Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2025/2026

„Protein bioinformatics II“

Retrieving protein sequences from databases

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

Prediction of proteases cutting

Predicting elements of protein secondary structure, domains

Predicting 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

…

Prediction of proteases cutting

protease = enzyme that catalyzes proteolysis (*e.g.* digestion)

Examples:

trypsin - digestive enzyme, present in duodenum)

- cleaves sequence „behind“ K(lysin) or R (arginin)

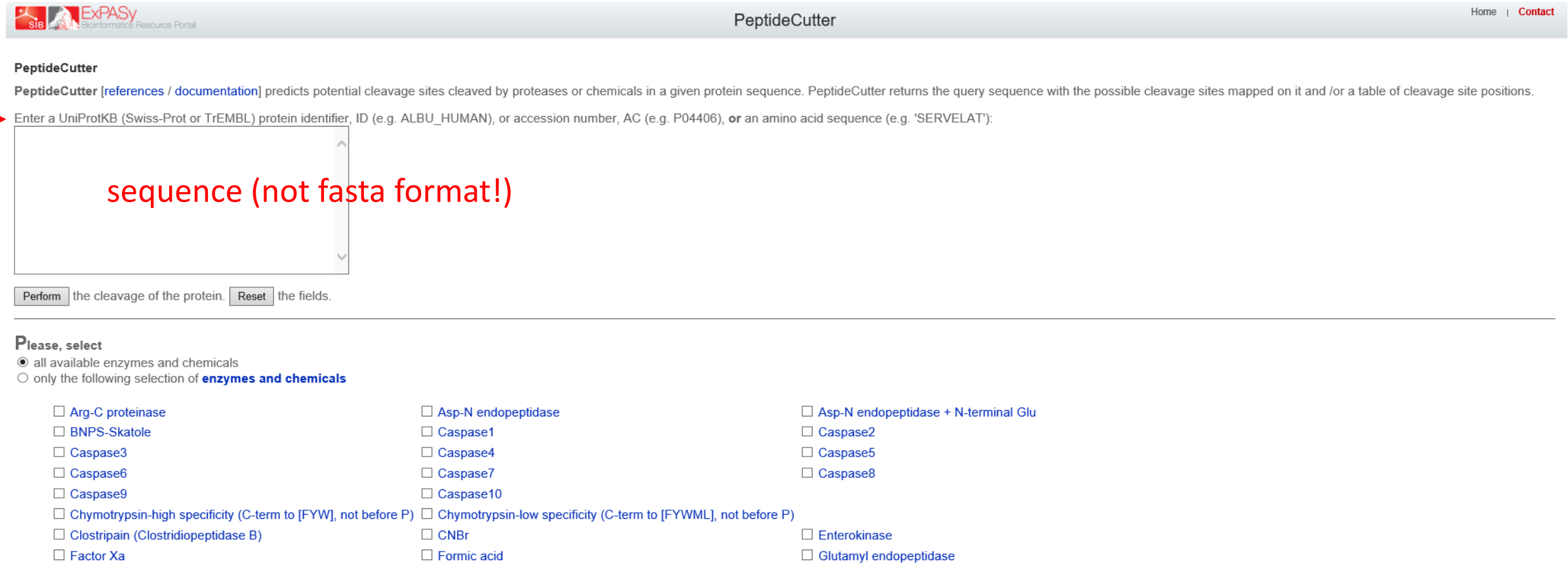
proteinase K - commonly used in molecular biology to digest protein and remove contamination from preparations of nucleic acid.

- cleaves ubiquitously

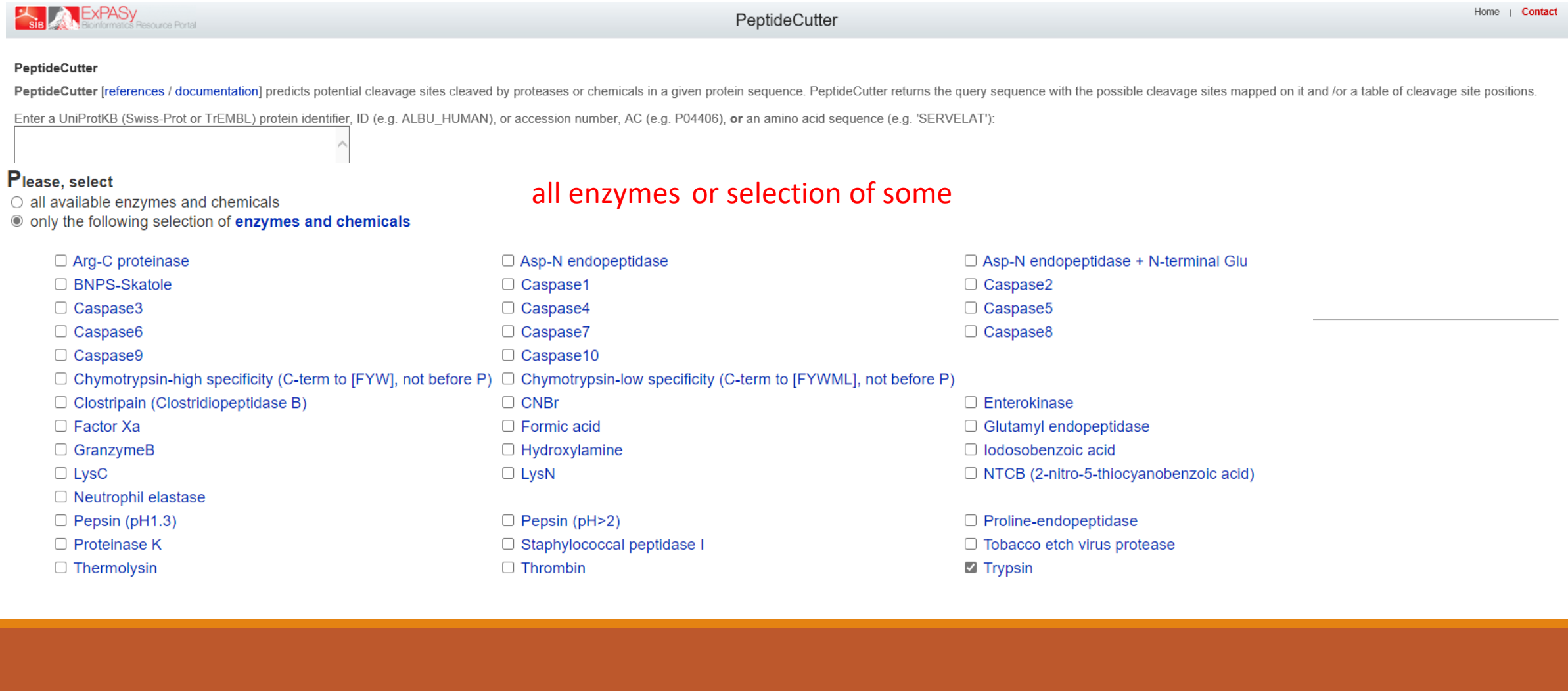
enterokinase - activation of zymogens (precursors of digestive enzymes like trysinogen)

- specific cleavage site (Asp-Asp-Asp-Asp-Lys)

Prediction of proteases cutting



Prediction of proteases cutting



Prediction of proteases cutting

Error

Fasta format provided (only raw format processed).

sequence (not fasta format!)

Prediction of proteases cutting

Name of enzyme	No. of cleavages	Positions of cleavage sites
Arg-C proteinase	9	4 5 15 53 119 139 201 211 273
Asp-N endopeptidase	12	40 54 61 83 95 133 163 198 216 229 244 266
Asp-N endopeptidase + N-terminal Glu	29	13 23 35 38 40 54 61 70 77 83 87 92 95 117 123 133 163 185 198 205 212 216 217 229 241 244 245 246 266
BNPS-Skatole	6	35 106 116 170 208 216
CNBr	7	1 22 45 132 155 165 239
Chymotrypsin-high specificity (C-term to [FYW], not before P)	30	18 20 35 43 47 66 76 100 106 107 116 117 121 125 127 129 133 138 156 179 182 191 208 216 222 223 229 233 237 252
Chymotrypsin-low specificity (C-term to [FYWML], not before P)	67	1 7 10 12 18 20 22 30 35 42 43 45 47 60 66 74 76 80 81 92 97 100 104 106 107 113 116 117 121 125 127 129 133 138 145 156 158 162 165 169 177 178 179 182 185 189 191 195 205 208 212 216 221 222 223 228 229 231 233 237 238 239 252 254 258 259 260
Clostripain	9	4 5 15 53 119 139 201 211 273
Enterokinase	1	248
Formic acid	12	41 55 62 84 96 134 164 199 217 230 245 267
Glutamyl endopeptidase	17	14 24 36 39 71 78 88 93 118 124 186 206 213 218 242 246 247
Iodosobenzoic acid	6	35 106 116 170 208 216
LysC	24	23 31 32 33 54 59 61 77 90 91 114 135 141 142 209 210 240 241 248 250 251 262 271 274
LysN	24	22 30 31 32 53 58 60 76 89 90 113 134 140 141 208 209 239 240 247 249 250 261 270 273
NTCB (2-nitro-5-thiocyanobenzoic acid)	1	179
Pepsin (pH1.3)	59	9 10 18 29 30 41 42 46 59 60 65 66 73 74 80 91 96 97 99 100 102 103 106 107 112 113 117 120 124 125 145 157 158 168 176 177 178 179 181 182 184 189 204 205 220 222 227 228 229 230 231 232 233 236 237 238 251 254 259
Pepsin (pH>2)	82	9 10 18 19 20 29 30 41 42 43 46 59 60 65 66 68 73 74 75 76 80 91 96 97 99 100 102 103 105 106 107 112 113 115 117 120 124 125 126 127 128 129 132 133 145 155 156 157 158 168 170 176 177 178 179 181 182 184 189 190 191 204 205 207 208 215 216 220 222 227 228 229 230 231 232 233 236 237 238 251 254 259
Proteinase K	142	2 6 7 8 9 10 11 14 16 18 20 21 24 25 26 27 28 29 30 35 36 37 38 39 42 43 44 47 50 51 56 57 60 64 66 68 70 71 73 74 75 76 78 81 85 86 87 88 92 93 94 95 97 98 99 100 102 104 106 107 109 111 112 113 116 117 118 120 121 122 124 125 126 127 128 129 130 131 133 138 143 144 145 147 148 149 156 158 161 167 168 169 170 172 176 177 179 182 184 185 186 189 190 191 193 196 198 200 202 204 205 206 208 212 213 215 216 218 219 221 222 223 224 228 229 231 233 235 237 238 242 243 246 247 252 254 256 260 264 266 270 272
Staphylococcal peptidase I	16	14 24 36 39 71 78 88 93 118 124 186 206 213 218 242 246
Thermolysin	90	1 5 6 7 8 9 10 17 20 21 25 26 27 28 29 37 43 44 46 49 50 59 63 65 69 72 73 74 80 85 86 91 94 97 98 99 103 106 110 111 112 116 119 120 121 125 129 130 131 137 142 143 144 146 154 157 160 166 167 168 171 175 176 178 181 183 184 188 192 197 201 203 204 211 214 220 222 227 228 232 234 236 237 238 251 253 255 259 269 271
Trypsin	33	4 5 15 23 31 32 33 53 54 59 61 77 90 91 114 119 135 139 141 142 201 209 210 211 240 241 248 250 251 262 271 273 274

all enzymes

These chosen enzymes do not cut:

Caspase1

Caspase10

Prediction of proteases cutting

The enzyme(s) that you have chosen:

- Trypsin

You have chosen to display all possible cleaving enzymes.

