Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2023/2024

„Protein bioinformatics II“

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

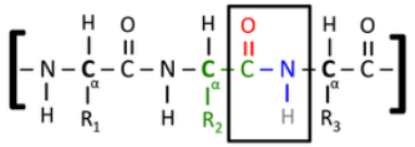
Finding 3-D structure and the domain organization of proteins

Finding all proteins that share a similar sequence and Classifying proteins into families

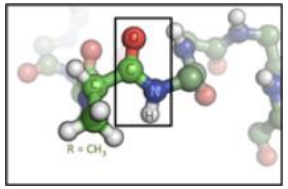
Finding evolutionary relationships between proteins, drawing proteins’ family trees

Computing the optimal alignment between two or more protein sequences

…



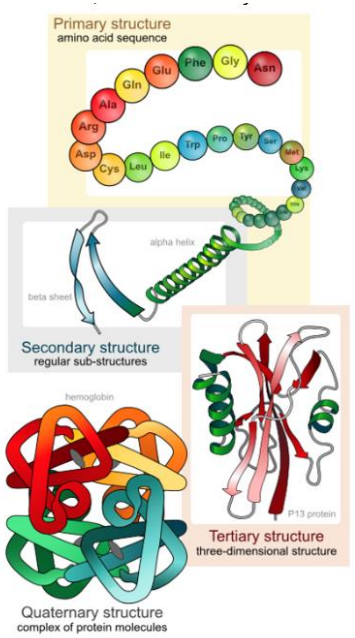
Proteins



20 Aminoacids – primary structure:

(Frederick Sanger-1958 Nobel prize for insulin sequencing)

Secondary structure
Tertiary structure
Quaternary structure



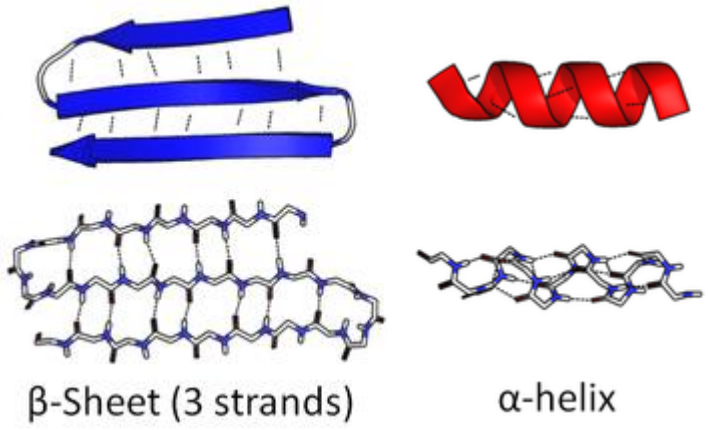
1-letter code	3-letter code	Amino acid	Possible codons
A	Ala	Alanine	GCA, GCC, GCG, GCT
B	Asx	Asparagine or Aspartic acid	AAC, AAT, GAC, GAT
C	Cys	Cysteine	TGC, TGT
D	Asp	Aspartic acid	GAC, GAT
E	Glu	Glutamic acid	GAA, GAG
F	Phe	Phenylalanine	TTC, TTT
G	Gly	Glycine	GGA, GGC, GGG, GGT
H	His	Histidine	CAC, CAT
I	Ile	Isoleucine	ATA, ATC, ATT
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	CTA, CTC, CTG, CTT, TTA, TTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAC, AAT
P	Pro	Proline	CCA, CCC, CCG, CCT
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	AGA, AGG, CGA, CGC, CGG, CGT
S	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
T	Thr	Threonine	ACA, ACC, ACG, ACT
V	Val	Valine	GTA, GTC, GTG, GTT
W	Trp	Tryptophan	TGG
X	X	Stop codon	TAA, TAG, TGA
Y	Tyr	Tyrosine	TAC, TAT
Z	Glx	Glutamine or Glutamic acid	CAA, CAG, GAA, GAG

SEQUENCE⇨STRUCTURE⇨FUNCTION

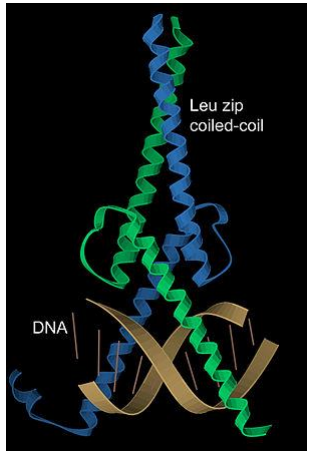
Protein domain

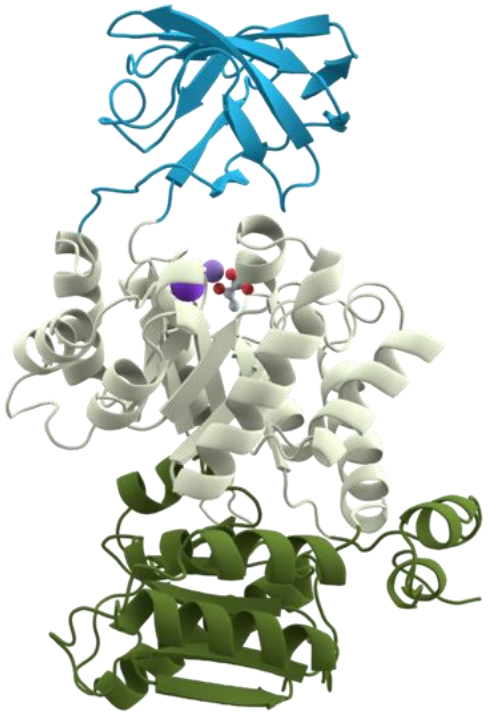
- region of a protein's polypeptide chain that folds independently from the rest
- forms a compact folded three-dimensional structure
- many proteins consist of several domains

Secondary



Tertiary



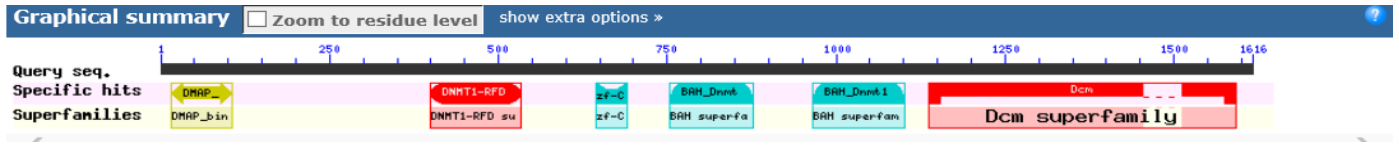


Conserved domain search

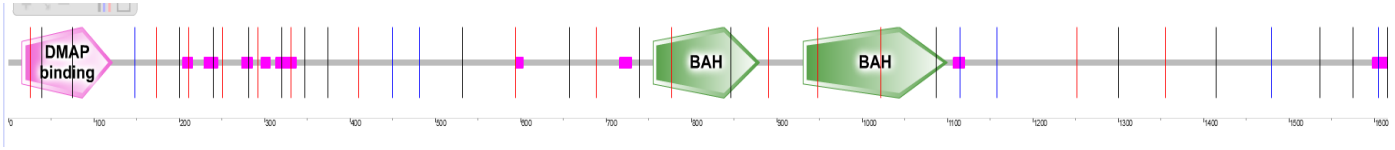
SEQUENCE➪STRUCTURE➪FUNCTION

- Conserved domain databases:

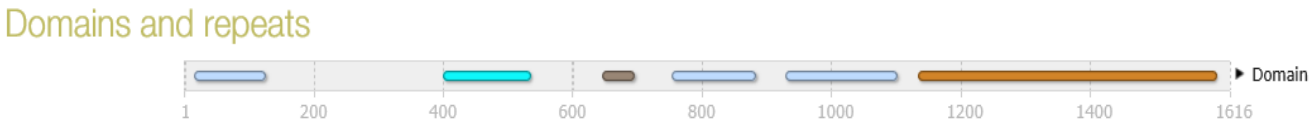
NCBI/CDD



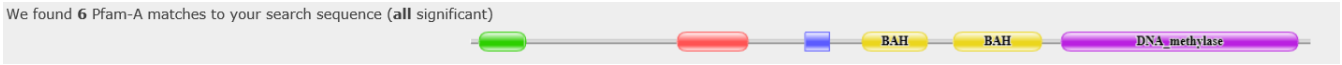
SMART



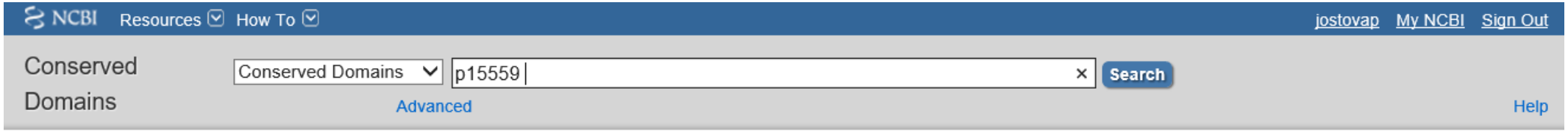
EMBL/InterPro

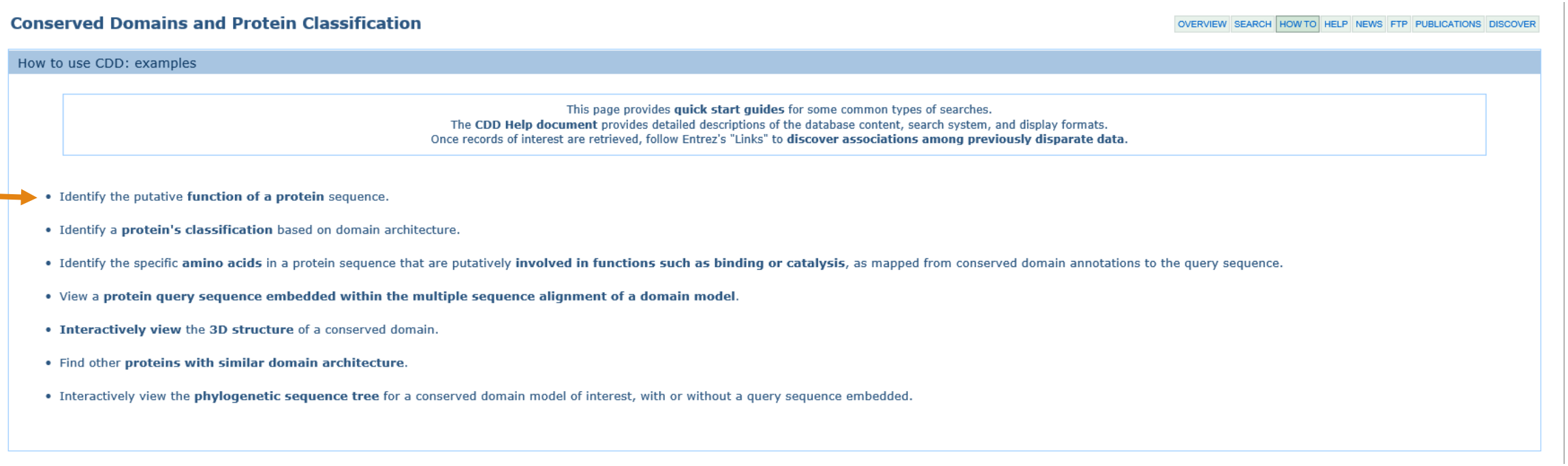


Pfam

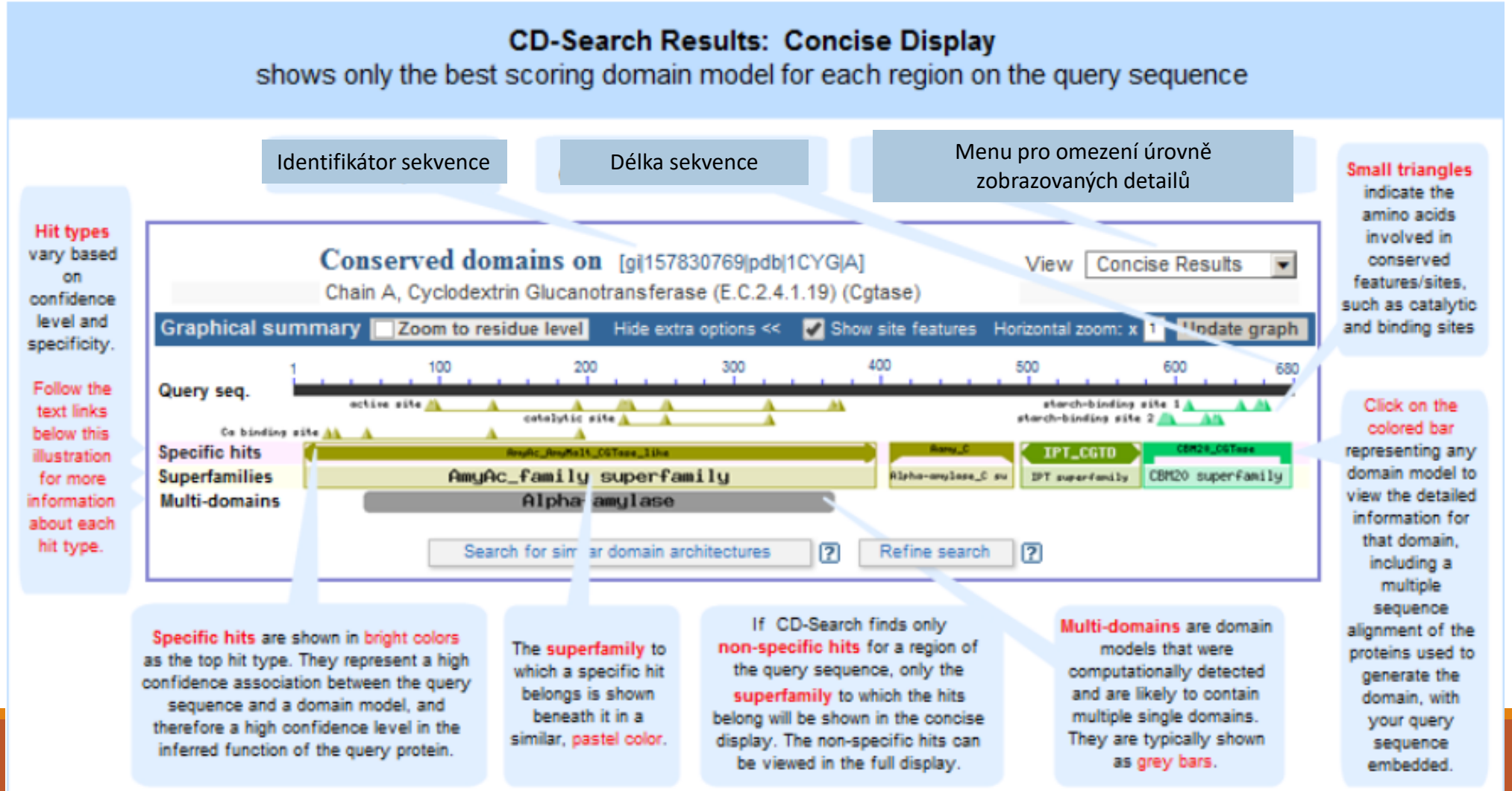


Conserved domain search - CD (NCBI)

