NFE2L2

Organism: Homo sapiens

Actions

- re-center network on this node
- add this node to input nodes
- show protein sequence
- homologs among STRING organisms
- Pathways, Functions, Resources (GeneCards)

480 596

◀ 5 of 5 ▶

homology model (Q16236 / 1zvuA)

identity: 16%

NQO1

Information

NAD(P)H dehydrogenase [qui
apparently serves as a quinon
connection with conjugation r
hydroquinons involved in detc
as well as in biosynthetic proc
vitamin K-dependent gamma-
glutamate residues in prothrombin synthesis;
Belongs to the NAD(P)H dehydrogenase (quinone)
family

Identifier: ENSP00000319788, NQO1

Organism: Homo sapiens

- show protein sequence
- homologs among STRING organisms
- Pathways, Functions, Resources (GeneCards)

◀ 1 of 2 ▶

homology model (P15559 / 5fuqA)

identity: 99.6%

Nodes:

Network nodes represent proteins

splice isoforms or post-translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.

Node Color

colored nodes:
query proteins and first shell of interactors

white nodes:
second shell of interactors

Node Content

empty nodes:
proteins of unknown 3D structure

filled nodes:
some 3D structure is known or predicted

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) Identify and download five similar sequnces.

2) Compare your sequence with the „same“ sequence from mouse, how identical are they?

2) Prepare multiple alignment of the five sequences, snip the phylogeny tree.

4) Is there a 3D structure? Snip one figure.

5) Is your protein an enzyme? Find E.C.

E.g use „výstřižky“ „snipping tool“

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

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?

Ex2: NQO1 isoforms

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

How many isofroms are there?

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

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?