These enzymes cleave the sequence:

Name of enzyme	No. of cleavages	Positions of cleavage sites
Trypsin	33	4 5 15 23 31 32 33 53 54 59 61 77 90 91 114 119 135 139 141 142 201 209 210 211 240 241 248 250 251 262 271 273 274

These are the cleavage sites of the chosen enzymes and chemicals mapped onto the entered protein sequence:

- You have chosen a block size of **60** for the map.
- Please note that the cleavage occurs at the **right side** (C-terminal direction) of the marked amino acid.
- You have the possibility to display the results of a single enzyme by **mouseclicking** on the respective enzyme name in the map.

or selection of some

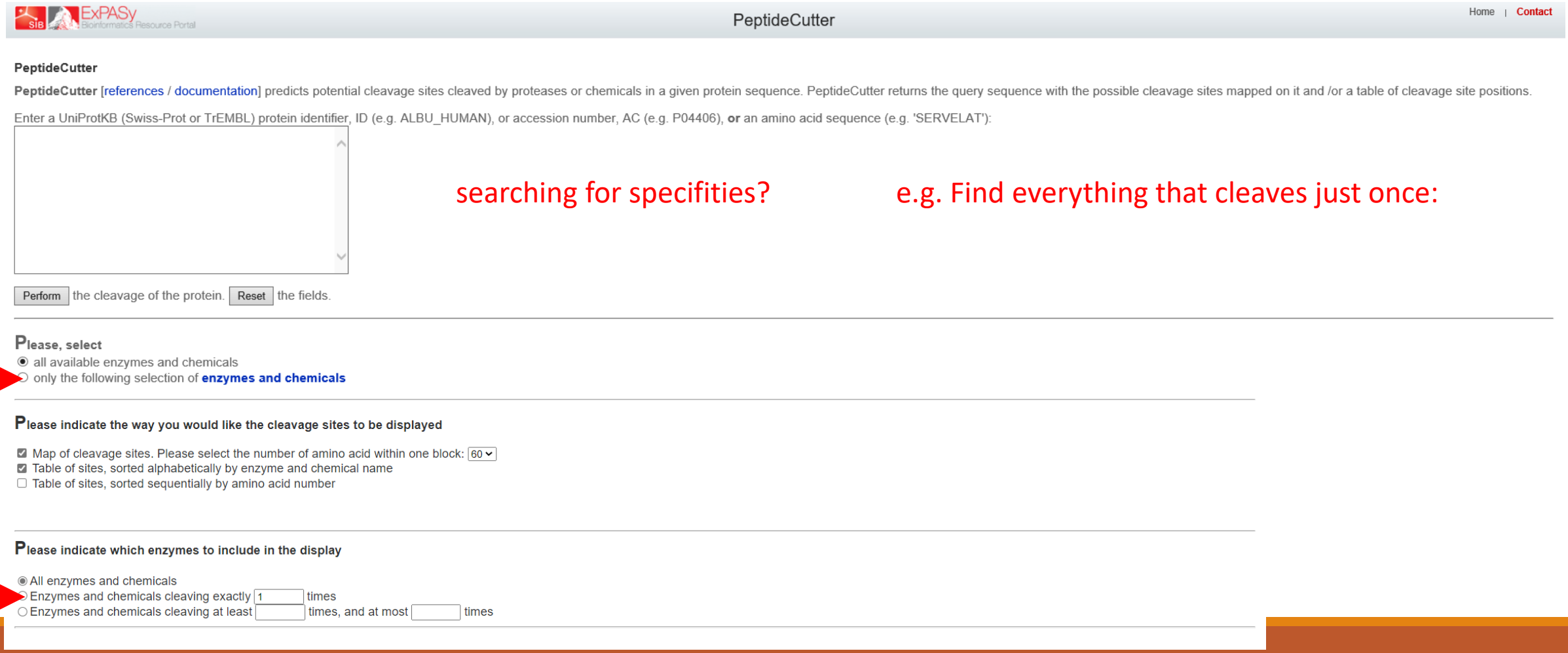
```
                                Tryps                 Tryps
    Tryps                       Tryps|              Tryps    |
   Tryps|     Tryps   Tryps    Tryps||             Tryps|    |

      ||        |      |       |||                   ||    |
   MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKL
1  ---------+---------+---------+---------+---------+---------+   60
```

```
                               Tryps                 Tryps
Tryps           Tryps         Tryps|               Tryps    |

    |             |            ||                     |    |
   KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV
61 ---------+---------+---------+---------+---------+---------+   120
```

Prediction of proteases cutting



searching for specifities?

e.g. Find everything that cleaves just once:

Prediction of proteases cutting

[*] *NOTE: Proline-endopeptidase was reported to cleave only substrates whose sequences do not exceed 30 amino acids. An unusual beta-propeller domain regulates proteolysis: see Fulop et al., 1998.*
You have chosen to display only those enzymes that cleave exactly 1 times. However, the following enzymes also cleave but not with the selected frequency:
Staphylococcal peptidase I , Pepsin (pH1.3) , Glutamyl endopeptidase , CNBr , Pepsin (pH>2) , Asp-N endopeptidase , Asp-N endopeptidase + N-terminal Glu , Formic acid , Iodosobenzoic acid , Arg-C proteinase , Thermolysin , Trypsin , Clostripain , Proteinase K , Chymotrypsin-high specificity (C-term to [FYW], not before P) , Chymotrypsin-low specificity (C-term to [FYWML], not before P) , LysC , BNPS-Skatole , LysN ,

These enzymes cleave the sequence:

Name of enzyme	No. of cleavages	Positions of cleavage sites
Enterokinase	1	248
NTCB (2-nitro-5-thiocyanobenzoic acid)	1	179

At these positions the following enzymes cleave:

- Please note that the size of the peptides are calculated as if **all chosen enzymes were present** during digestion. If you want to obtain the size of the peptides resulting from the cleavage of only one enzyme, please, deselect the others.
- Please be aware of the fact that the present version of the PeptideCutter program does not take into consideration any kind of **modification** neither of the protein sequence nor of modifications evoked by the cleavage. Mass computations are based on average masses of the occurring amino acid residues, and giving peptide masses as [M]. If you want to select different parameters, we recommend to use PeptideMass.

Position of cleavage site	Name of cleaving enzyme(s)	Resulting peptide sequence (see explanations)	Peptide length [aa]	Peptide mass [Da]
179	NTCB (2-nitro-5-thiocyanobenzoic acid)	MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHF	179	19997.201
248	Enterokinase	CGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFL MKKEVQDEEK	69	8032.136
274	**end of sequence**	NKKFGLSVGHHLGKSIPTDNQIKARK	26	2874.342

These are the cleavage sites of the chosen enzymes and chemicals mapped onto the entered protein sequence:

- You have chosen a block size of **60** for the map.
- Please note that the cleavage occurs at the **right side** (C-terminal direction) of the marked amino acid.
- You have the possibility to display the results of a single enzyme by **mouseclicking** on the respective enzyme name in the map.

```
    MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKL
1   ---------+---------+---------+---------+---------+---------+   60

    KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV
61  ---------+---------+---------+---------+---------+---------+   120
```

Prediction of proteases cutting

Prediction of proteases cutting

Name of enzyme	No. of cleavages	Positions of cleavage sites
Trypsin	33	4 5 15 23 31 32 33 53 54 59 61 77 90 91 114 119 135 139 141 142 201 209 210 211 240 241 248 250 251 262 271 273 274

At these positions the following enzymes cleave:

- Please note that the size of the peptides are calculated as if **all chosen enzymes were present** during digestion. If you want to obtain the size of the peptides resulting from the cleavage of only one enzyme, please, deselect the others.
- Please be aware of the fact that the present version of the PeptideCutter program does not take into consideration any kind of **modification** neither of the protein sequence nor of modifications evoked by the cleavage. Mass computations are based on average ma of the occurring amino acid residues, and giving peptide masses as [M]. If you want to select different parameters, we recommend to use PeptideMass.

Position of cleavage site	Name of cleaving enzyme(s)	Resulting peptide sequence (see explanations)	Peptide length [aa]	Peptide mass [Da]
4	Trypsin	MVGR	4	461.580
5	Trypsin	R	1	174.203
15	Trypsin	ALIVLAHSER	10	1108.306
23	Trypsin	TSFNYAMK	8	961.100
31	Trypsin	EAAAAALK	8	743.858
32	Trypsin	K	1	146.189
33	Trypsin	K	1	146.189
53	Trypsin	GWEVVESDLYAMNFNPIISR	20	2340.636
54	Trypsin	K	1	146.189
59	Trypsin	DITGK	5	532.594
61	Trypsin	LK	2	259.349
77	Trypsin	DPANFQYPAESVLAYK	16	1812.997
90	Trypsin	EGHLSPDIVAEQK	13	1422.558
91	Trypsin	K	1	146.189
114	Trypsin	LEAADLVIFQFPLQWFGVPAILK	23	2616.141
119	Trypsin	GWFER	5	693.760
135	Trypsin	VFIGEFAYTYAAMYDK	16	1889.153
139	Trypsin	GPFR	4	475.548
141	Trypsin	SK	2	233.268
142	Trypsin	K	1	146.189
201	Trypsin	AVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYS	59	6287.190
209	Trypsin	IQILEGWK	8	986.179
210	Trypsin	K	1	146.189
211	Trypsin	R	1	174.203
240	Trypsin	LENIWDETPLYFAPSSLFDLNFQAGFLMK	29	3407.885
241	Trypsin	K	1	146.189
248	Trypsin	EVQDEEK	7	875.888

Try PeptideCutter

Analyze your sequence

How many times is your sequence cut by trypsin (HW3)

How long is the longest product after trypsin digest?

Is there any enzyme that cuts just once?

„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

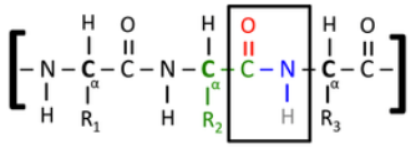
→ **Finding 3-D structure and the domain organization of proteins**

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

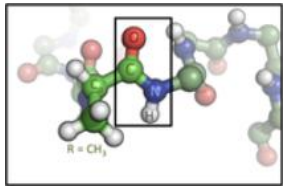
Finding evolutionary relationships between proteins, drawing proteins’ family trees

Computing the optimal alignment between two or more protein sequences

…



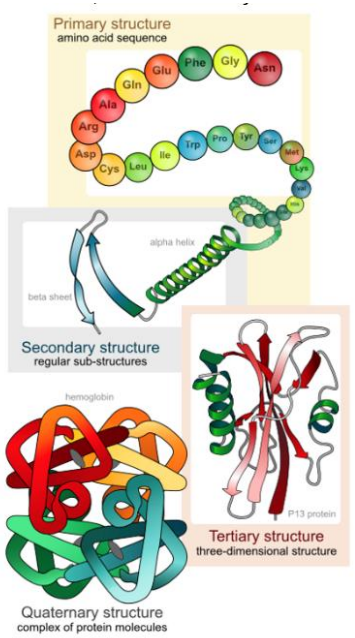
Proteins



20 Aminoacids – primary structure:

(Frederick Sanger-1958 Nobel prize for insulin sequencing)

Secondary structure
Tertiary structure
Quaternary structure

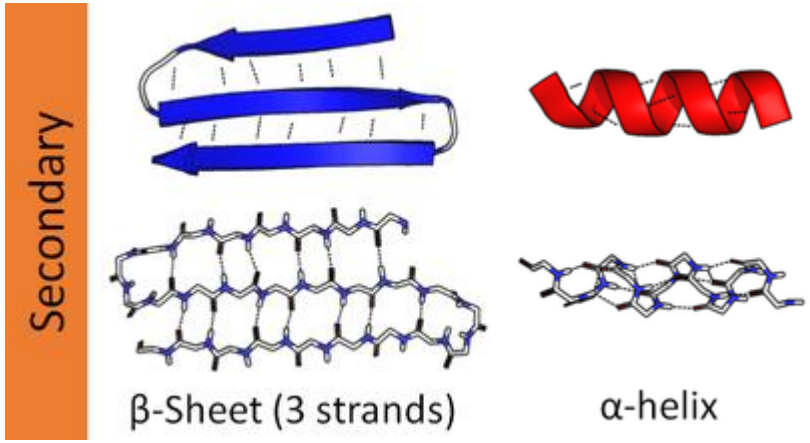


SEQUENCE ⇨ STRUCTURE ⇨ FUNCTION

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

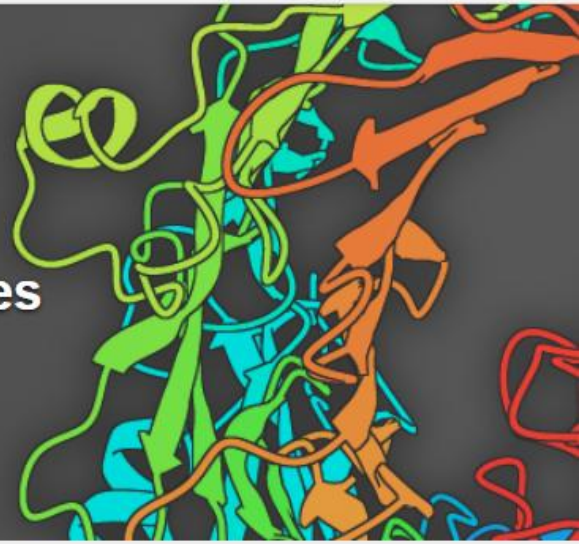
Protein domain

- 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



CATH / Gene3D v4.4

151 million protein domains classified into 6,573 superfamilies



Search by keywords, PDB code, GO term, etc Search

Putative CATH annotations for predicted structural domains in AlphaFold DB are available in The Encyclopedia of Domains (TED). Annotations for the 21 model organisms predicted by AlphaFold (v2) are available to download (doi:10.1038/s42003-023-04488-9). Core classification files for the latest version of CATH-Plus (v4.4) are available to download. Daily updates of our very latest classifications are also available.