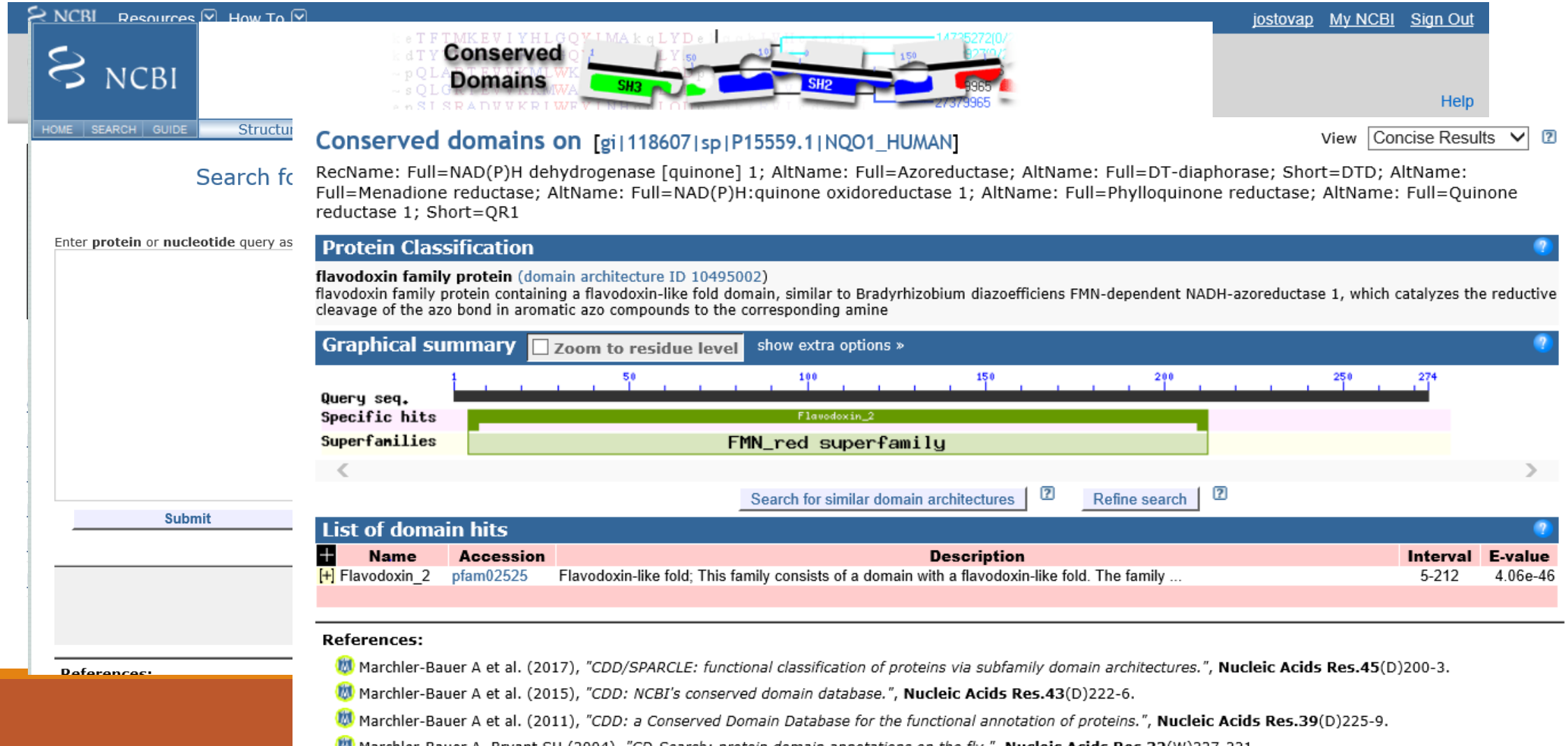




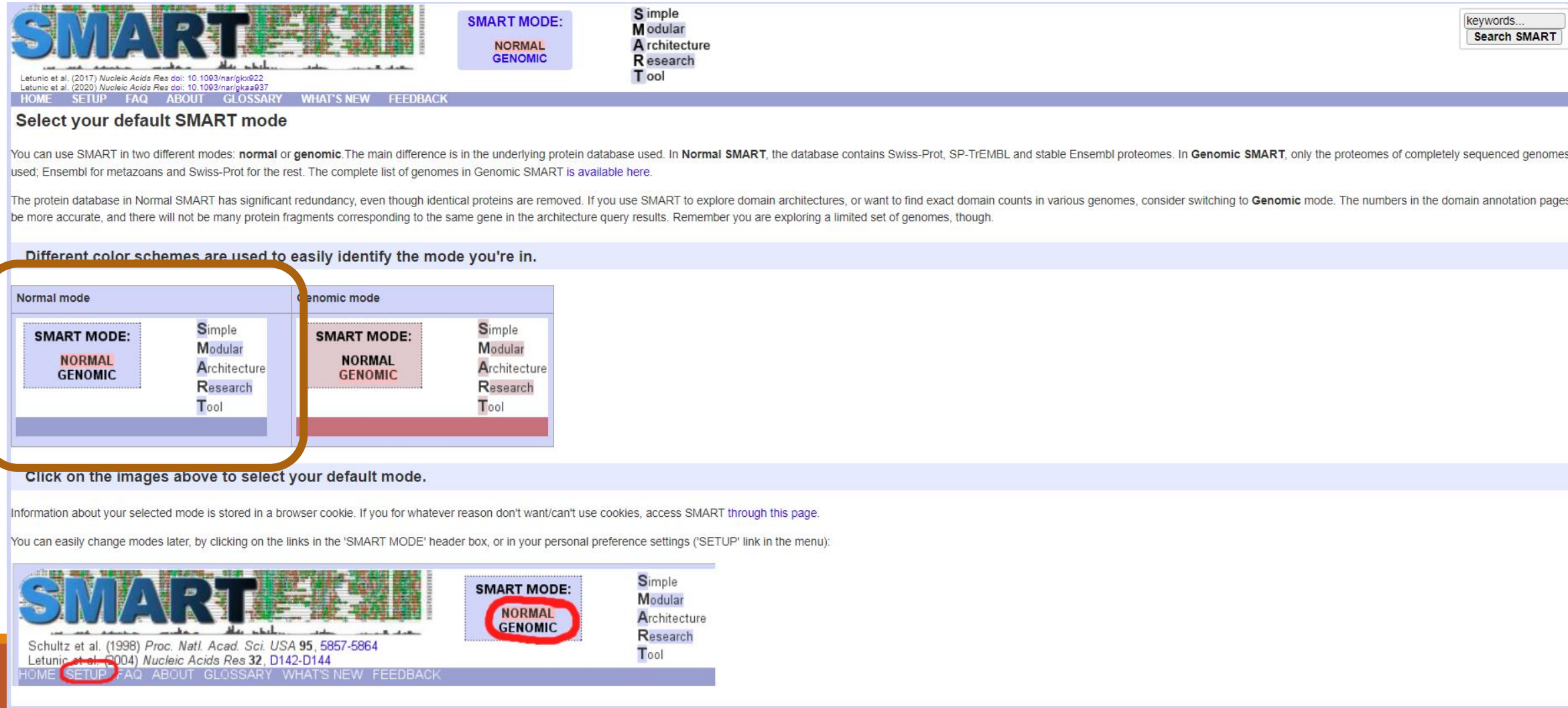
Conserved domain search - CD (NCBI)



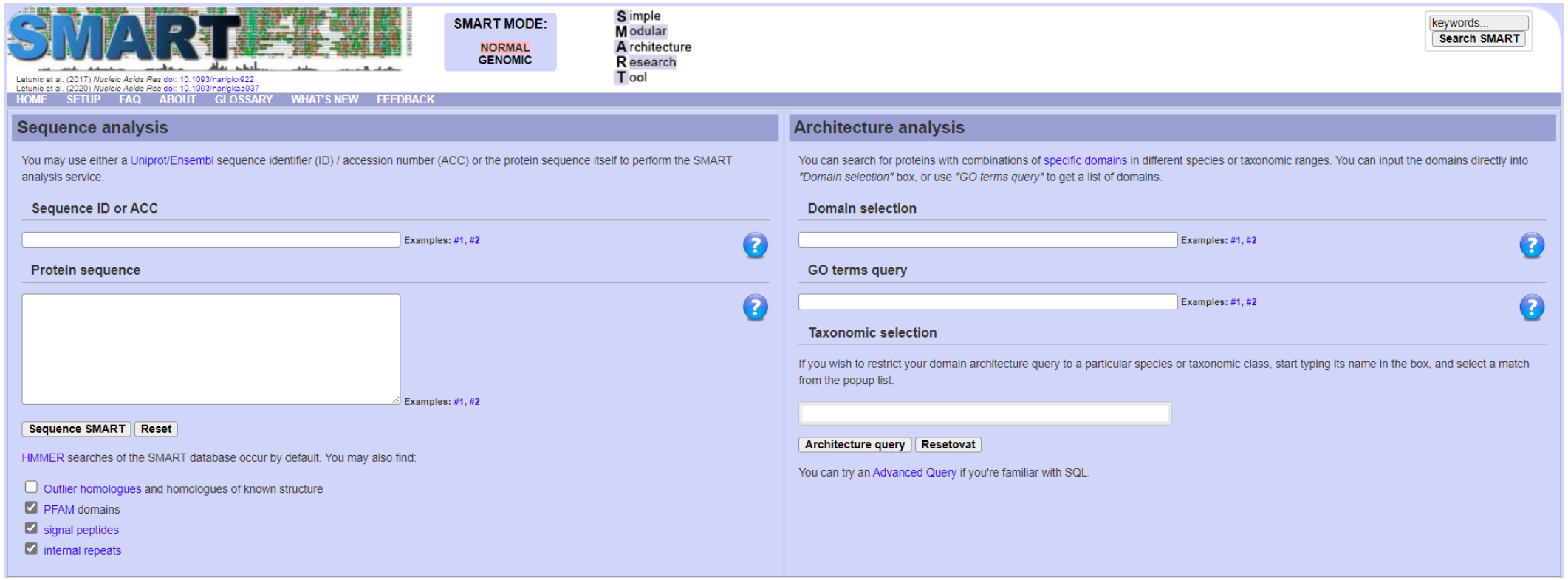
Conserved domain search / NQO1



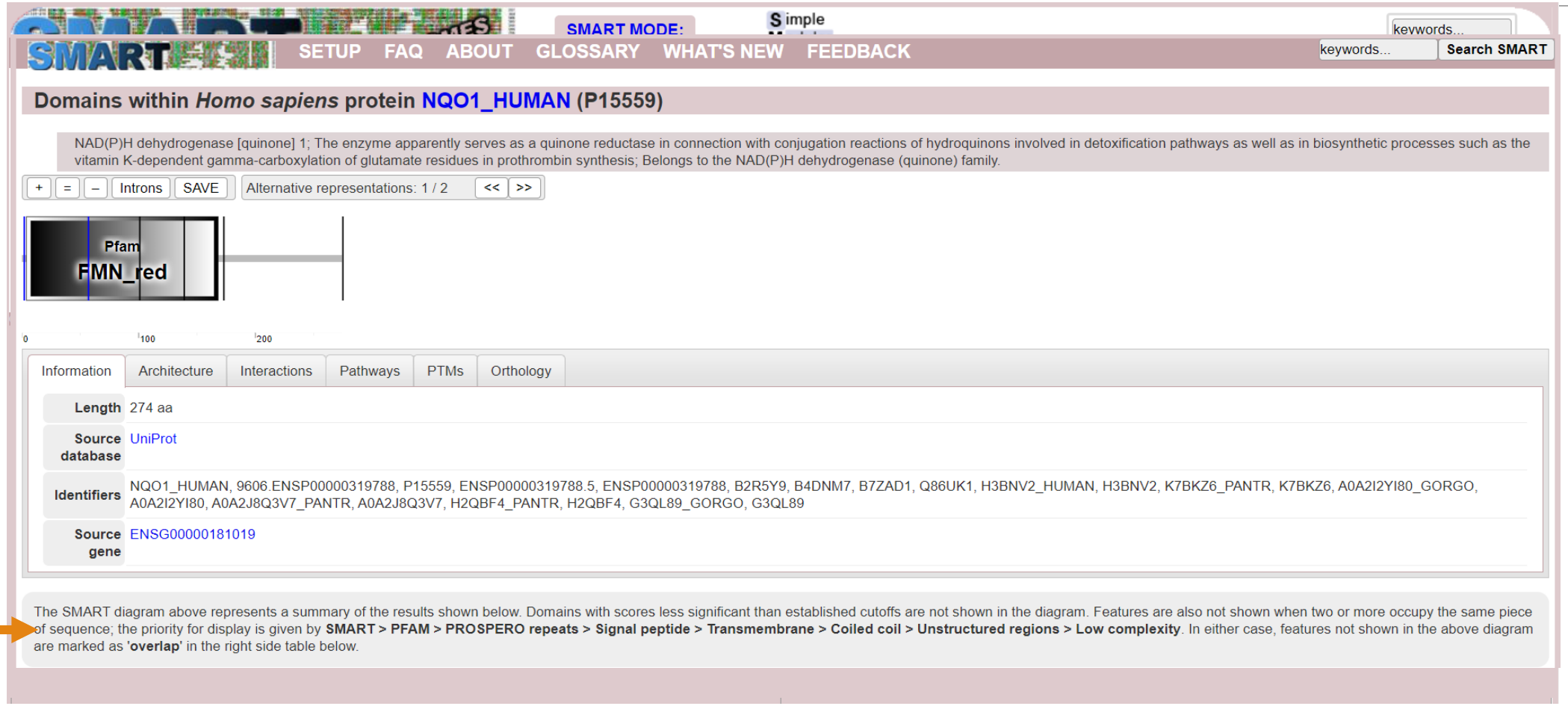
Conserved domain search - SMART



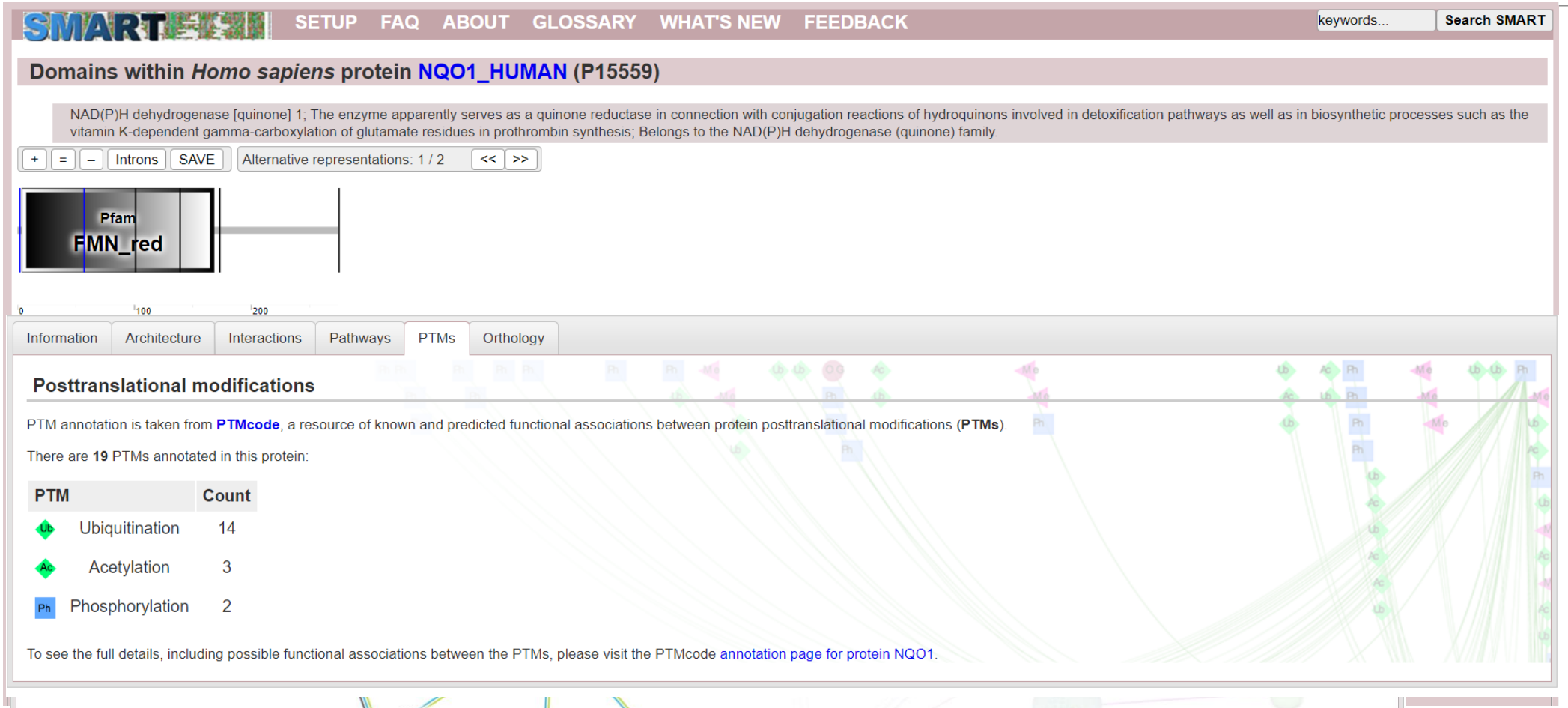
Conserved domain search - SMART



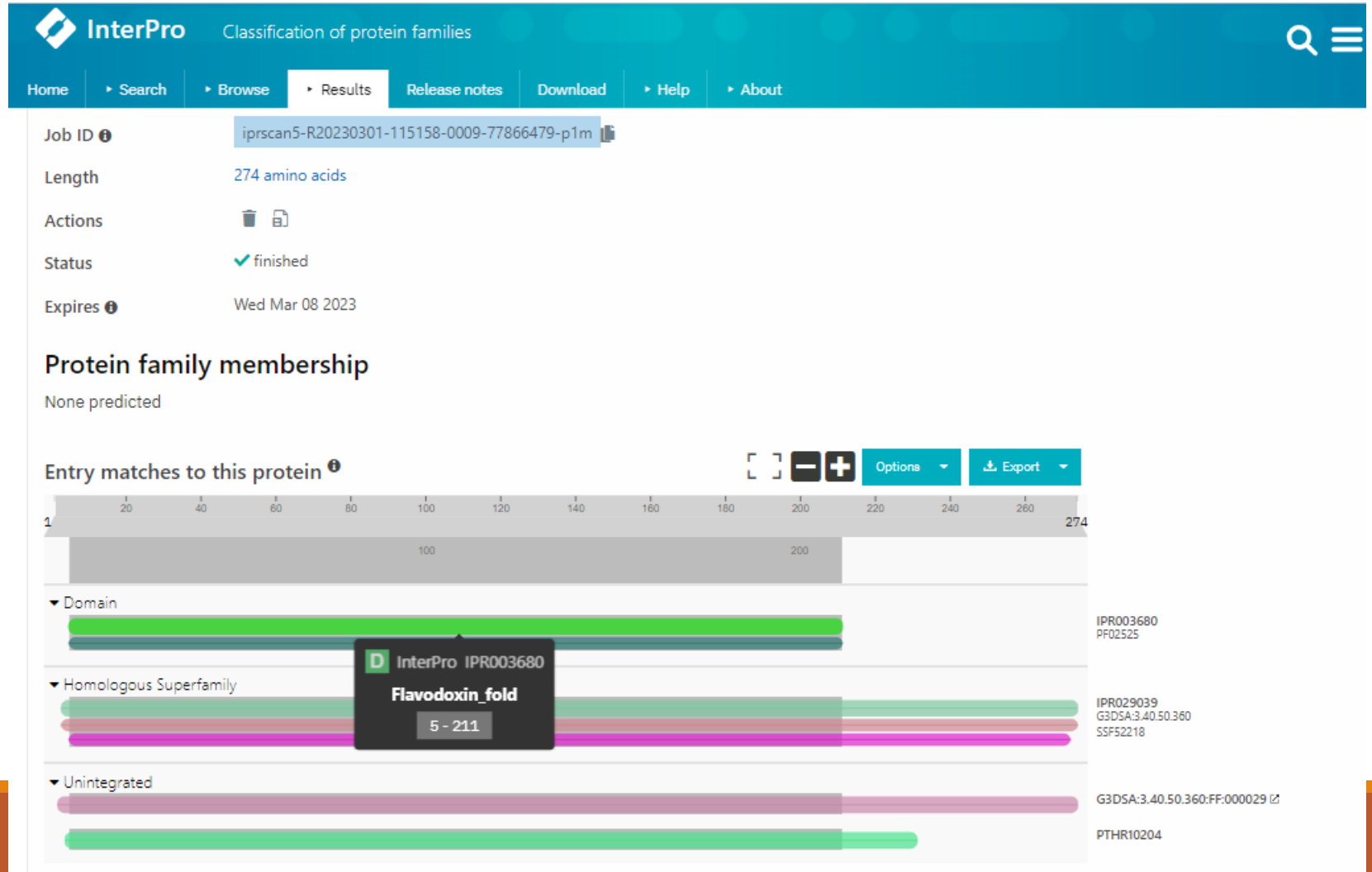
Conserved domain search - SMART



Conserved domain search - SMART



Conserved domain search - InterPro



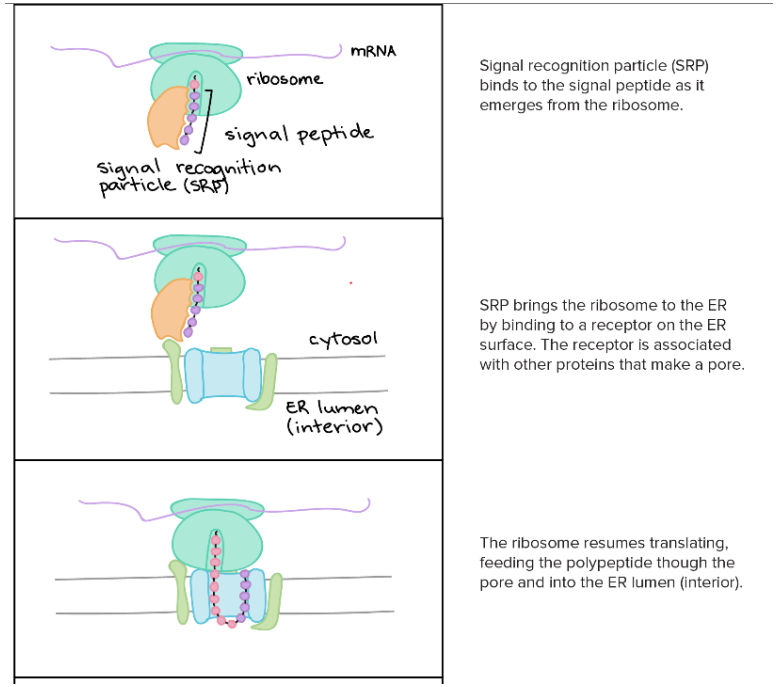
Practical part

Try
CD / SMART/ InterPro
search

Find domains in your sequnce

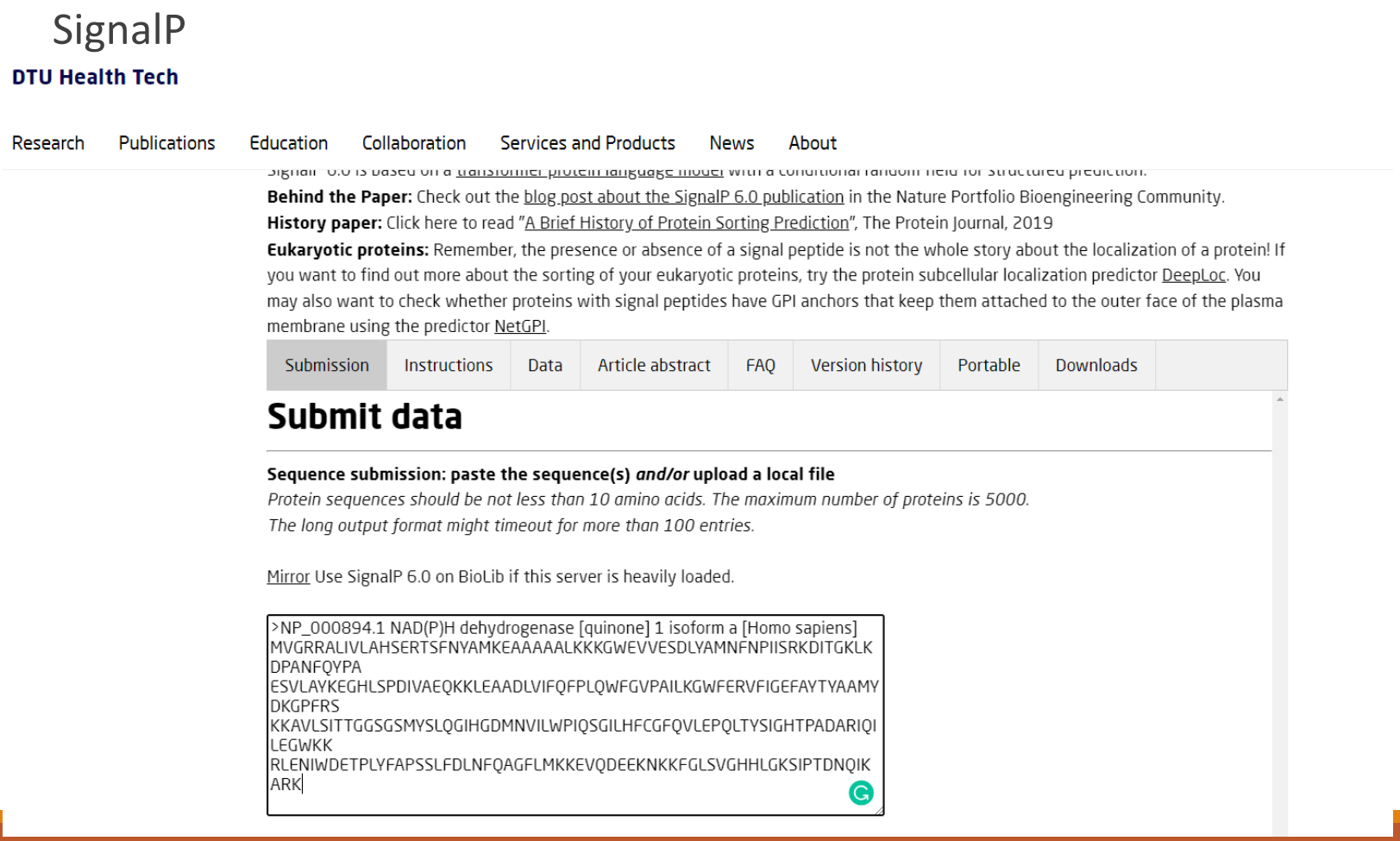
ER signal peptide prediction

Endoplasmic reticulum signal peptide: 15-60 amino acids on protein N-terminus



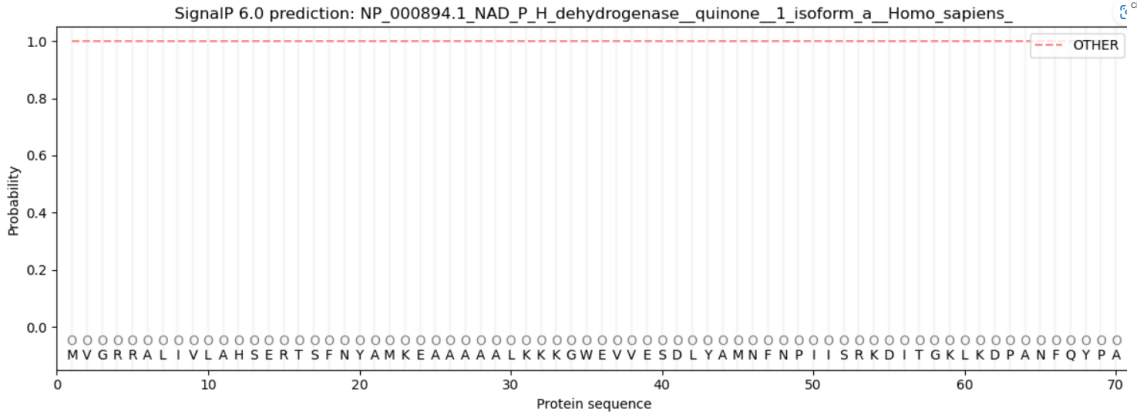
https://www.khanacademy.org/science/biology/gene-expression-central-dogma#central-dogma-transcription