3D Structure

Find out what 3D structure your protein adopts

Find out more Go

Protein Evolution

Learn about a particular protein family and how it evolved

Find out more

Protein Function

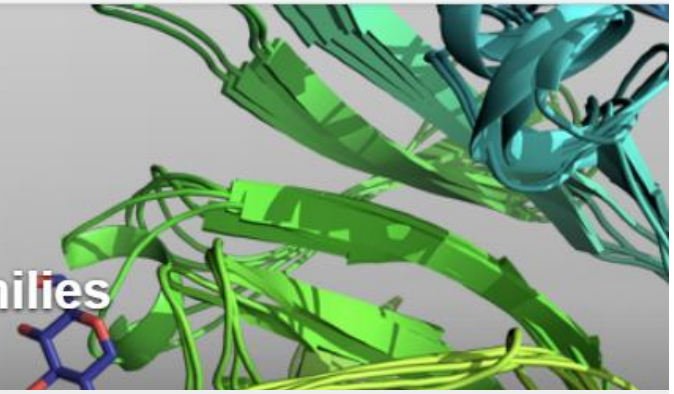
Investigate the function of your protein

Find out more Go

C A T H Home Search Browse Download About Support Search CATH by keywords or ID

CATH / Gene3D v4.3

151 million protein domains classified into 5,841 superfamilies



What is CATH-Gene3D?

CATH is a classification of protein structures downloaded from the Protein Data Bank. We group protein domains into superfamilies when there is sufficient evidence they have diverged from a common ancestor.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence
- Search CATH by PDB structure
- Browse CATH Hierarchy
- CATH Release Statistics
- CATH Tutorials

Gene3D uses the information in CATH to predict the locations of structural domains on millions of protein sequences available in public databases. This allows us to include additional annotations to the CATH-Gene3D database such as functional information and active site residues.

- Go to Gene3D
- Compare Genomes
- Download Gene3D Data
- Learn how Gene3D is created

If you have any questions, comments or suggestions please get in touch via Twitter, ask a question in our online forum or visit our support page.

Latest Release Statistics

Info

	CATH-Plus 4.3.0	CATH (daily snapshot)
PDB Release	01-07-2019	about 15 hours ago
		Gene3D v21
Protein Sequences		82,665,384
CATH Domain Predictions		151,013,797

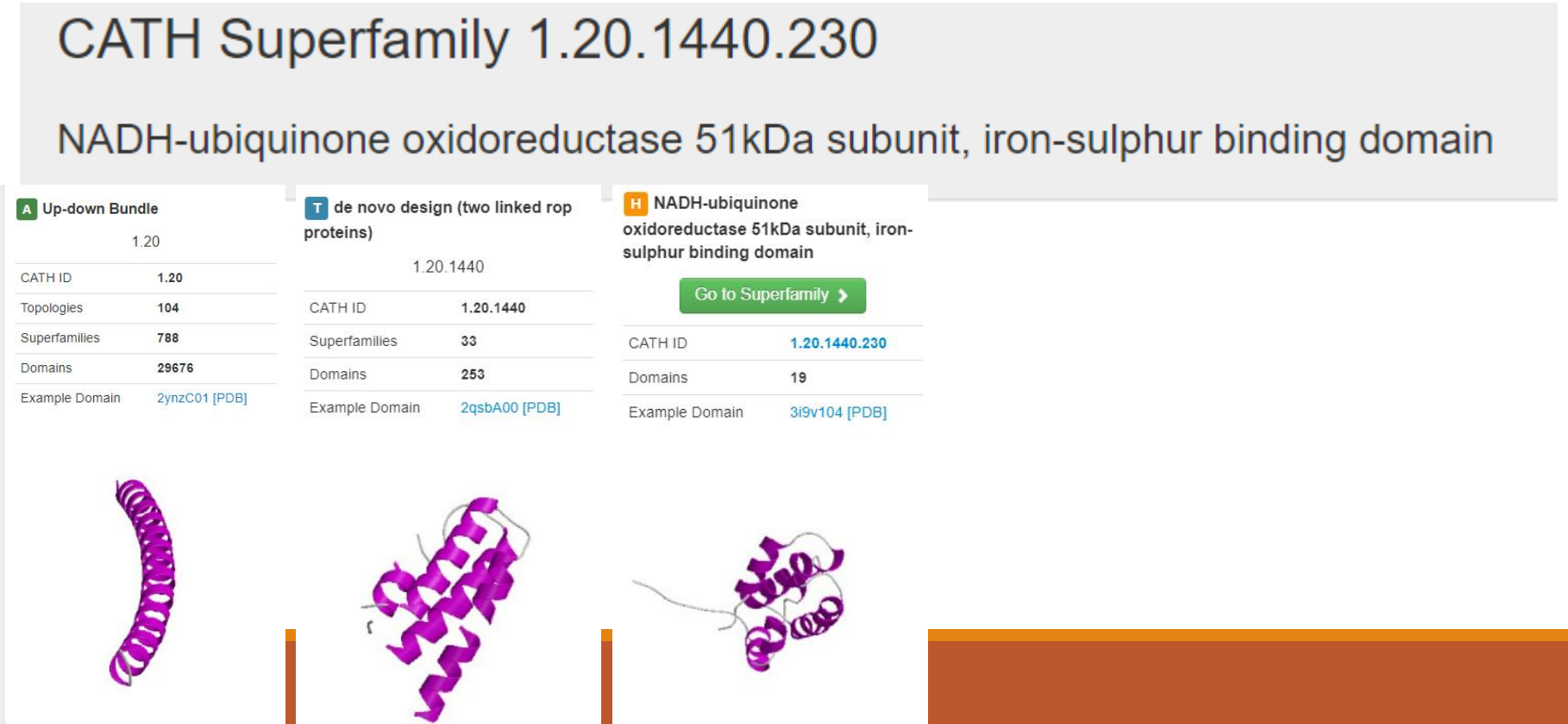
Top of CATH Hierarchy (4 Classes)

C	1	Mainly Alpha	5 Architectures, 404 Folds, 2033 Superfamilies, 103788 Domains
C	2	Mainly Beta	21 Architectures, 244 Folds, 1290 Superfamilies, 124032 Domains
C	3	Alpha Beta	14 Architectures, 634 Folds, 2337 Superfamilies, 262275 Domains
C	4	Few Secondary Structures	1 Architectures, 108 Folds, 181 Superfamilies, 5716 Domains
C	6	Special	2 Architectures, 82 Folds, 790 Superfamilies, 4427 Domains

CATH – try it

- look for a superfamily of your protein

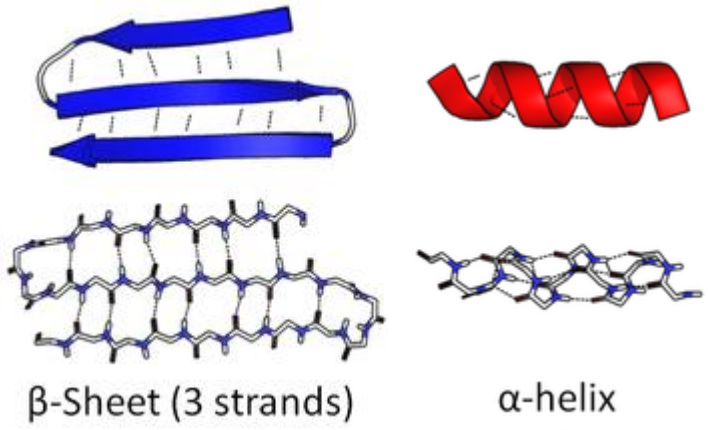
NQO1:



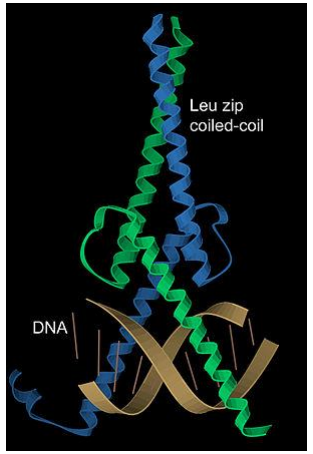
Protein domain

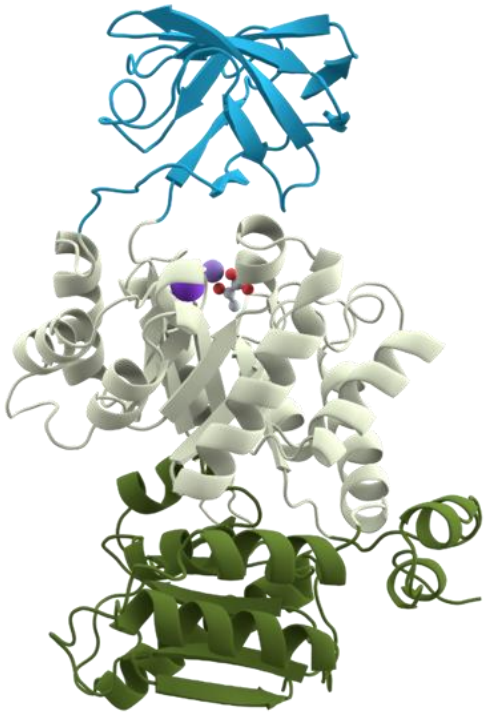
- region of a protein's polypeptide chain that folds independently from the rest
- forms a compact folded three-dimensional structure
- many proteins consist of several domains

Secondary



Tertiary



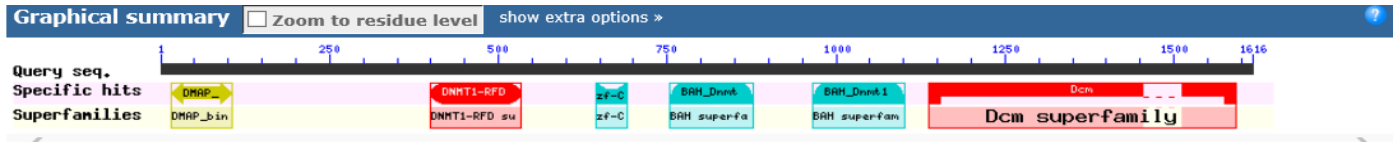


Conserved domain search

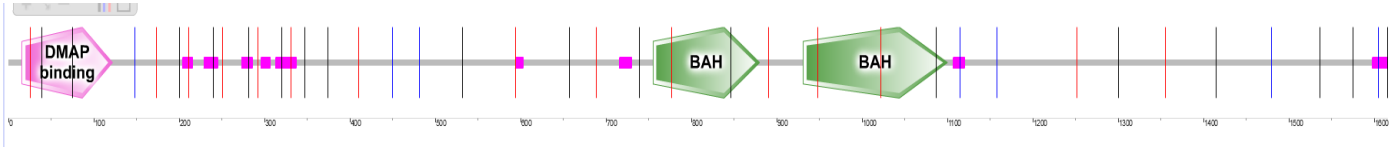
SEQUENCE➪STRUCTURE➪FUNCTION

- Conserved domain databases:

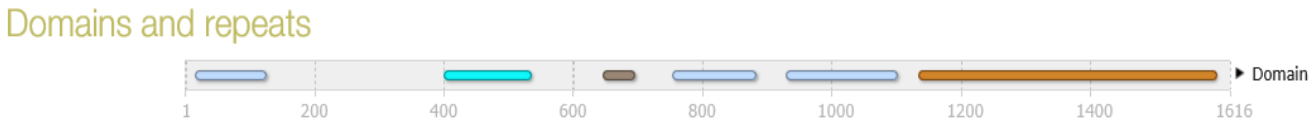
NCBI/CDD



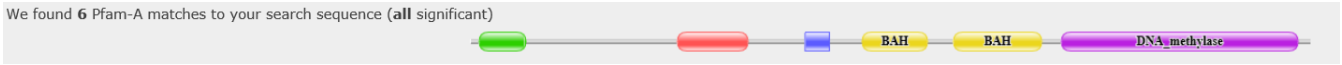
SMART



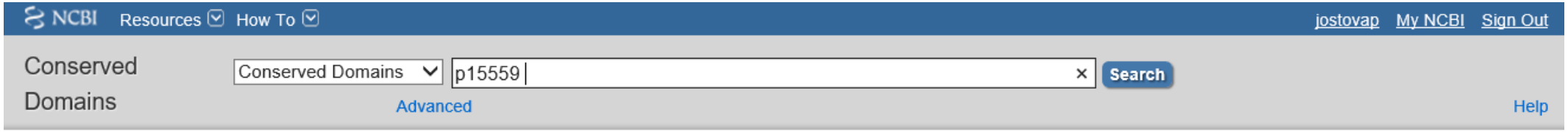
EMBL/InterPro

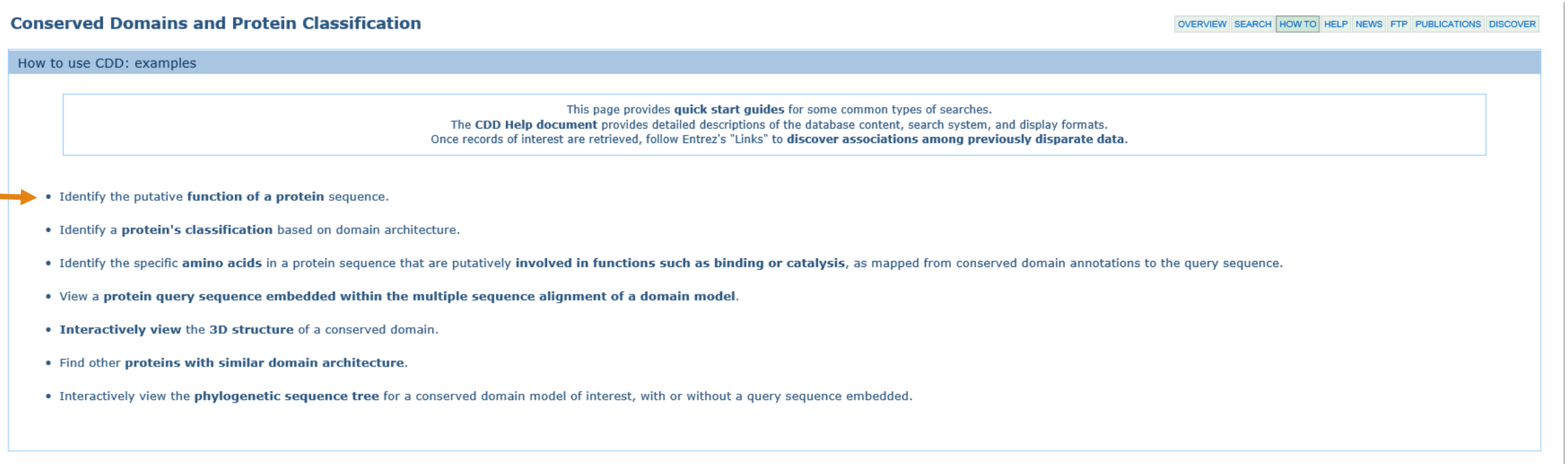


Pfam

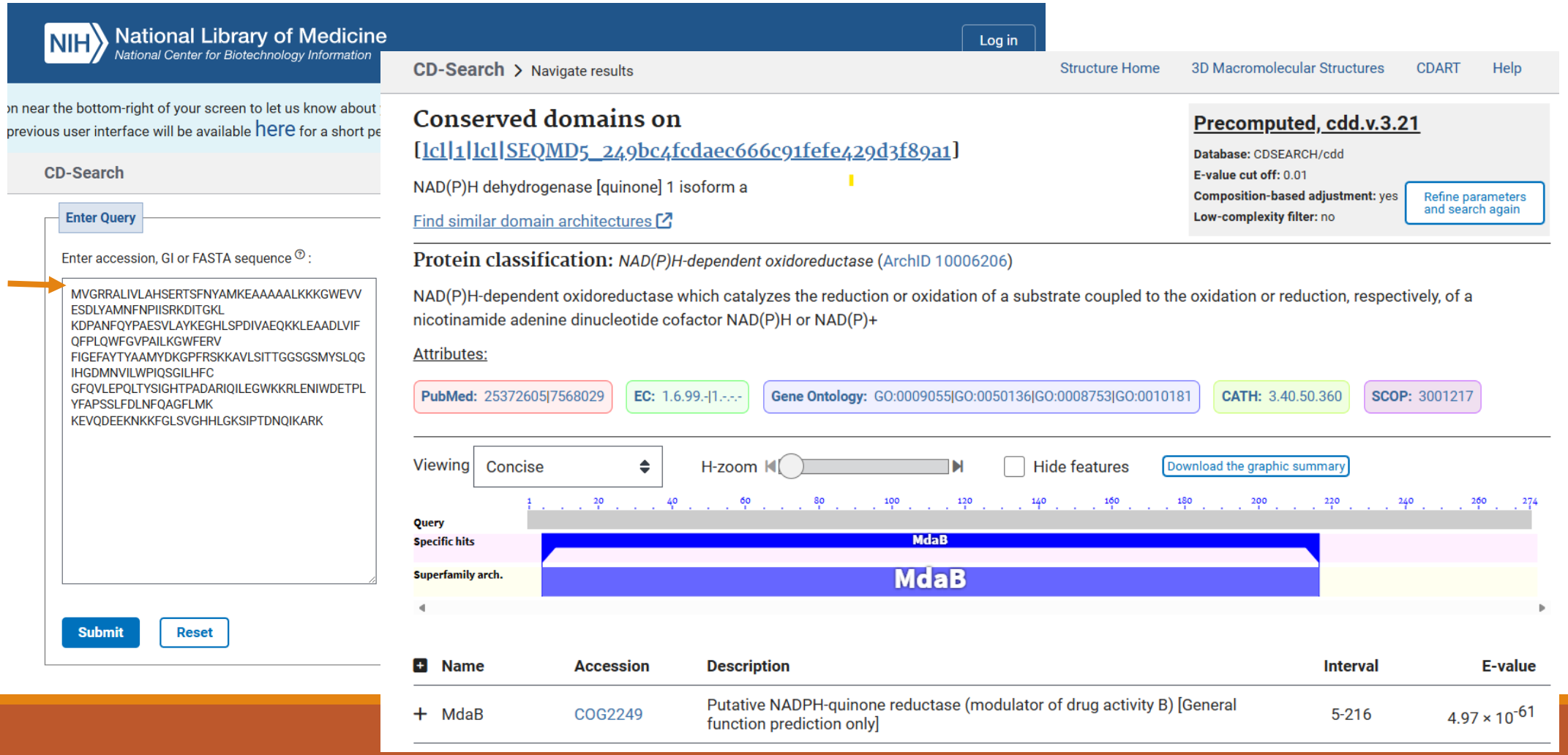


Conserved domain search - CD (NCBI)