Signal peptides

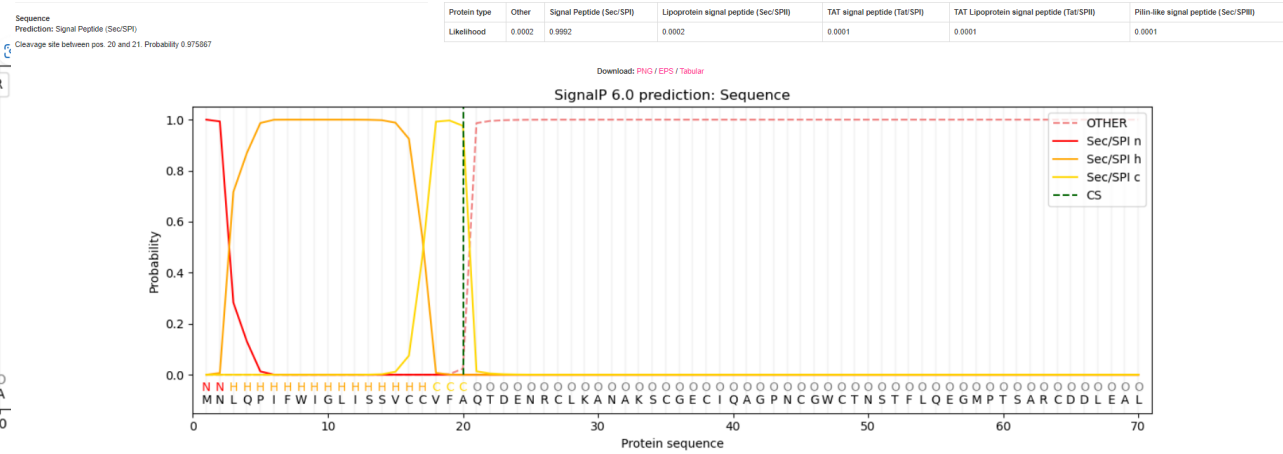


Signal peptides

SignalP

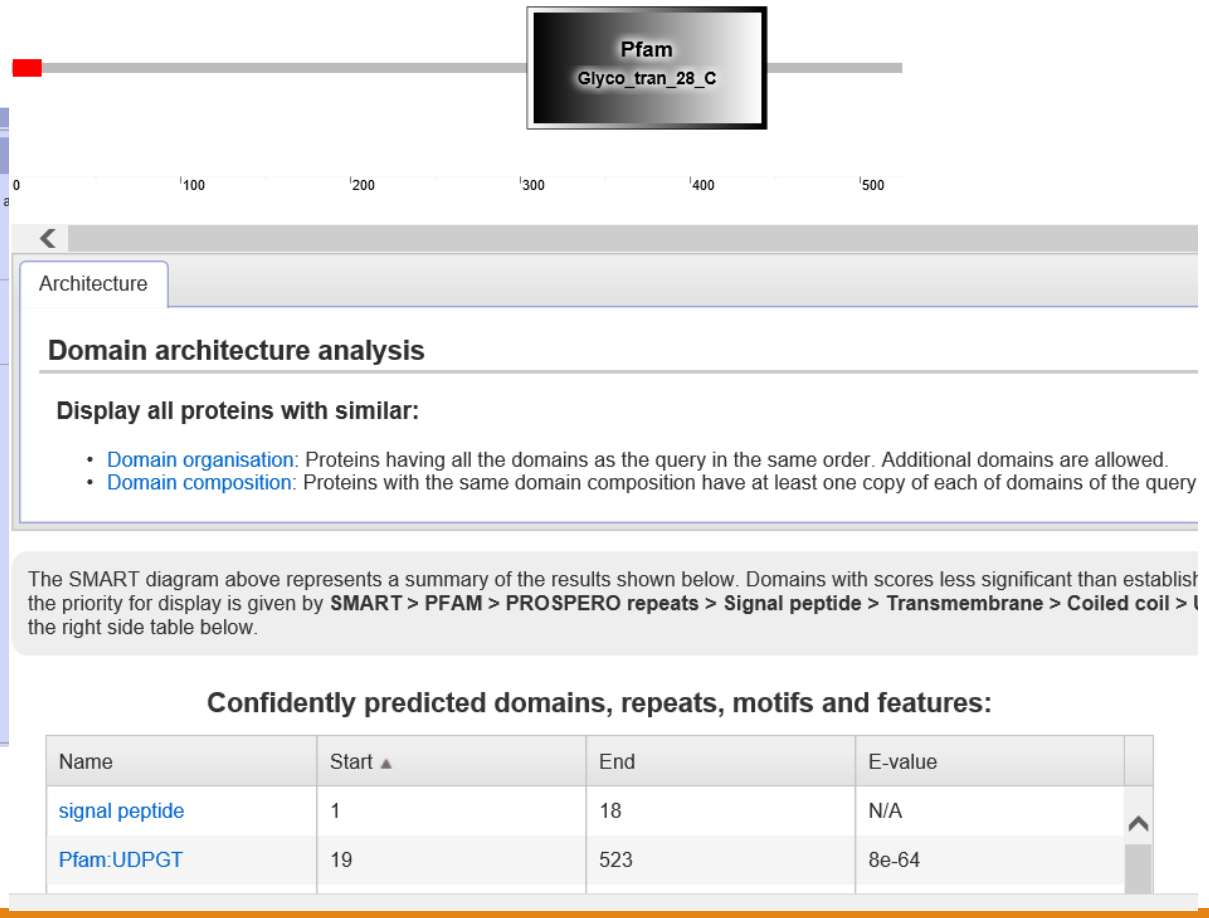


protein does not have signal peptide



Protein **has** signal peptide
(with certain probability)

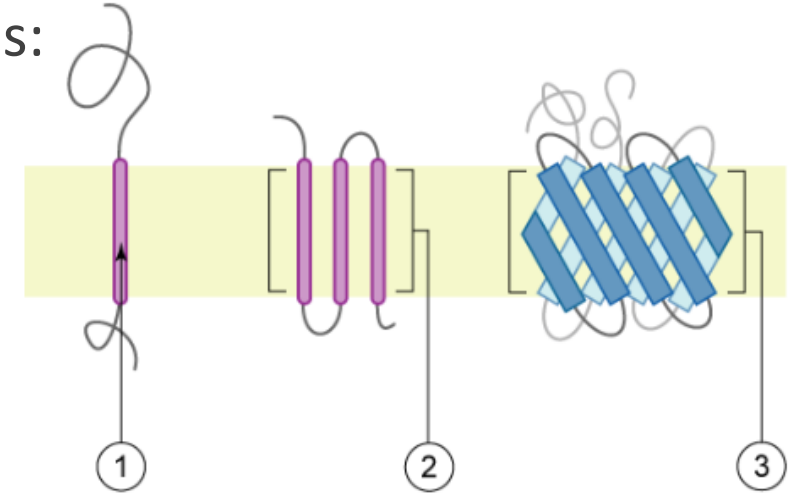
Signal peptides

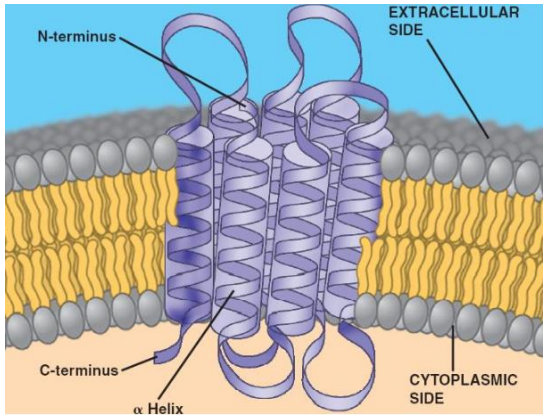


search for signal peptide in your sequnce

Prediction of transmembrane helices

Transmembrane proteins:



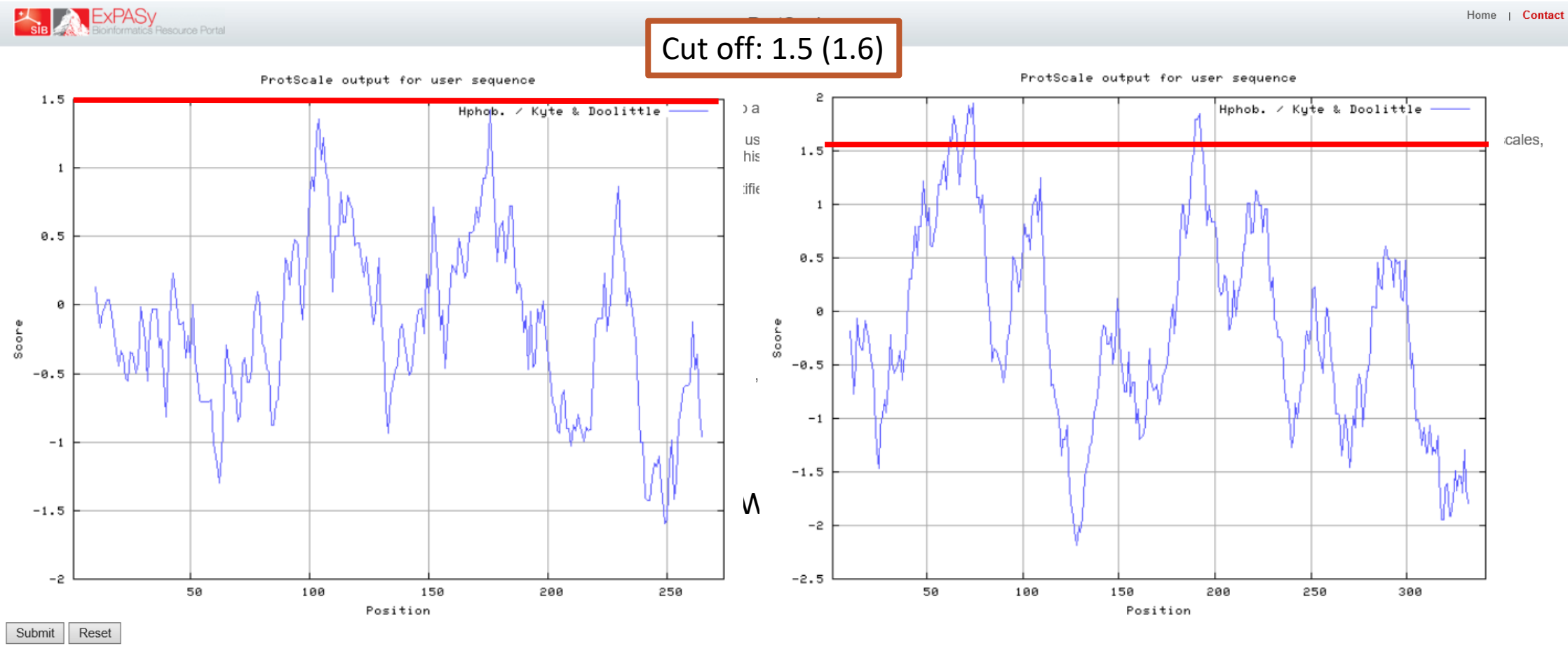


Amino acid Hydrofobicity

- various programs – different alghoritms – different results
- Topological predictions (estimation of in and out topology)

Prediction of transmembrane helices

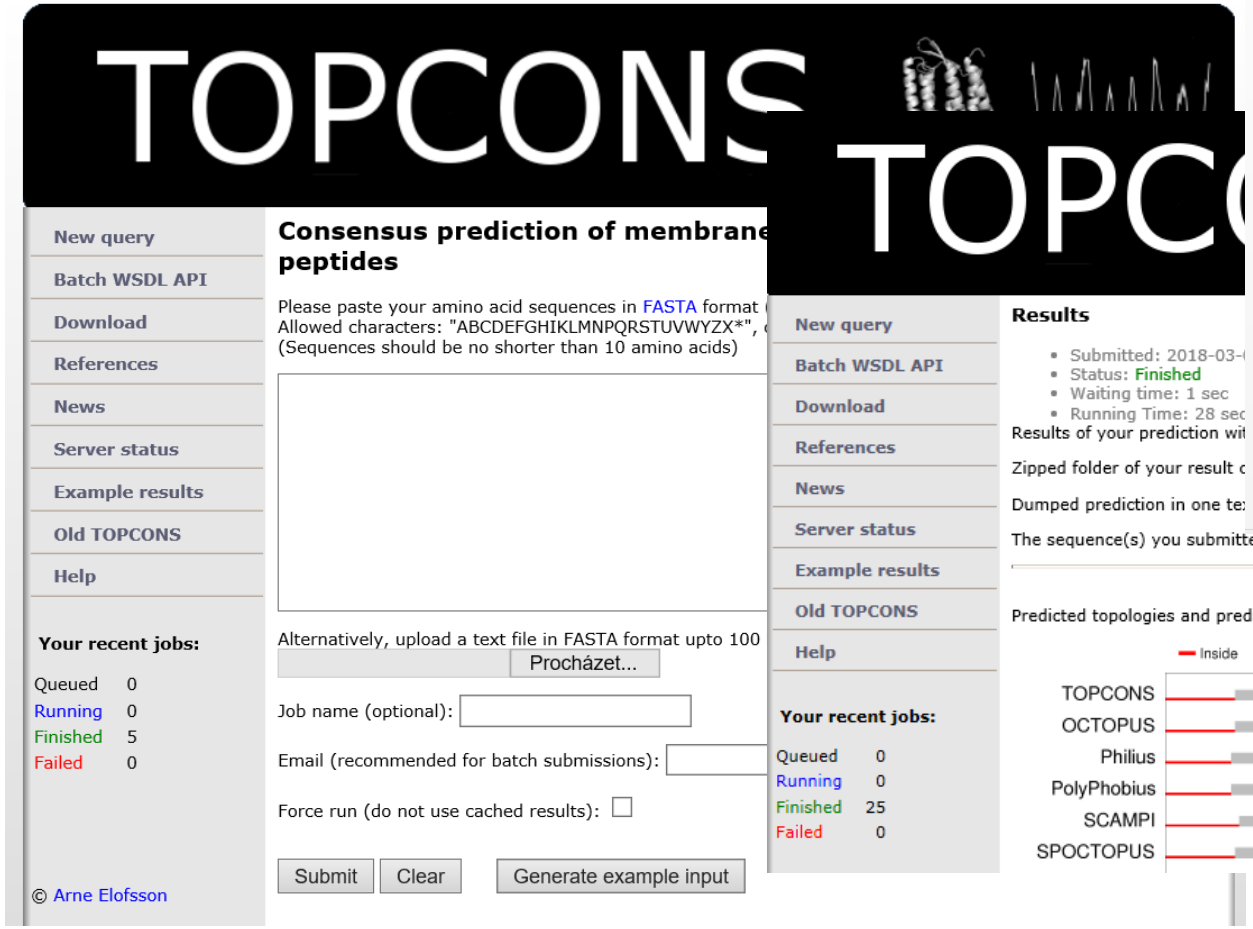
Profile of amino acids hydrofobicity

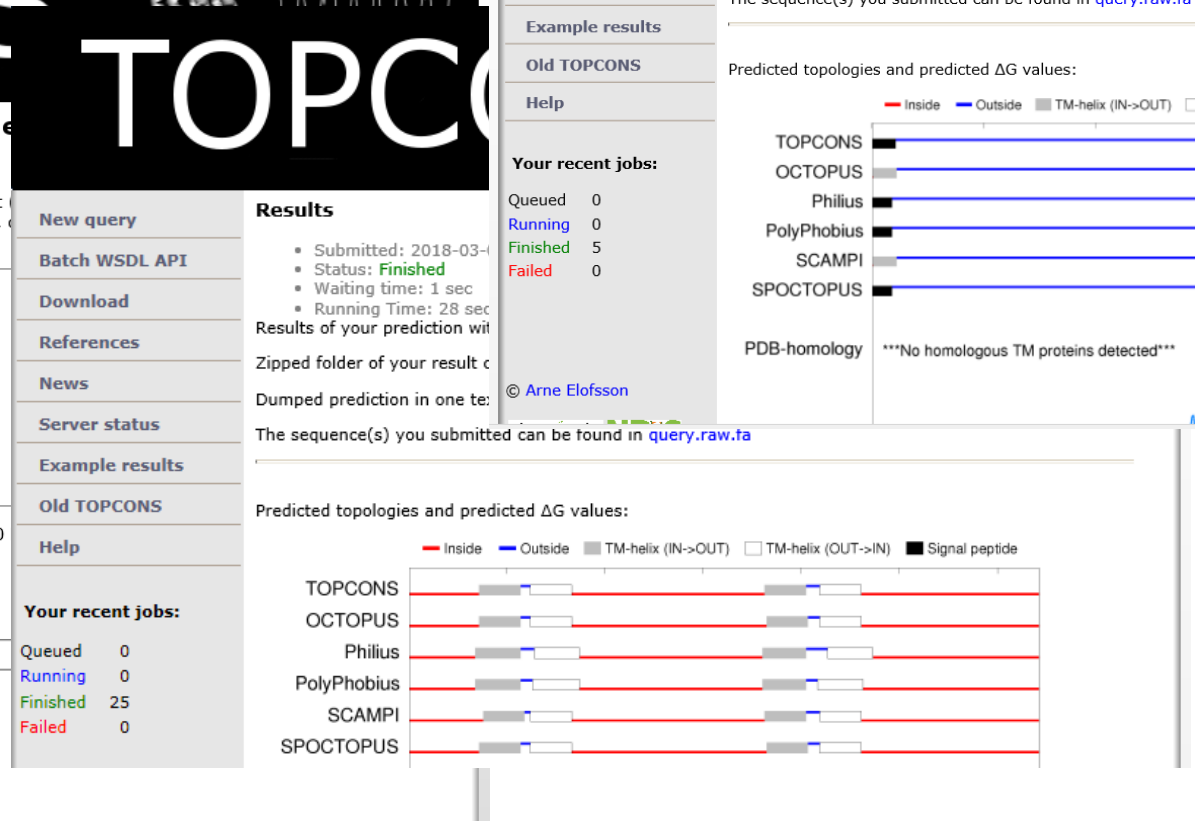


TMHMM


CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS
EVENTS
NEWS
RESEARCH GROUPS
CBS PREDICTION SERVERS
CBS DATA SETS
PUBLICATIONS
EDUCATION
STAFF
CONTACT
ABOUT CBS
INTERNAL
CBS BIOINFORMATICS TOOLS
CBS COURSES
OTHER BIOINFORMATICS LINKS
CENTER FOR BIOLOGICAL SEQU
CBS >> CBS Prediction Servers >> TMHMM
TMHMM Server v. 2.0
Prediction of transmembrane helices in proteins
SUBMISSION
Submission of a local file in FASTA form
Procházet...
OR by pasting sequence(s) in FASTA for
Output format:
Extensive, with graphics
Extensive, no graphics
One line per protein
Other options:
Use old model (version 1)
Submit
Clear
Restrictions:
At most 10,000 sequences and 4,000,000 a
Confidentiality:
The sequences are kept confidential and wi
TMHMM result
HELP with output formats
no TM helix
WEBSEQUENCE Length: 274
WEBSEQUENCE Number of predicted TMHs: 0
WEBSEQUENCE Exp number of AAs in TMHs: 0.20324
WEBSEQUENCE Exp number, first 60 AAs: 0
WEBSEQUENCE Total prob of N-in: 0.04315
WEBSEQUENCE TMHMM2.0 outside 1 274
TMHMM posterior probabilities for WEBSEQUENCE
probability
transmembrane
inside
outside
TMHMM result
HELP with output formats
five predicted TM helices
WEBSEQUENCE Length: 321
WEBSEQUENCE Number of predicted TMHs: 5
WEBSEQUENCE Exp number of AAs in TMHs: 108.47546
WEBSEQUENCE Exp number, first 60 AAs: 21.62676
WEBSEQUENCE Total prob of N-in: 0.05151
WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE TMHMM2.0 outside 1 34
WEBSEQUENCE TMHMM2.0 TMhelix 35 57
WEBSEQUENCE TMHMM2.0 inside 58 63
WEBSEQUENCE TMHMM2.0 TMhelix 64 86
WEBSEQUENCE TMHMM2.0 outside 87 95
WEBSEQUENCE TMHMM2.0 TMhelix 96 118
WEBSEQUENCE TMHMM2.0 inside 119 179
WEBSEQUENCE TMHMM2.0 TMhelix 180 202
WEBSEQUENCE TMHMM2.0 outside 203 216
WEBSEQUENCE TMHMM2.0 TMhelix 217 236
WEBSEQUENCE TMHMM2.0 inside 237 321
TMHMM posterior probabilities for WEBSEQUENCE
probability
transmembrane
inside
outside

TOPCONS


TOPCONS
New query
Batch WSDL API
Download
References
News
Server status
Example results
Old TOPCONS
Help
Your recent jobs:
Queued 0
Running 0
Finished 5
Failed 0
© Arne Elofsson
Consensus prediction of membrane peptides
Please paste your amino acid sequences in FASTA format
Allowed characters: "ABCDEFGHIKLMNPQRSTUVWYZX*",
(Sequences should be no shorter than 10 amino acids)
Alternatively, upload a text file in FASTA format upto 100
Procházet...
Job name (optional):
Email (recommended for batch submissions):
Force run (do not use cached results):
Submit
Clear
Generate example input


TOPCONS
New query
Batch WSDL API
Download
References
News
Server status
Example results
Old TOPCONS
Help
Your recent jobs:
Queued 0
Running 0
Finished 25
Failed 0
Results
Submitted: 2018-03-
Status: Finished
Waiting time: 1 sec
Running Time: 28 sec
Results of your prediction wit
Zipped folder of your result
Dumped prediction in one te
The sequence(s) you submitted can be found in query.raw.fa
Predicted topologies and predicted ΔG values:
Inside
Outside
TM-helix (IN->OUT)
TM-helix (OUT->IN)
Signal peptide
TOPCONS
OCTOPUS
Philius
PolyPhobius
SCAMPI
SPOCTOPUS

TOPCONS
New query
Batch WSDL API
Download
References
News
Server status
Example results
Old TOPCONS
Help
Your recent jobs:
Queued 0
Running 0
Finished 5
Failed 0
© Arne Elofsson
Results
Submitted: 2018-03-05 15:59:14
Status: Finished
Waiting time: 0 sec
Running Time: 0 sec
Results of your prediction with jobid: rst_BKOlKK
Zipped folder of your result can be found in rst_BKOlKK.zip
Dumped prediction in one text file can be found in query.result.txt
The sequence(s) you submitted can be found in query.raw.fa
Predicted topologies and predicted ΔG values:
Inside
Outside
TM-helix (IN->OUT)
TM-helix (OUT->IN)
Signal peptide
TOPCONS
OCTOPUS
Philius
PolyPhobius
SCAMPI
SPOCTOPUS
PDB-homology
No homologous TM proteins detected
15

Phobius

Phobius

A combined transmembrane topology and signal peptide predictor

Normal prediction | Constrained prediction | PolyPhobius | Instructions

Normal prediction

Paste your protein sequence here in Fasta format:

Or: Select the sequence file you wish to use Zvolit soubor Nevybrán žádný soubor

Select output format:

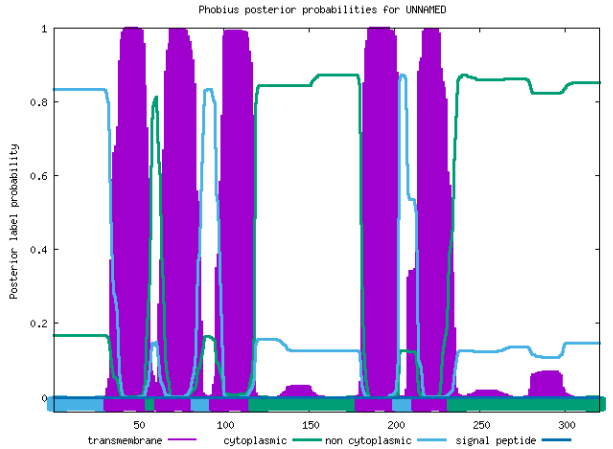
- ○ Short
- ○ Long without Graphics
- ◉ Long with Graphics

Odeslat Resetovat

Phobius prediction

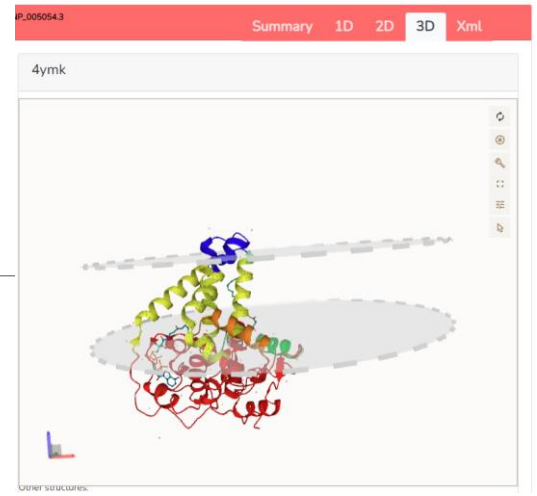
Prediction of UNNAMED

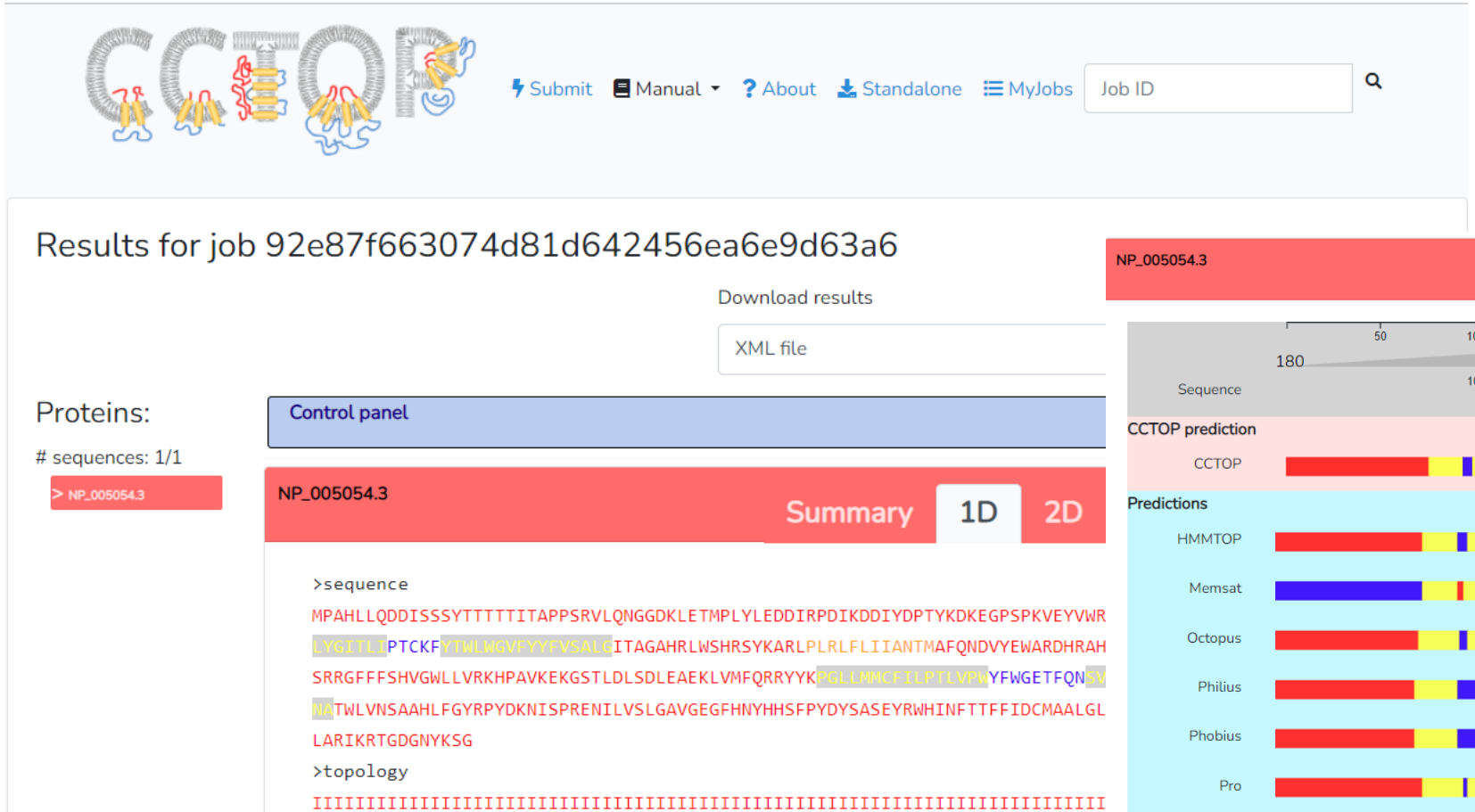
```
ID   UNNAMED
FT   TOPO_DOM     1     33       NON CYTOPLASMIC.
FT   TRANSMEM    34     57
FT   TOPO_DOM    58     63       CYTOPLASMIC.
FT   TRANSMEM    64     84
FT   TOPO_DOM    85     95       NON CYTOPLASMIC.
FT   TRANSMEM    96    118
FT   TOPO_DOM   119    180       CYTOPLASMIC.
FT   TRANSMEM   181    202
FT   TOPO_DOM   203    213       NON CYTOPLASMIC.
FT   TRANSMEM   214    234
FT   TOPO_DOM   235    320       CYTOPLASMIC.
//
```

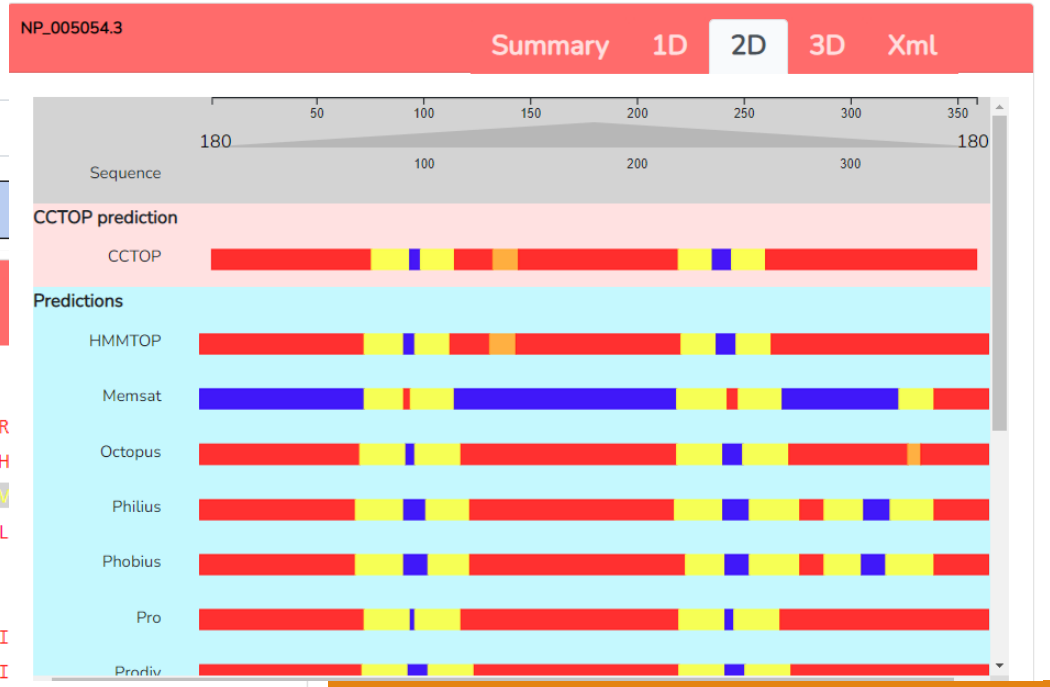


The probability data used in the plot is found here, and the gnuplot script is here.

CCTOP

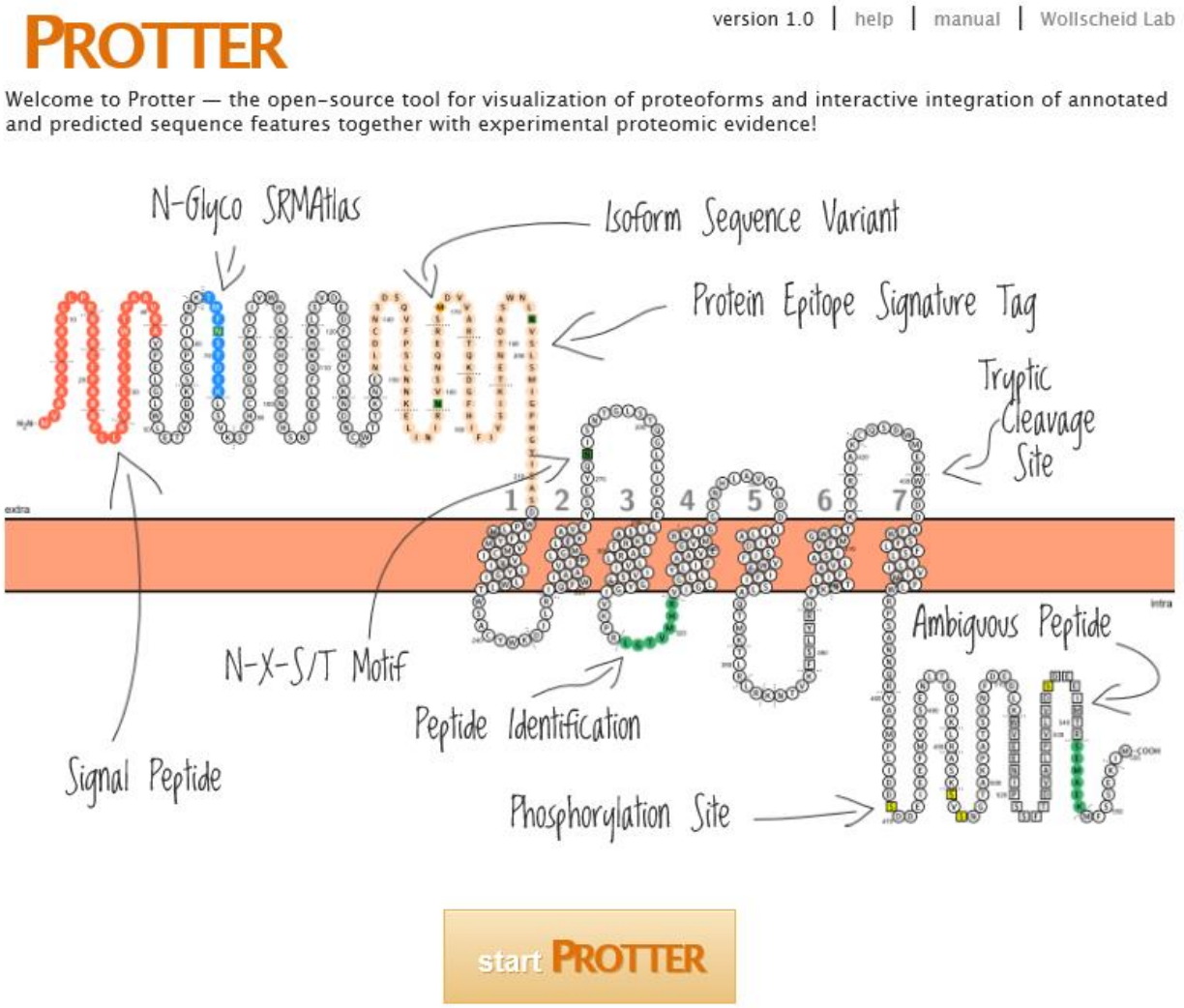

Summary
1D
2D
3D
Xml
4ymk


Submit
Manual
About
Standalone
MyJobs
Job ID
Results for job 92e87f663074d81d642456ea6e9d63a6
Download results
XML file
Proteins:
sequences: 1/1
> NP_005054.3
Control panel
NP_005054.3
Summary
1D
2D
>sequence
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVEYVWR
LYGITLIPTCKFVTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
SRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGETFQNSV
NATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFTTFFIDCMAALGL
LARIKRTGDGNYKSG
>topology
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
MMMMMMMOOOOOMMMMMMMMMMMMMMMIIIIIIIIIIIIIIIIIILLLLLLLLLLLIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
MMMMMMMMMMMMMMMMOOOOOOOOOMMMMMMMMMMMMMM
MMIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIII


NP_005054.3
Summary
1D
2D
3D
Xml
50
100
150
200
250
300
350
180
180
100
200
300
Sequence
CCTOP prediction
CCTOP
Predictions
HMMTOP
Memsat
Octopus
Philius
Phobius
Pro
Prodiv

PROTTER-figure!