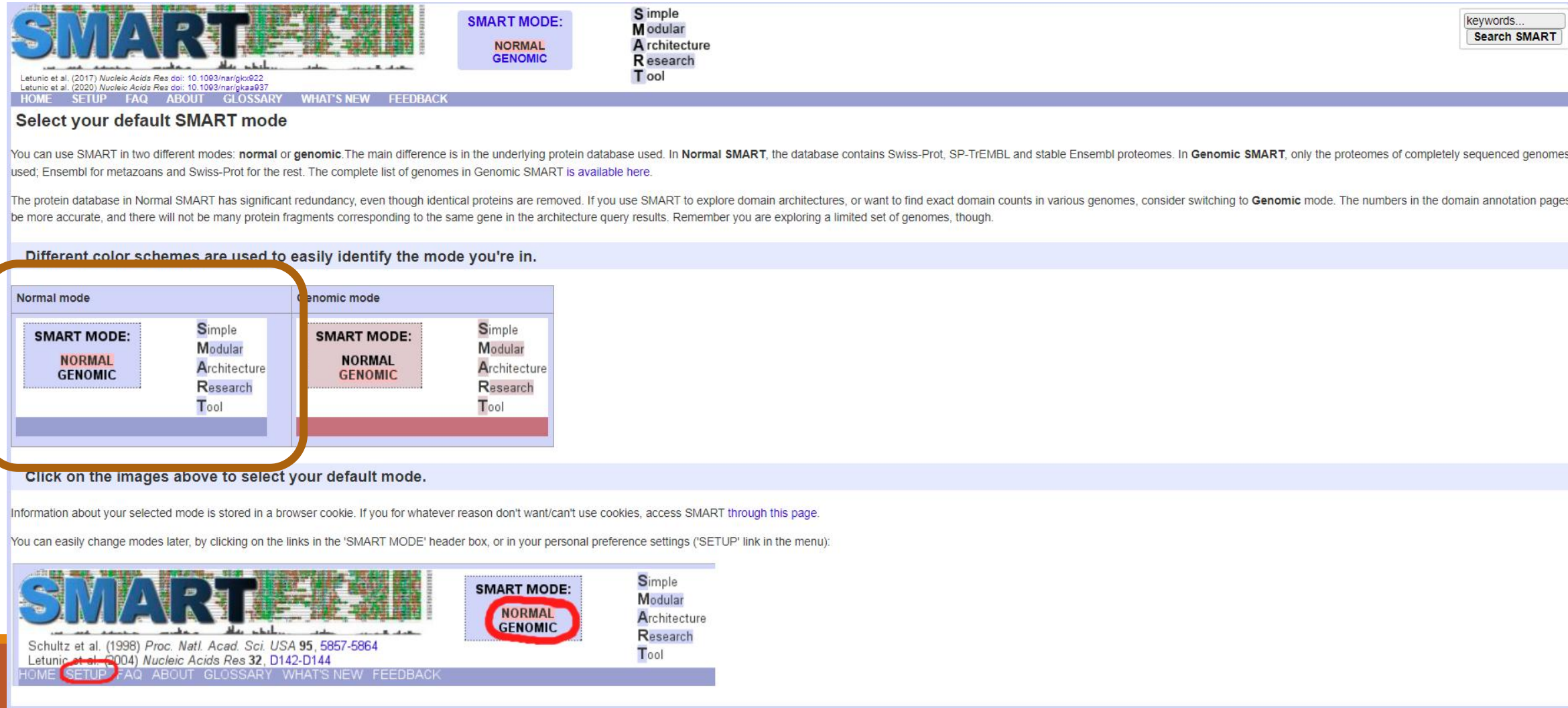




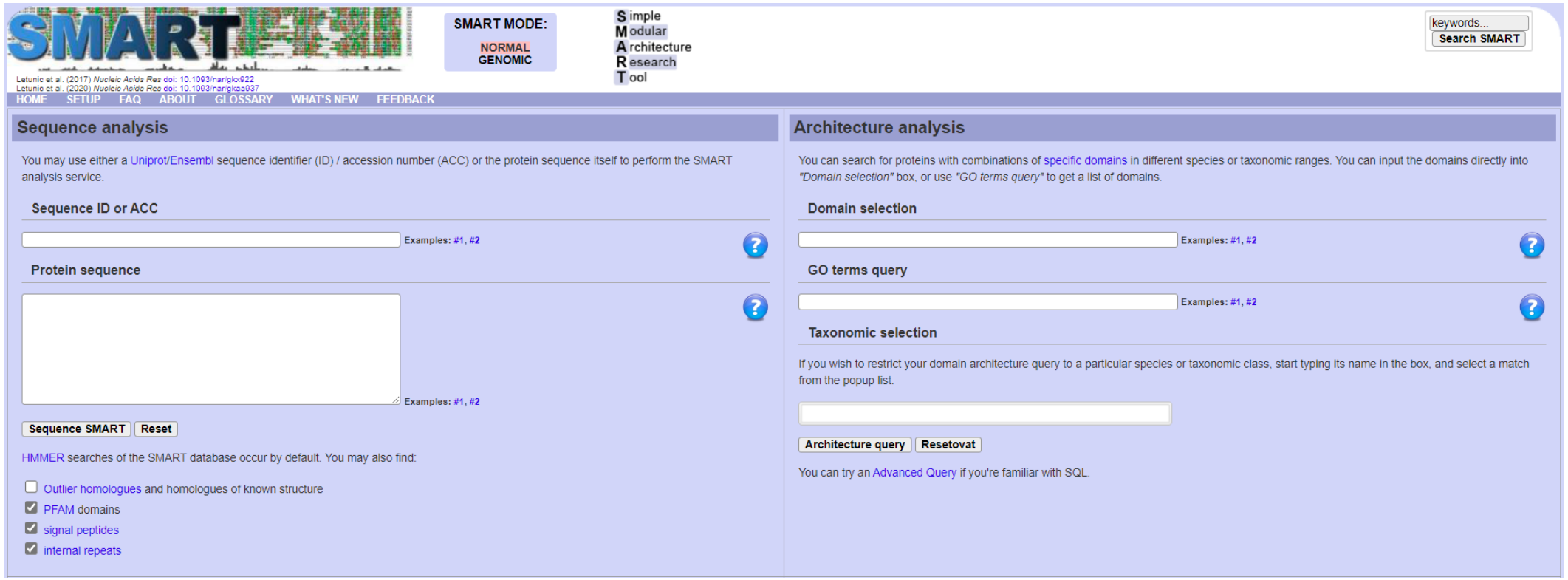
Conserved domain search - CD (NCBI)