-creates figure from the UniProt data



PROTTER-figure!

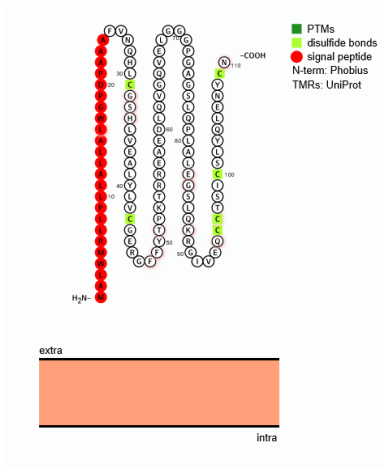
-creates figure from the UniProt data

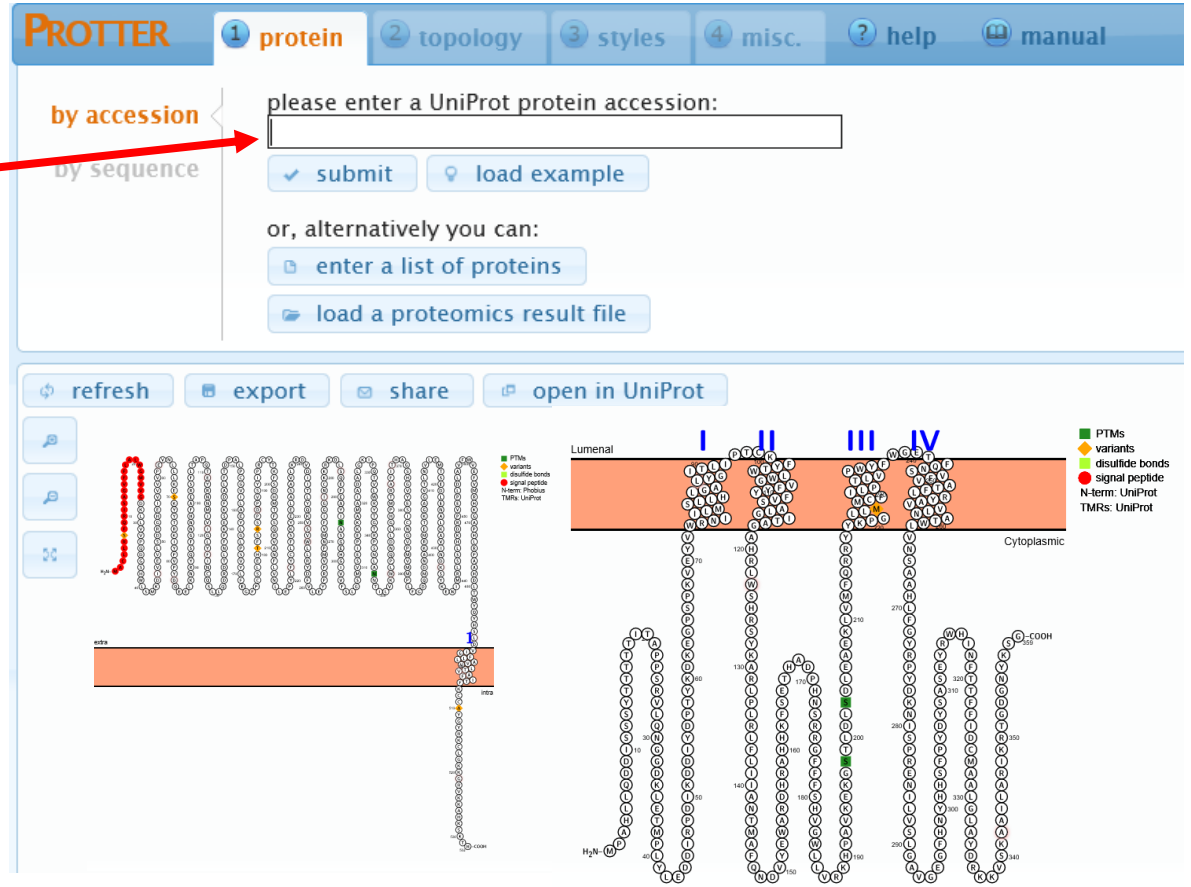
-uses UniProt ID:

UGT1A6 (P19224)

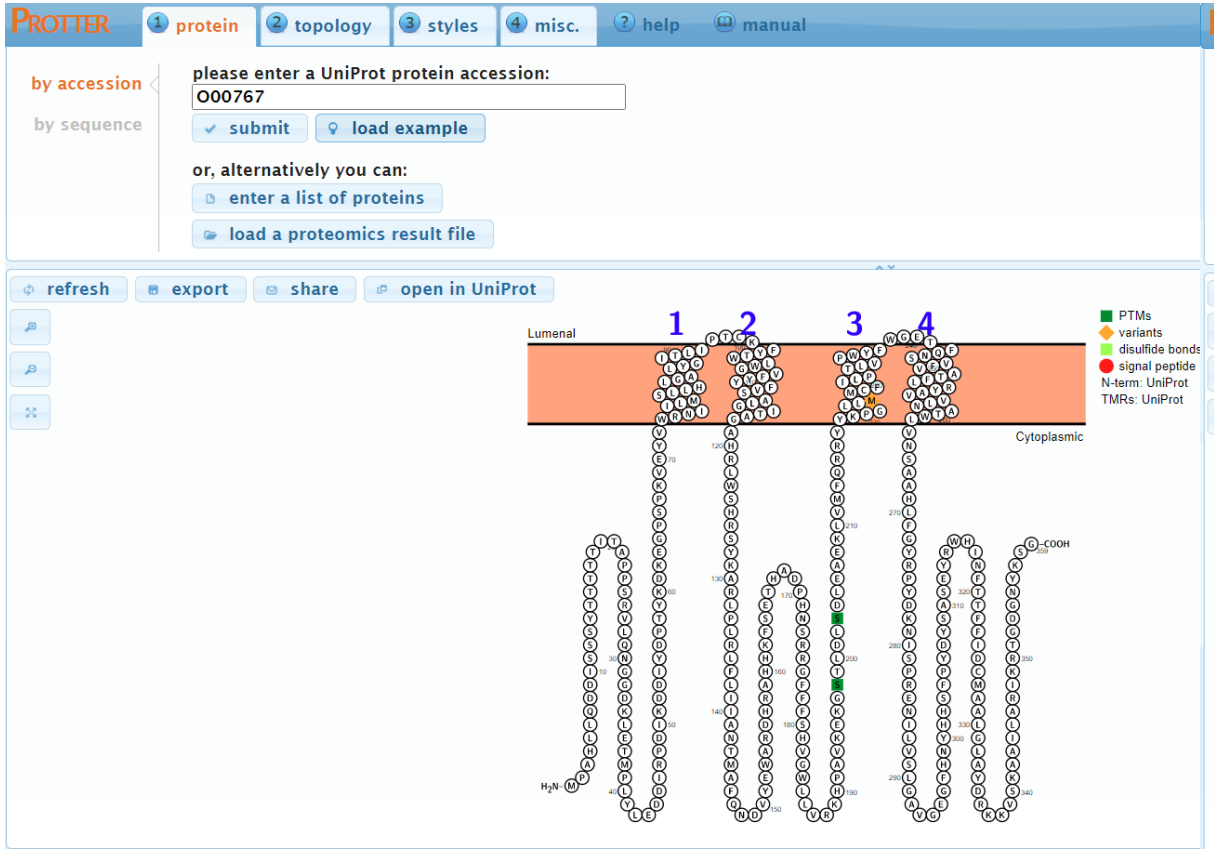
Desaturase (O00767)

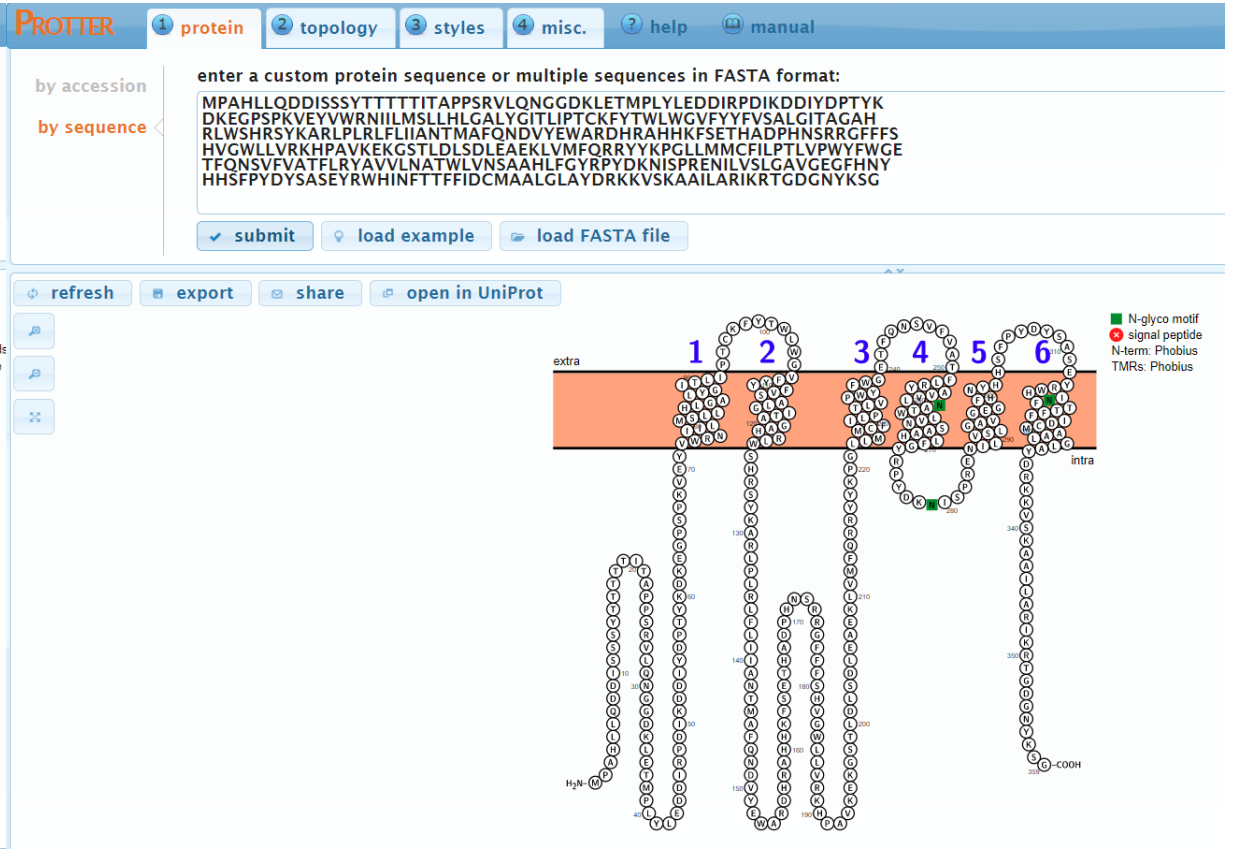
(prepro) insulin (P01308)





PROTTER-figure!



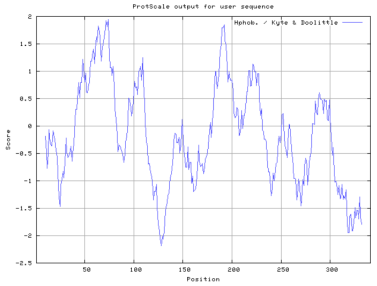


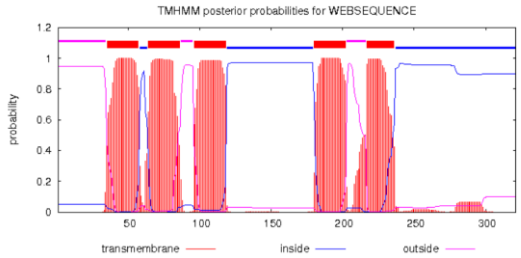
Prediction of transmembrane helices

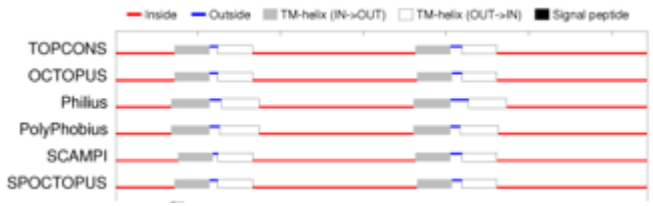
ProtScale

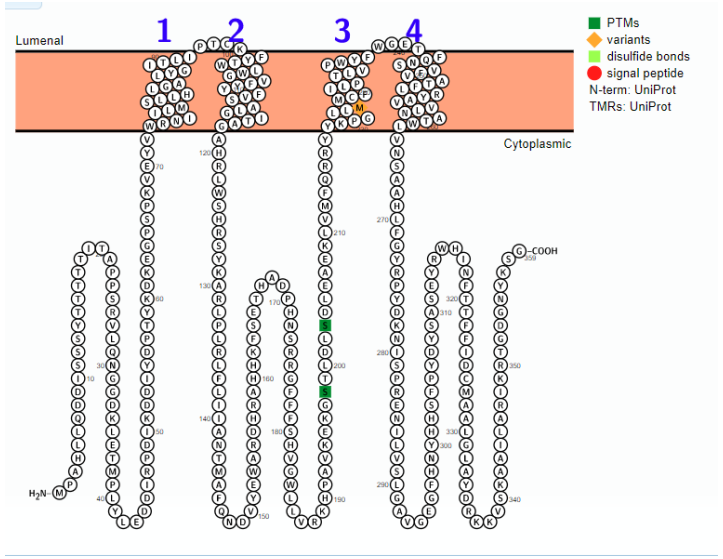
TMHMM

TOPCONS









- **Always try more programs!**

Try more programs.