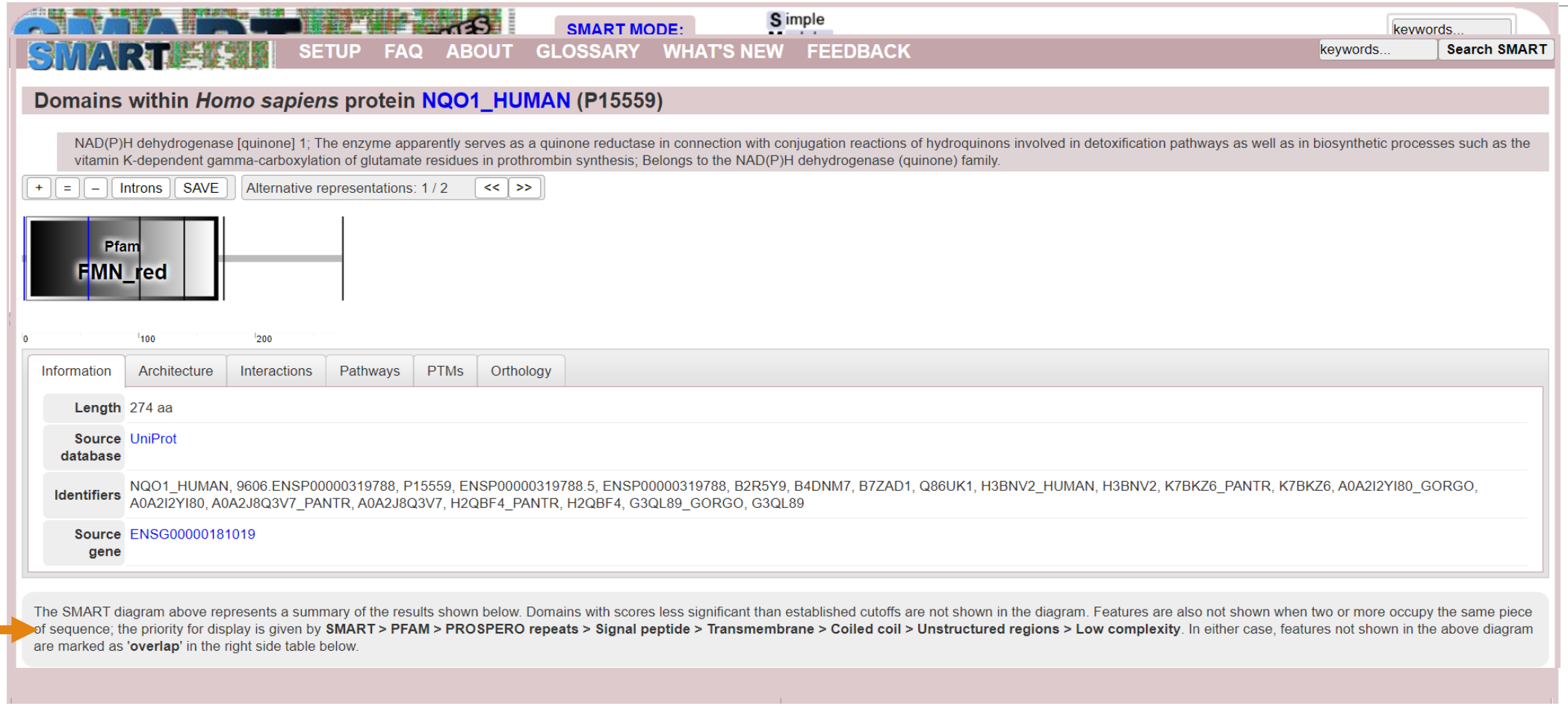
Conserved domain search - SMART



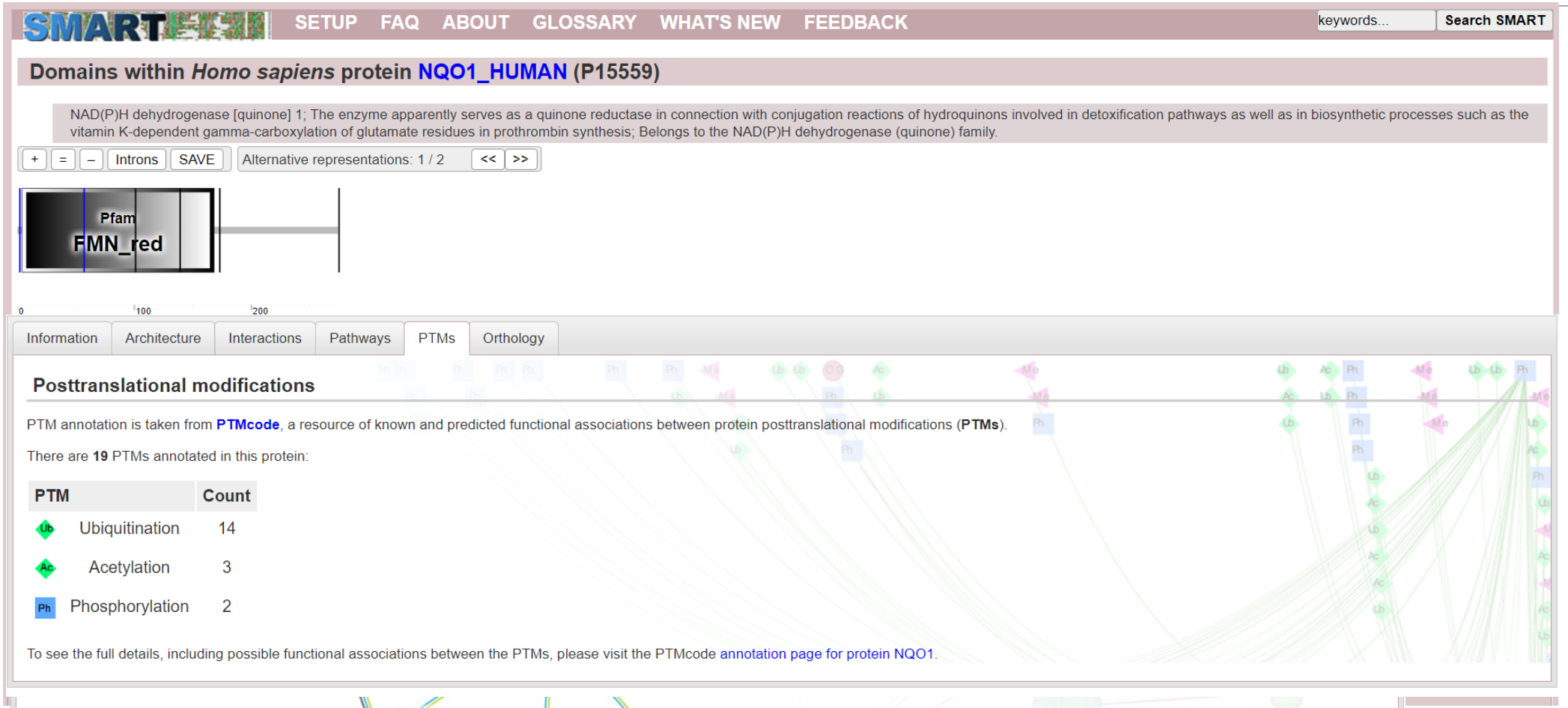
Conserved domain search - SMART



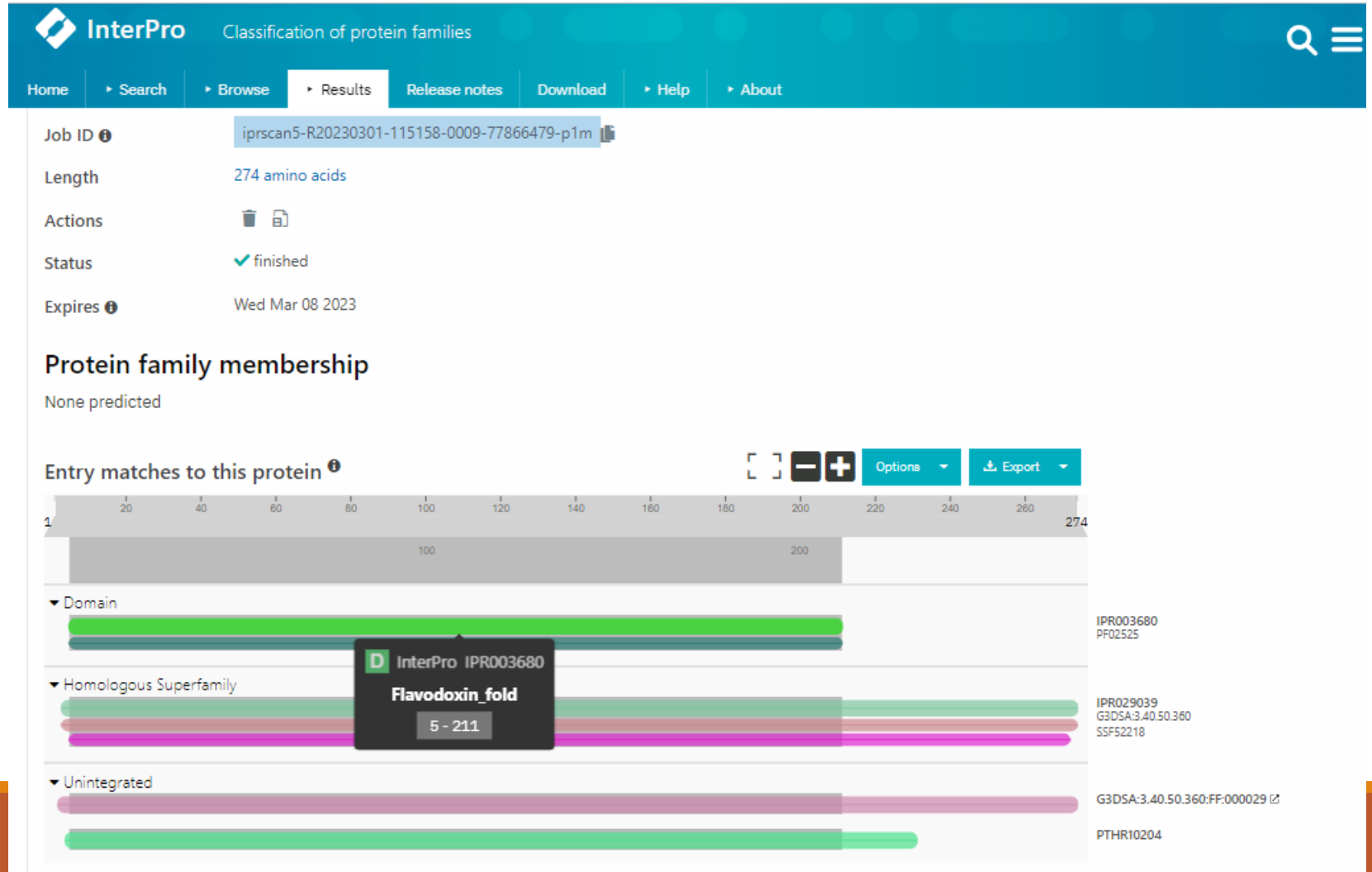
Conserved domain search - SMART



Conserved domain search - SMART



Conserved domain search - InterPro



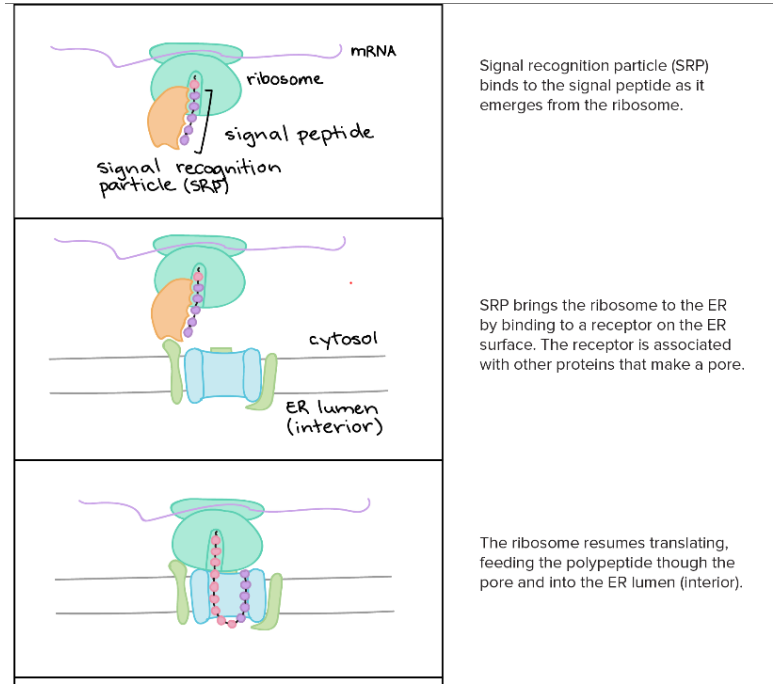
Practical part

Try
CD / SMART/ InterPro
search

Find domains in your sequnce

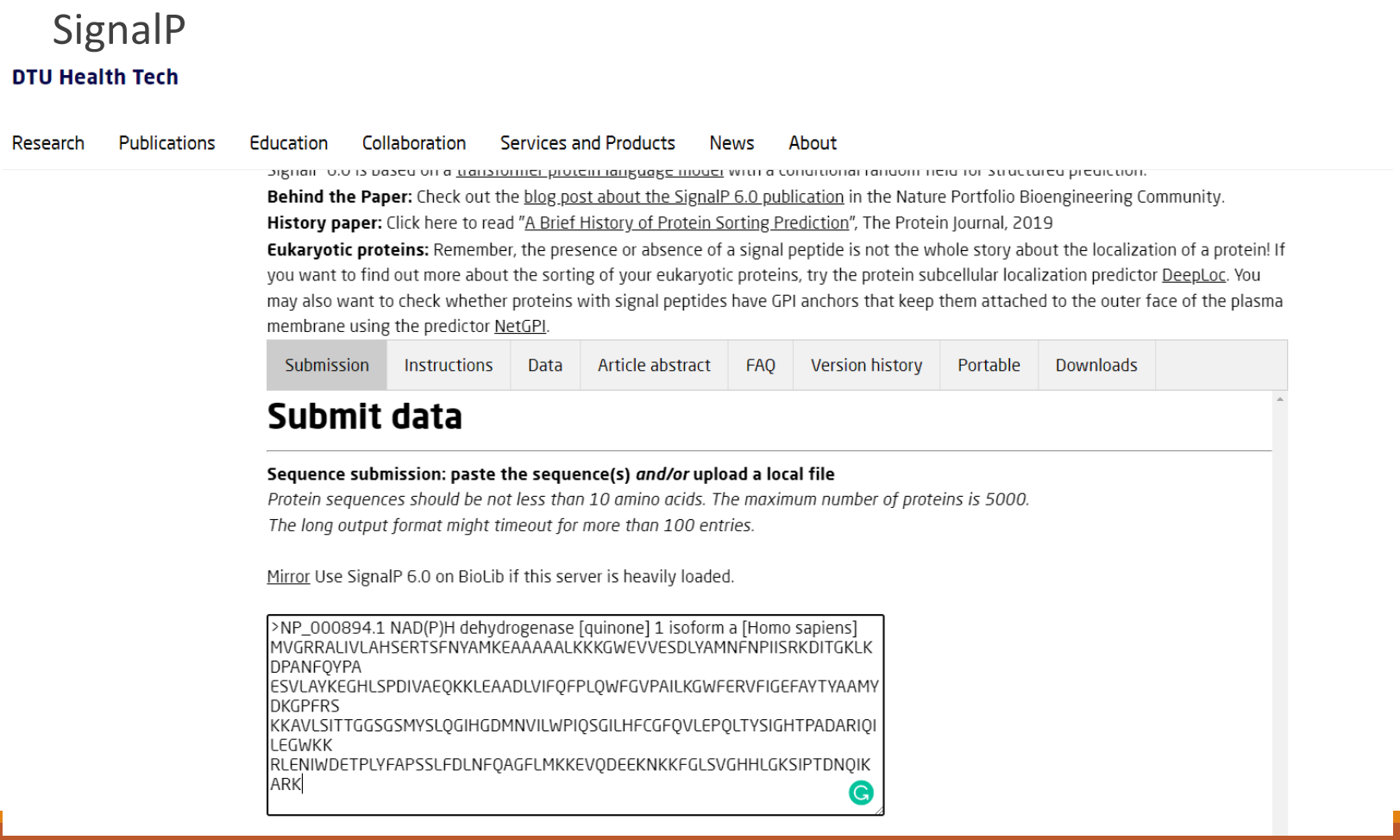
ER signal peptide prediction

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



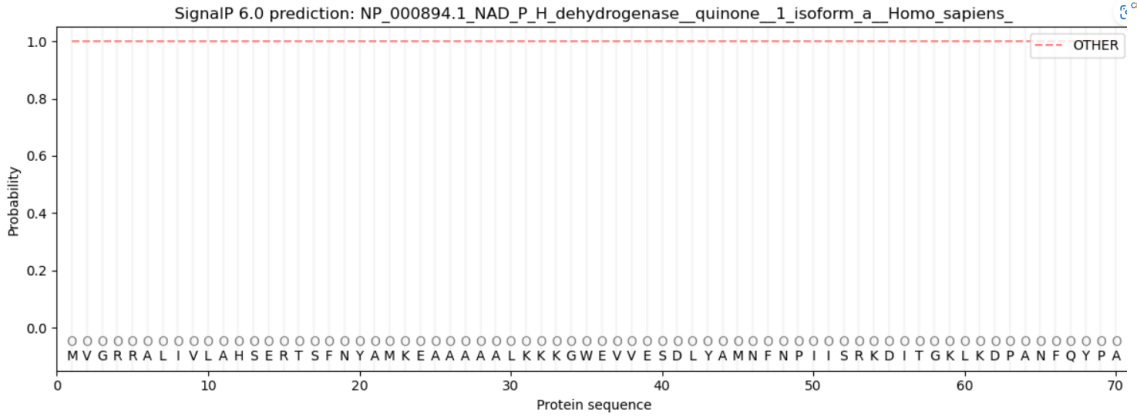
https://www.khanacademy.org/science/biology/gene-expression-central-dogma#central-dogma-transcription