Does your sequence have any TMHs? and/or signal peptide?

„Protein bioinformatics I“

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 and the domain organization of proteins

Finding all proteins that share a similar sequence and Classifying proteins into families

Finding evolutionary relationships between proteins, drawing proteins’ family trees

Computing the optimal alignment between two or more protein sequences

…

Searching for similar sequences

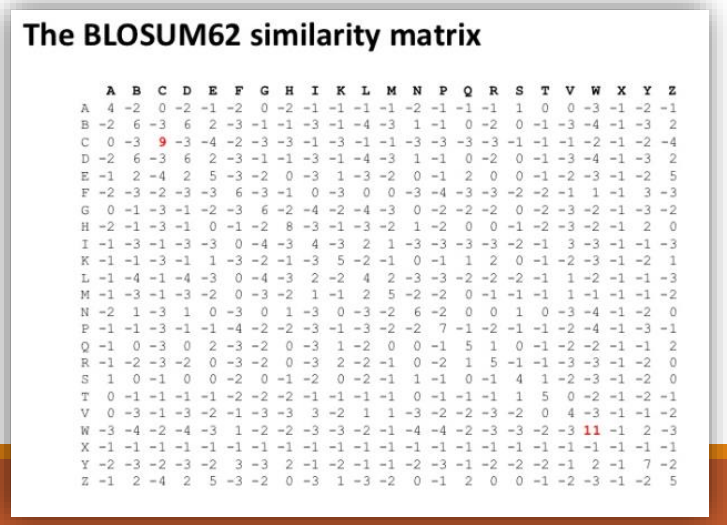
Similarity x Homology

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.

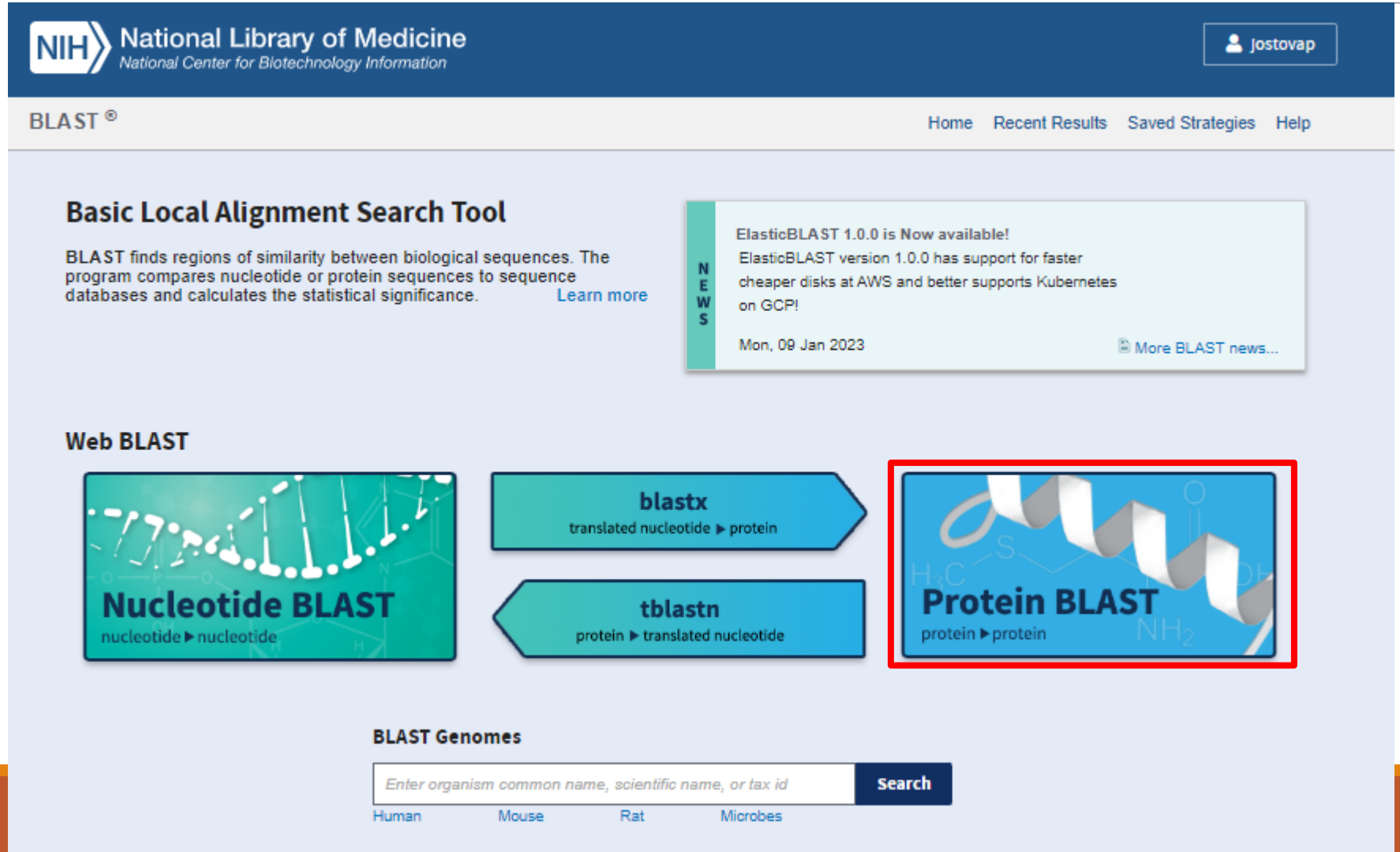
Similarity matrix:

„Leucine is more similar to Isoleucine than Histidine“



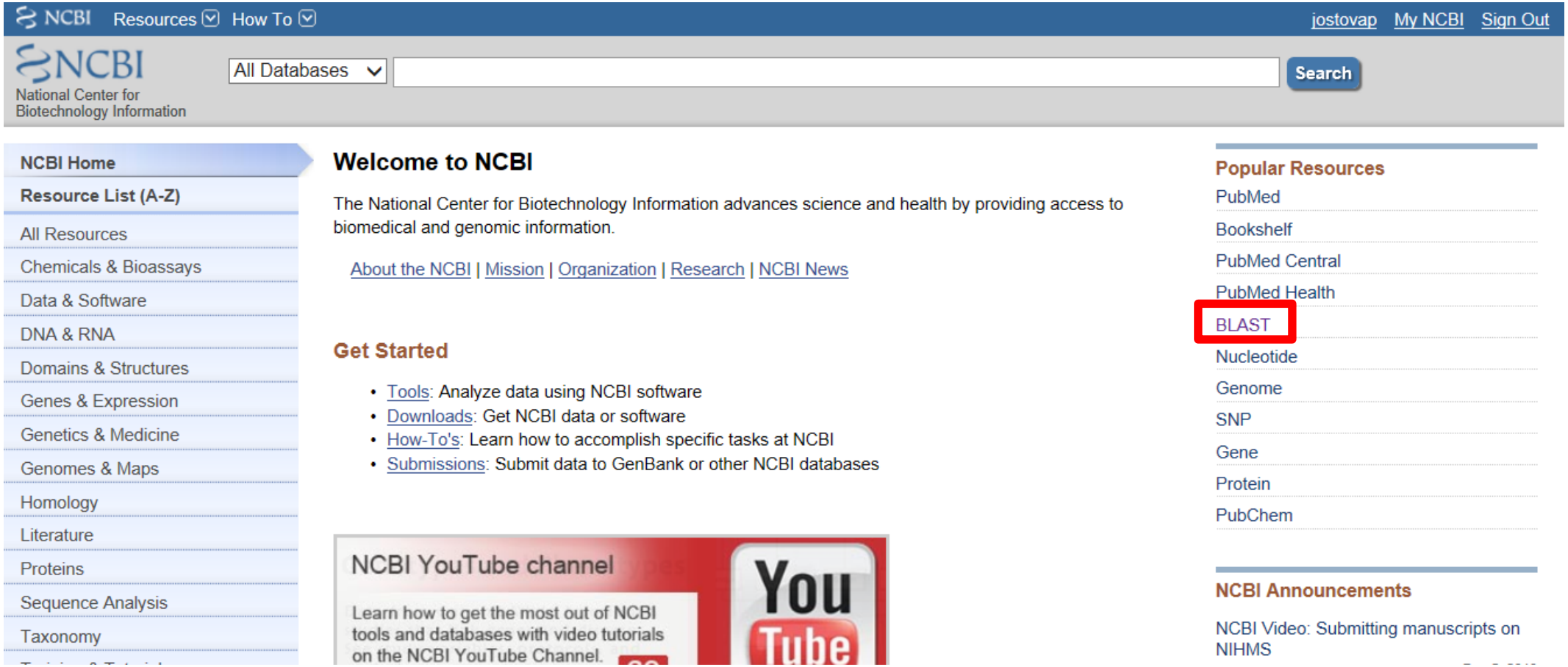
NCBI/BLAST

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



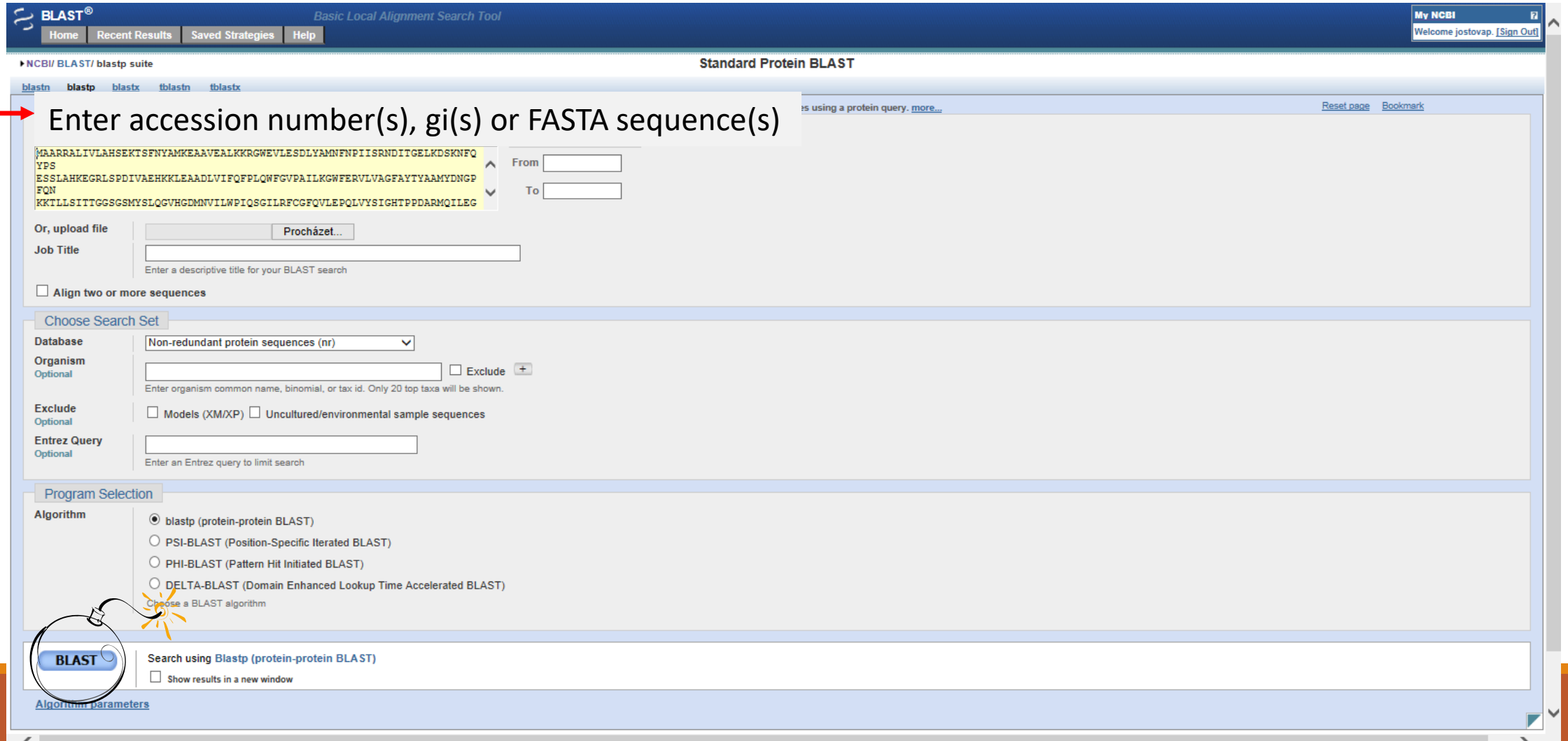
NCBI/BLAST

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

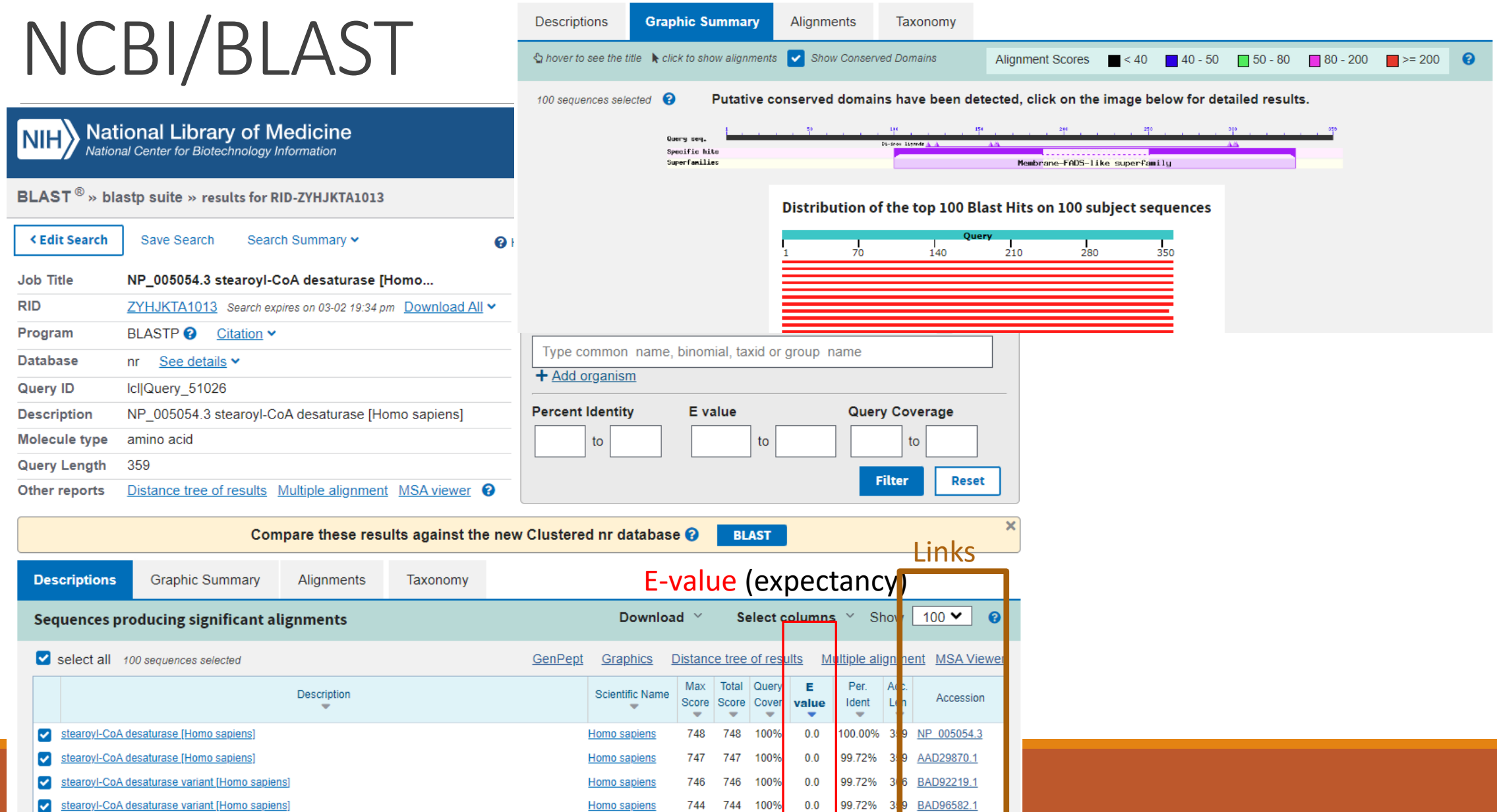


http://www.youtube.com/watchBLAST

NCBI/BLAST



NCBI/BLAST


Descriptions
Graphic Summary
Alignments
Taxonomy
hover to see the title
click to show alignments
Show Conserved Domains
Alignment Scores
< 40
40 - 50
50 - 80
80 - 200
>= 200
100 sequences selected
Putative conserved domains have been detected, click on the image below for detailed results.
Query seq.
Specific hits
Superfamilies
Membrane-FADS-like superfamily
Distribution of the top 100 Blast Hits on 100 subject sequences
Query
1
70
140
210
280
350
National Library of Medicine
National Center for Biotechnology Information
BLAST® » blastp suite » results for RID-ZYHJKTA1013
Edit Search
Save Search
Search Summary
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
Type common name, binomial, taxid or group name
Add organism
Percent Identity
E value
Query Coverage
to
Filter
Reset
Compare these results against the new Clustered nr database
BLAST
Descriptions
Graphic Summary
Alignments
Taxonomy
Links
E-value (expectancy)
Sequences producing significant alignments
Download
Select columns
Show
100
select all
100 sequences selected
GenPept
Graphics
Distance tree of results
Multiple alignment
MSA Viewer
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%
NP_005054.3
stearoyl-CoA desaturase [Homo sapiens]
Homo sapiens
747
747
100%
0.0
99.72%
AAD29870.1
stearoyl-CoA desaturase variant [Homo sapiens]
Homo sapiens
746
746
100%
0.0
99.72%
BAD92219.1
stearoyl-CoA desaturase variant [Homo sapiens]
Homo sapiens
744
744
100%
0.0
99.72%
BAD96582.1

NCBI/BLAST

Descriptions | Graphic Summary | **Alignments** | Taxonomy

Alignment view: Pairwise | Restore defaults | Download

100 sequences selected

Download | GenPept | Graphics | Next | Previous | Descriptions

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

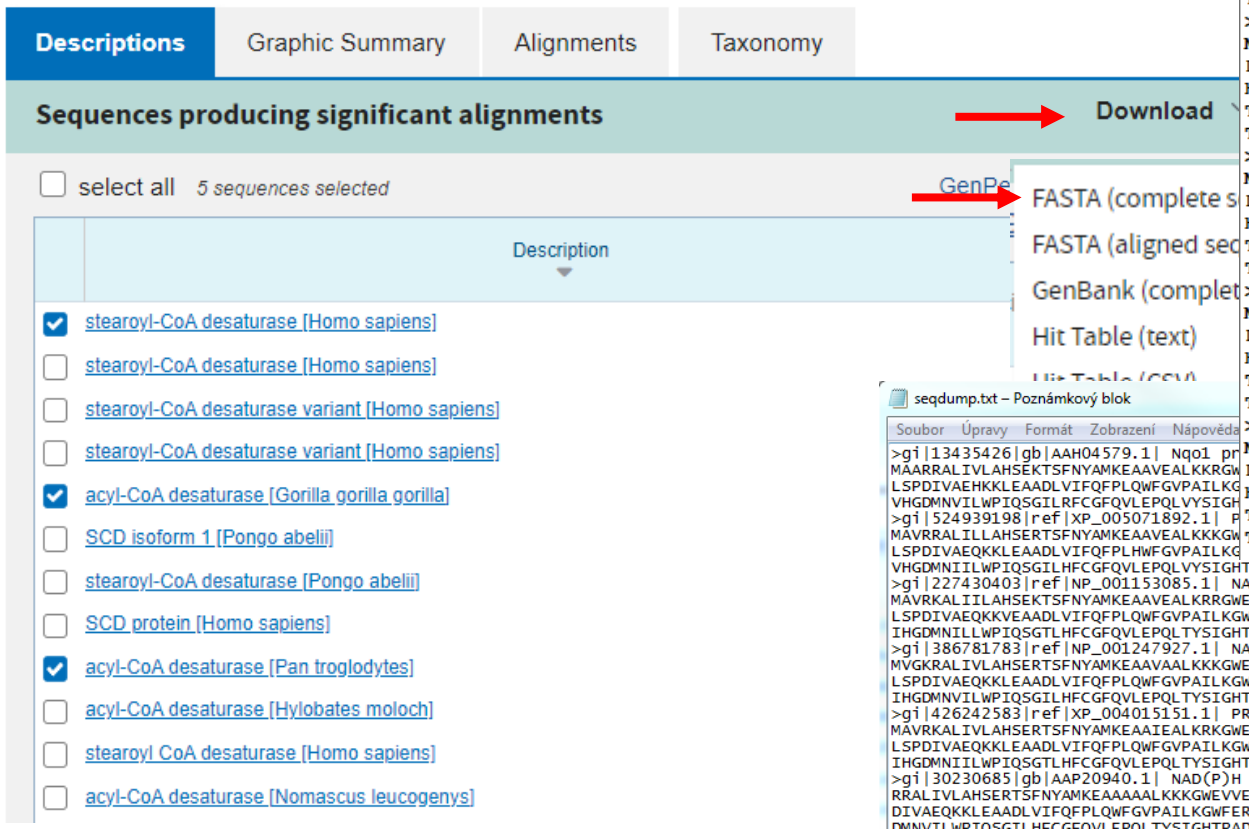
Gene - associated gene details

Genome Data Viewer - aligned genomic context

Identical Proteins - Identical proteins to NP_005054.3

→change sequences (FASTA) names into organism only

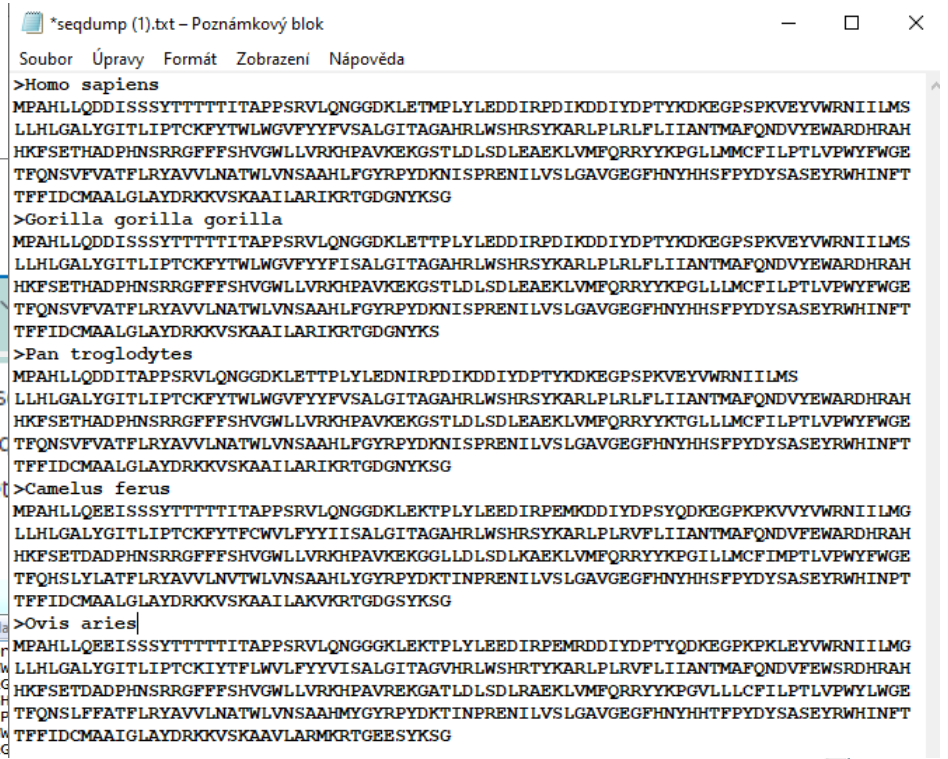
NCBI/BLAST



√

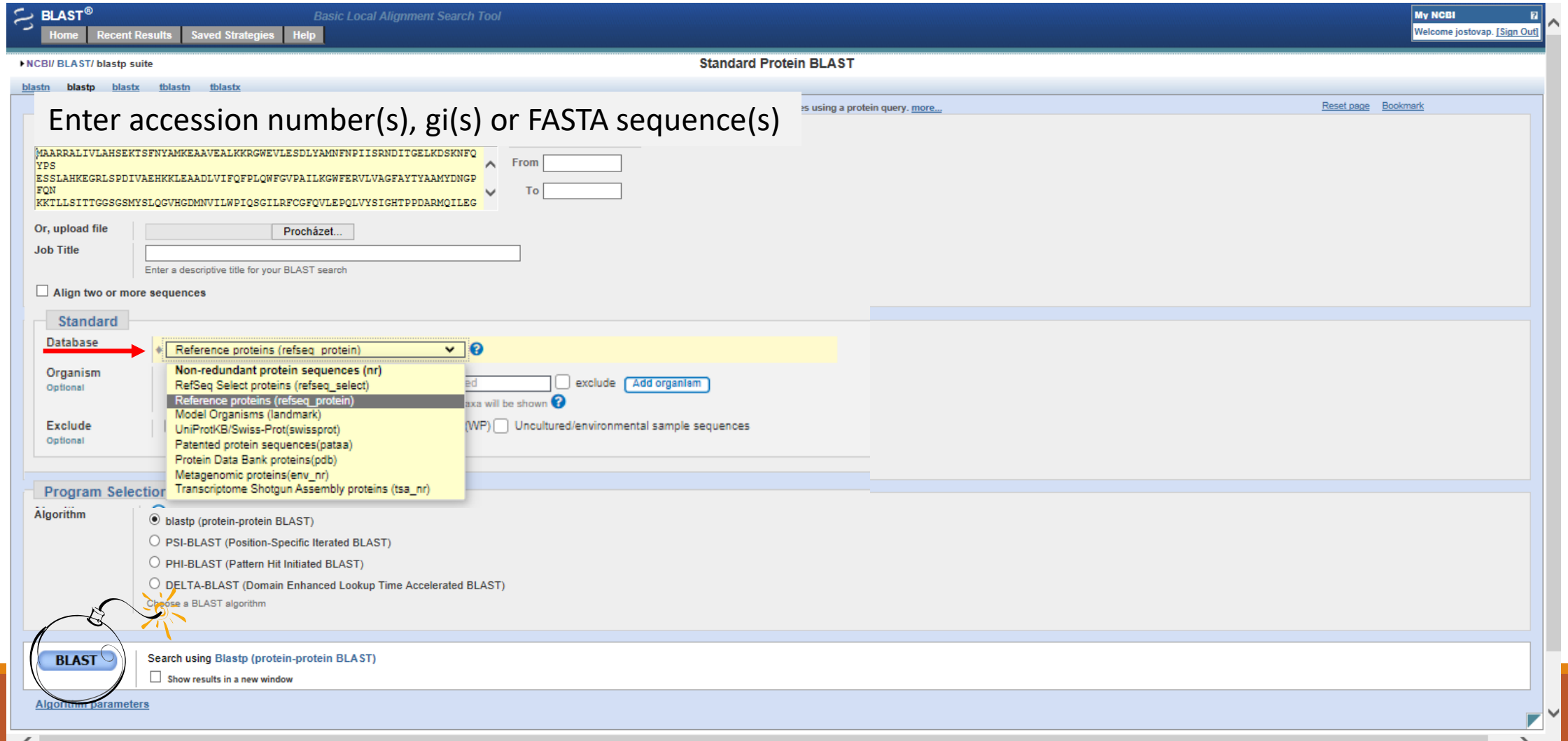
√

√



http://www.youtube.com/watchBLAST

NCBI/BLAST (reference proteins)



Try BLAST.

Download five similar sequences from different organisms.

Homework 3

1) How many times will be the whole sequence cut by trypsin?
2) Does your sequence have a typical domain?
3) Does your sequence have transmembrane helix?
4) Does your sequence have ER retention signal?
5) Find and download five similar sequnces.

E.g use „výstřižky“  „snipping tool“

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

Homework 3 - example

1) Trypsin 33

DÚ3

2) 

3) TMMOD

# Annotation	NON-TM PROTEIN	
# Length	274	
# Number of predicted TMHs	0	
# Exp number of AAs in TMHs	0.000000	
# Exp number, first 60 AAs	0.000000	
# Total prob of N-in	0.493689	
outside	1	274

show posterior probabilities

5)

>Macaca mulatta
MVGKRALIVLAHSERTSFNYAMKEAAVEALKKKGWEVAESDLYAMNFNPIISRKDITGKLKDPANFQYAAESTLAYKEGR
LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFVGEFAYTLAAMYDKGPFQSKKAVLSITTGGSGSMYSLQG
IHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK
KEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK

>Sus scrofa
MAVRKALIILAHSEKTSFNYAMKEAAVEALKRKGWEVAVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETALAYKEGR
LSPDIVAEQKKVEAADLVIFQFPLQWFGVPAILKGWFERVLIGEFAYTYAAMYDKGPFRNKKAVLSITTGGSGSMYSLQG
IHGDMNILLWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARIQILEEWKKRLENIWDETPLYFAPSSLFDLNFQAGFLMK
KQVQDEQKSNKFGLSVGHHLGKSIPTDNQVKARK

>Bos taurus
MAVRKALIVLAHSERTSFNYAMKEAAIEALKRKGWEVTVSDLYAMNFNPVISRKDITGKLKDPGNFQYPAETVLAYKEGR
LSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVLIGEFAYKYAAMYDKGPFRNKGAVLSITTGGSGSMYSLHG
IHGDMNIILWPIQSGTLHFCGFQVLEPQLTYSIGHTPEDARVQILEGWKKRLENIWDEMPLYFAPSSLFDLNFQAGFLMK
KEVQDEQKSKKFGLSVGHHLGKSIPMDNQIKAIK

>Mus musculus
MAARRALIVLAHSEKTSFNYAMKEAAVEALKKRGWEVLESDLYAMNFNPIISRNDITGELKDSHNFQYPSESSLAYKEGR
LSPDIVAEHKKLEAADLVIFQFPLQWFGVPAILKGWFERVLVAGFAYTYAAMYDKGPFQNKKTLLSITTGGSGSMYSLQG
VHGDMNVILWPIQSGILRFCGFQVLEPQLVYSIGHTPPDARMQILEGWKKRLETVWEETPLYFAPSSLFDLNFQAGFLMK
KEVQEEQKGNKFGLSVGHHLGKSIPADNQIKARK

>Alligator sinensis
MAAKGALIVLAHAEKTSFNHAMKDAAVDALQKKGWSVAVSDLYAMGFNPVQSREDITGKPKDPNNFNYAVEMGLAWKEGR
LSSDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFDRVFTGGFAYSYASMYDQGPFQKKKVVLSFTTGGAGSMYTPNG
INGDVNILLWPIQNGTLYFCGFQVLEPQIAYSIGHTPEDVRSQILMGWKERLGSIWEEKPLSFVPSSSFEMSFPKGGFKLK
AETQEQQKDQKYGLSVGQHLGKAIPFDNQVKAQRK

4)