Signal peptides

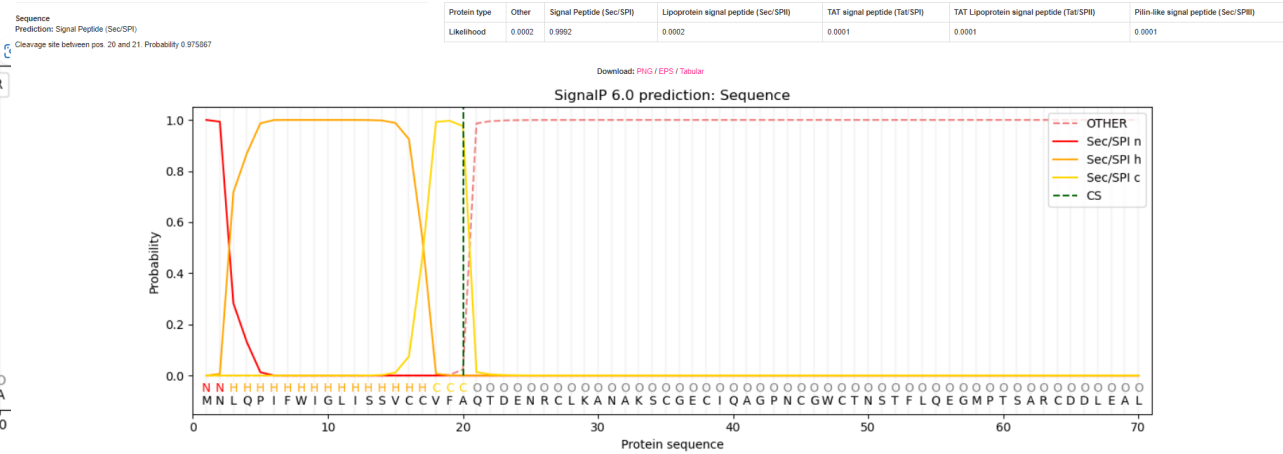


Signal peptides

SignalP

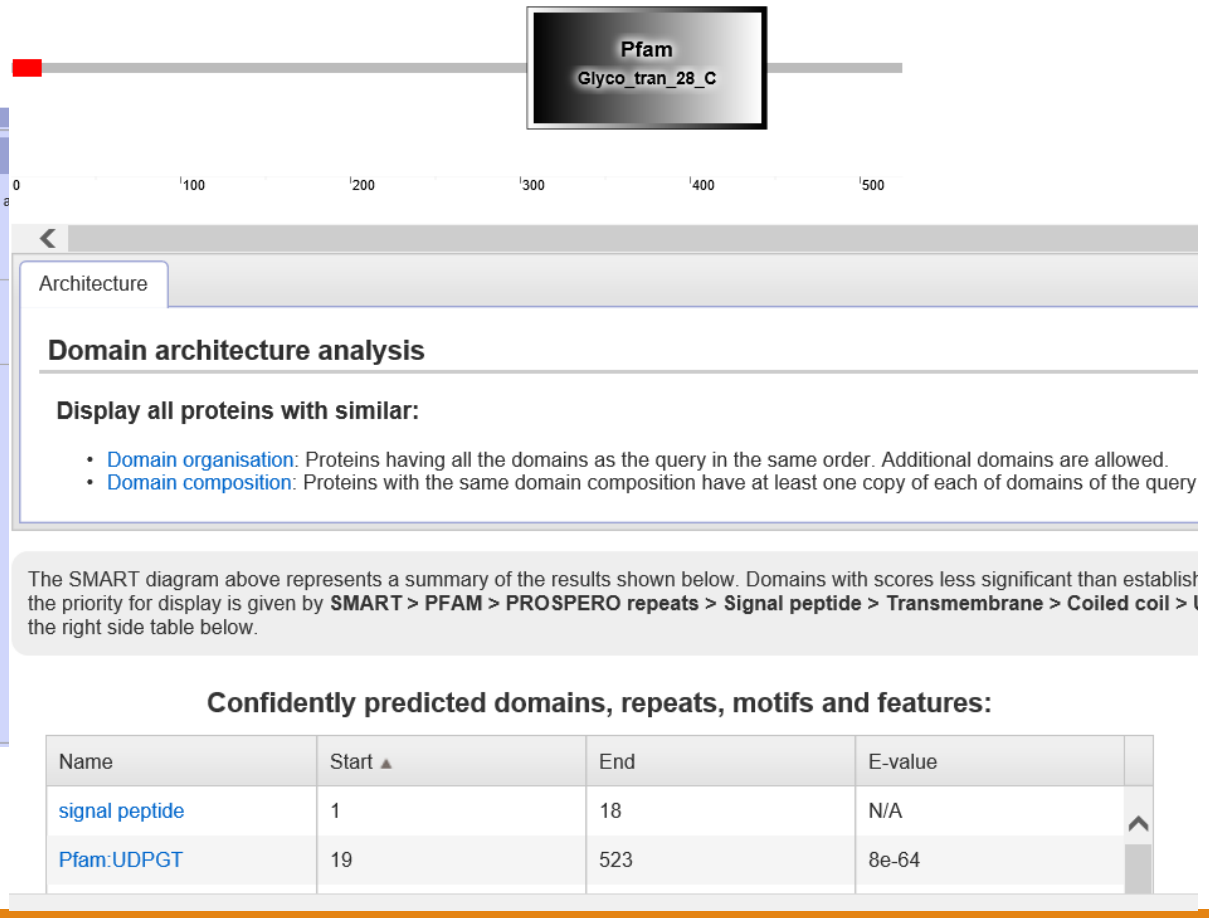


protein does not have signal peptide



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

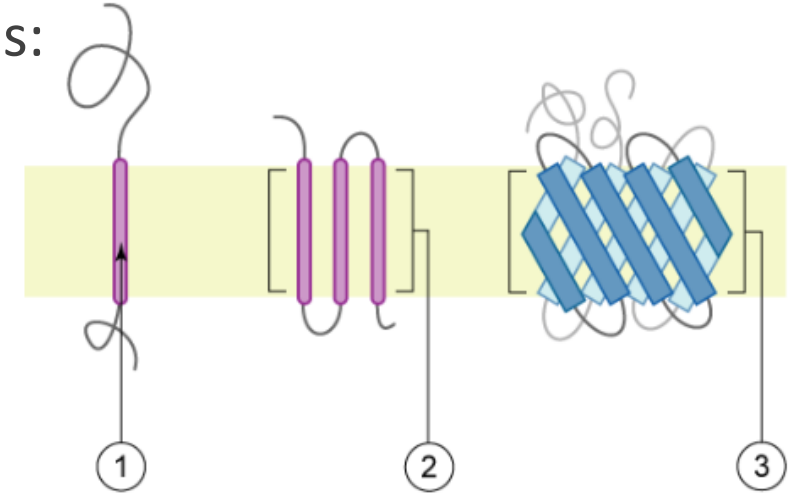
Signal peptides

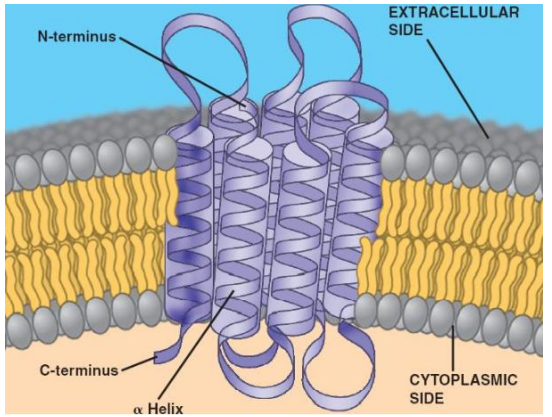


search for signal peptide in your sequnce

Prediction of transmembrane helices

Transmembrane proteins:



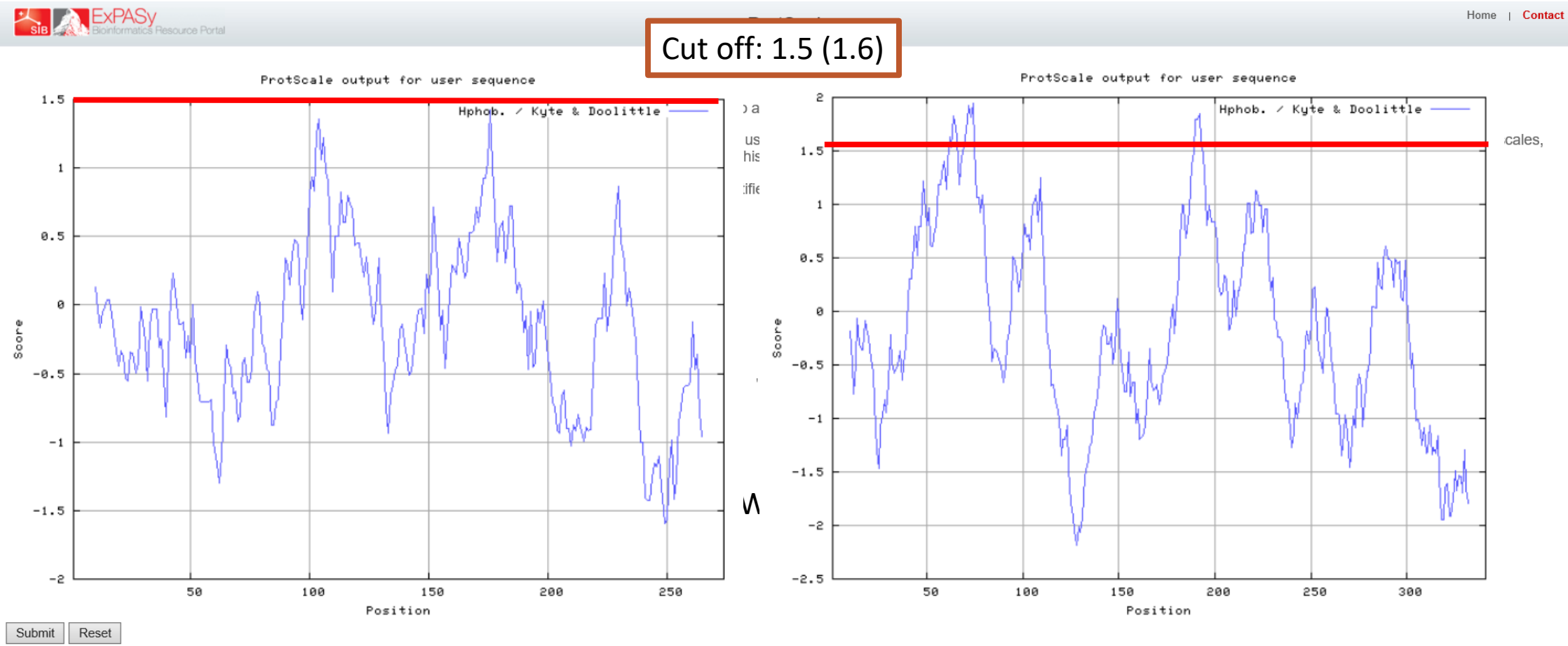


Amino acid Hydrofobicity

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

Prediction of transmembrane helices

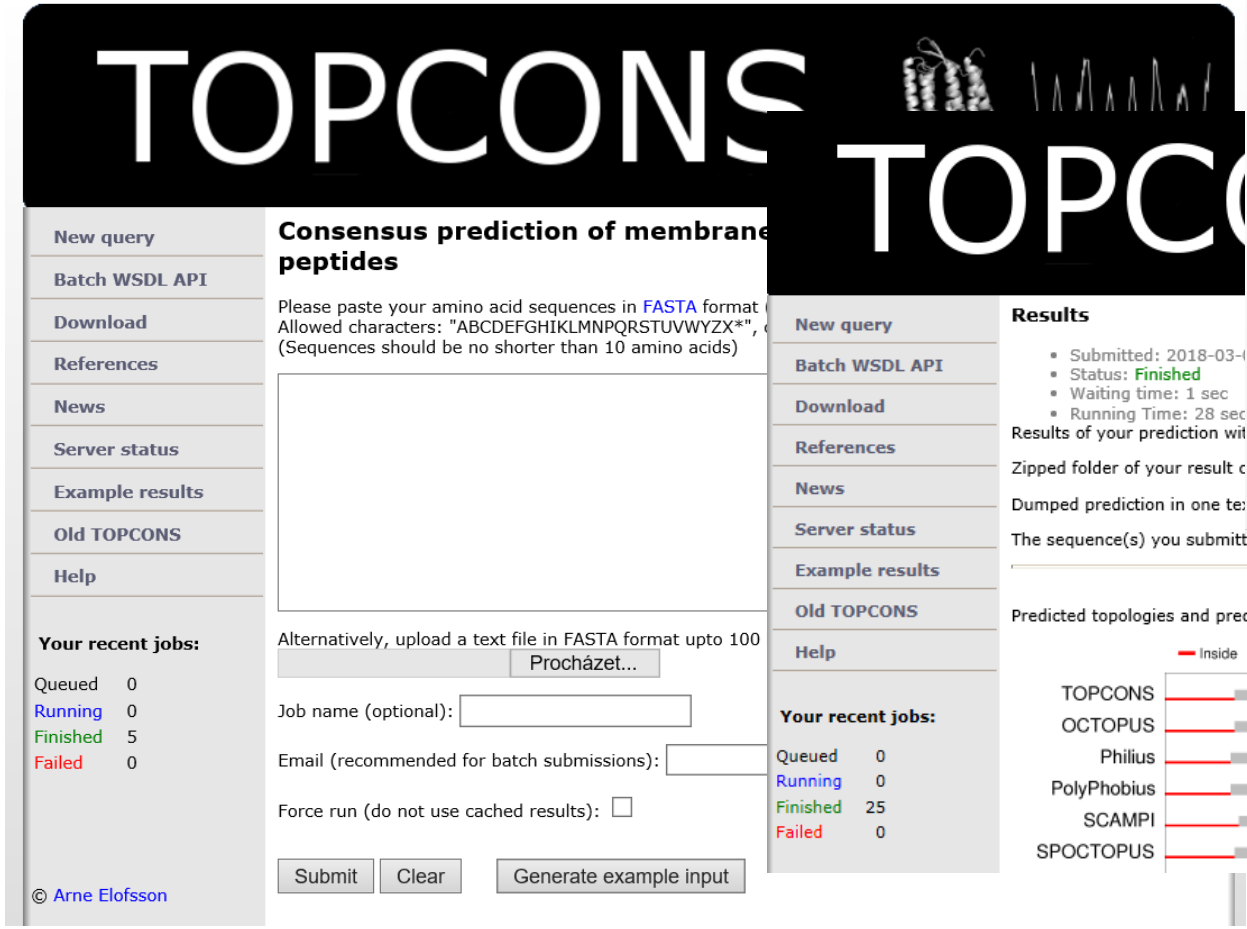
Profile of amino acids hydrofobicity

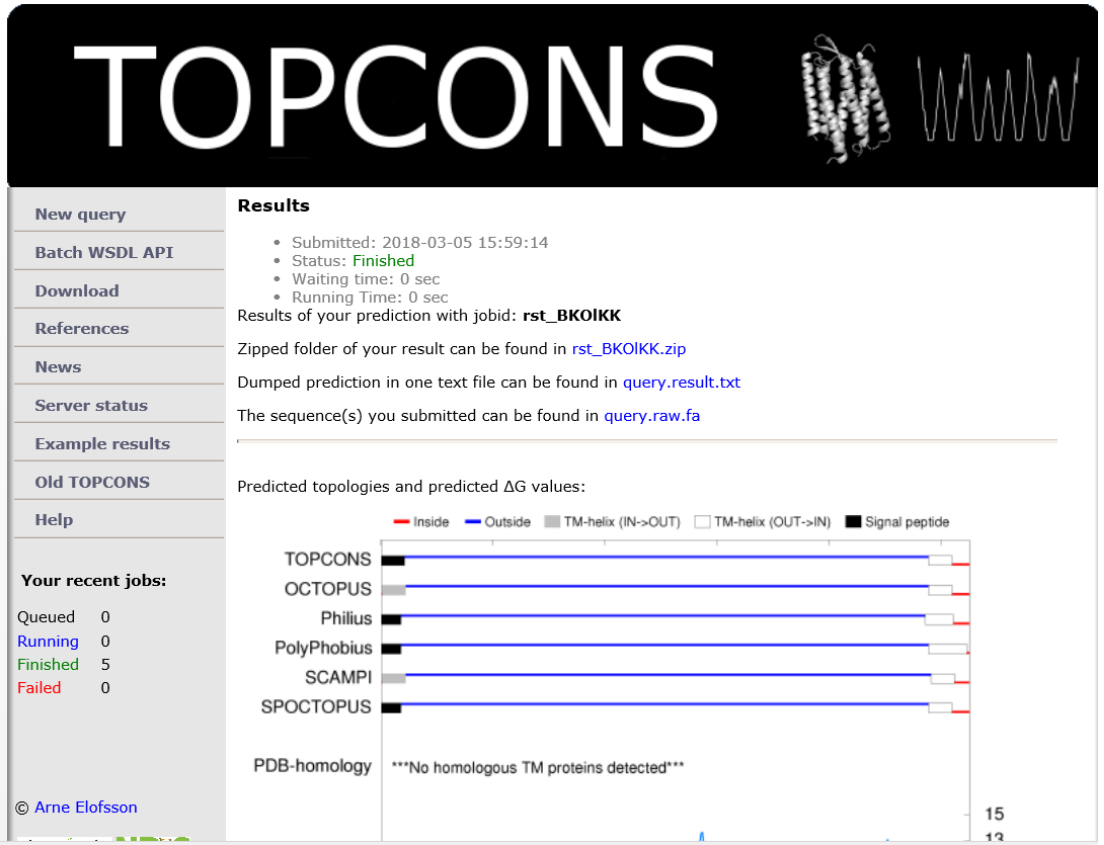


TMHMM


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 c
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:

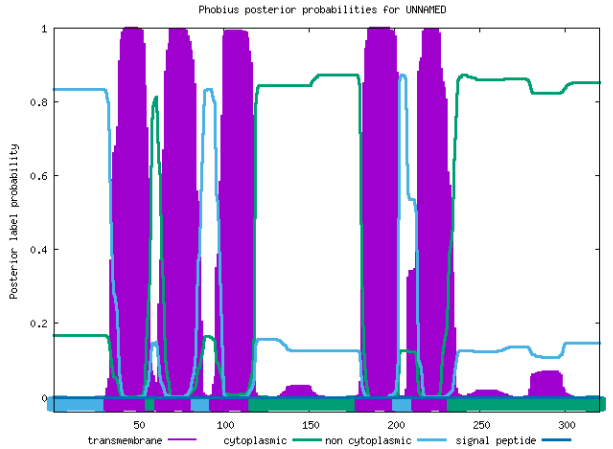
- ○ Short
- ○ Long without Graphics
- ◉ Long with Graphics

Odeslat Resetovat

Phobius prediction

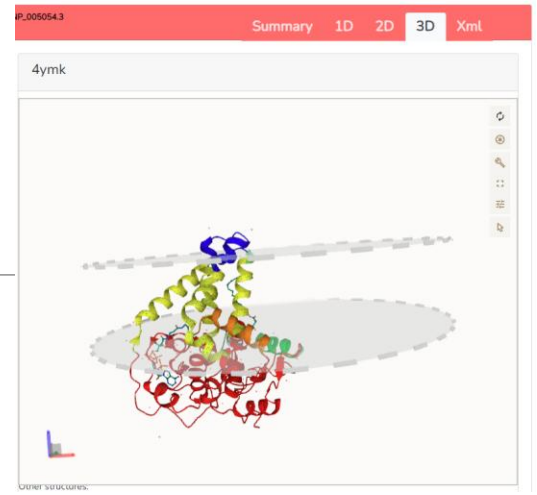
Prediction of UNNAMED

```
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



Submit Manual About Standalone MyJobs Job ID

Results for job 92e87f663074d81d642456ea6e9d63a6

Download results

XML file

Proteins:

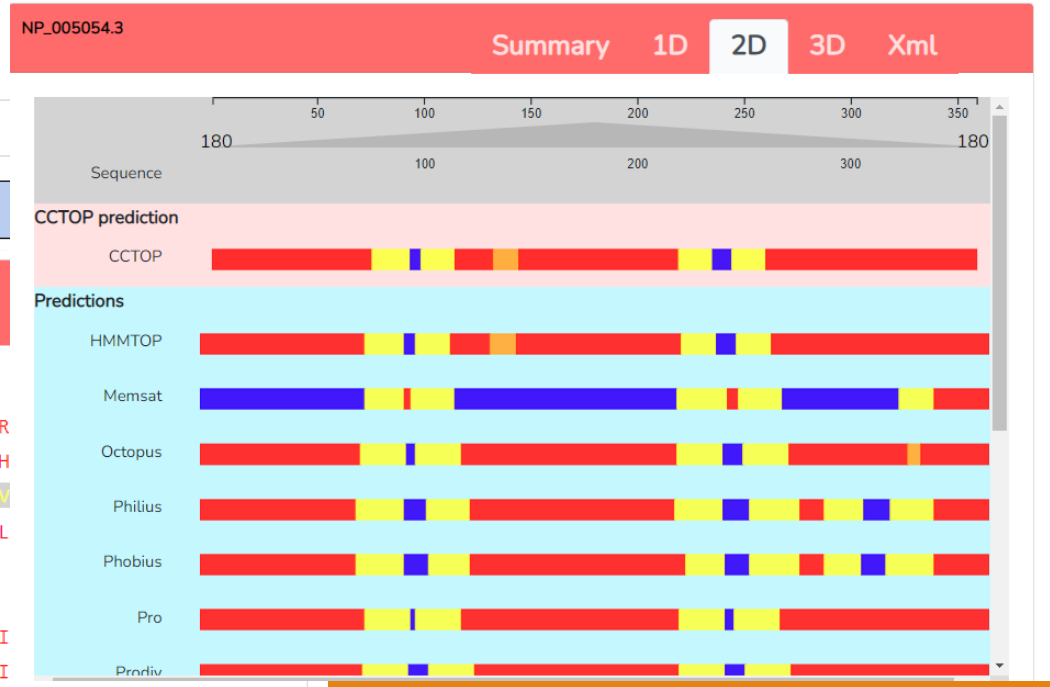
sequences: 1/1

> NP_005054.3

Control panel

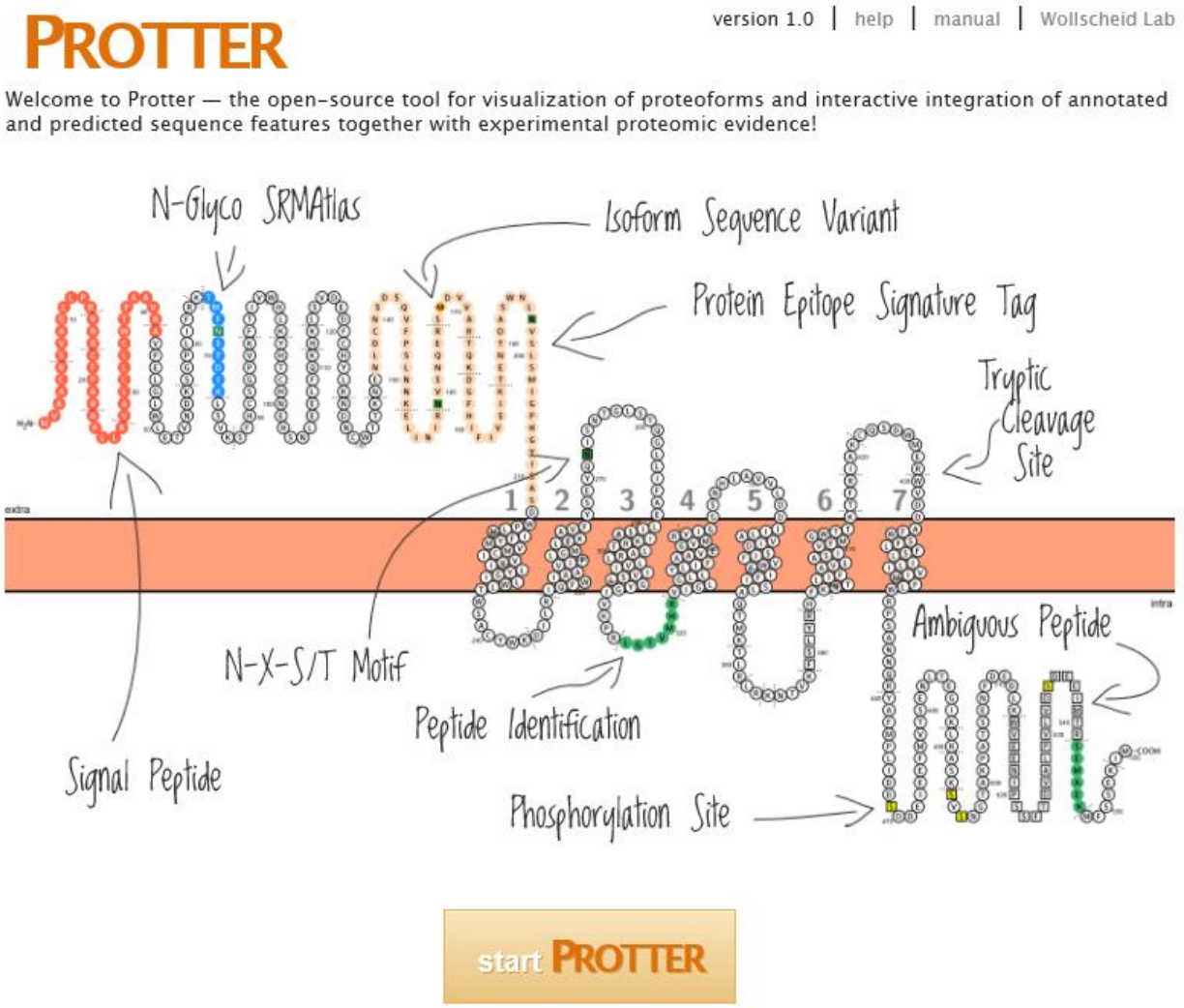
NP_005054.3 Summary 1D 2D

```
>sequence
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVEYVWR
LYGITLIPTCKFVTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
SRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGETFQNSV
NATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFTTFFIDCMAALGL
LARIKRTGDGNYKSG
>topology
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
MMMMMMMOOOOOMMMMMMMMMMMMMMMIIIIIIIIIIIIIIIIIILLLLLLLLLLLIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
MMMMMMMMMMMMMMMMOOOOOOOOOMMMMMMMMMMMMMM
MMIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIII
```



PROTTER-figure!

-creates figure from the UniProt data



PROTTER-figure!

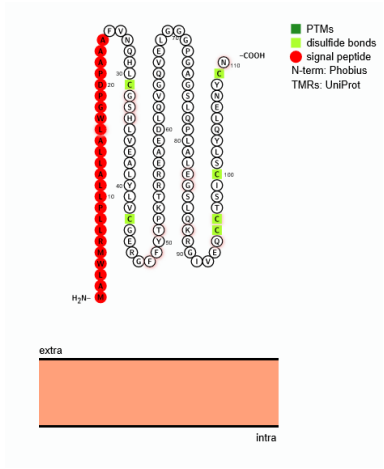
-creates figure from the UniProt data

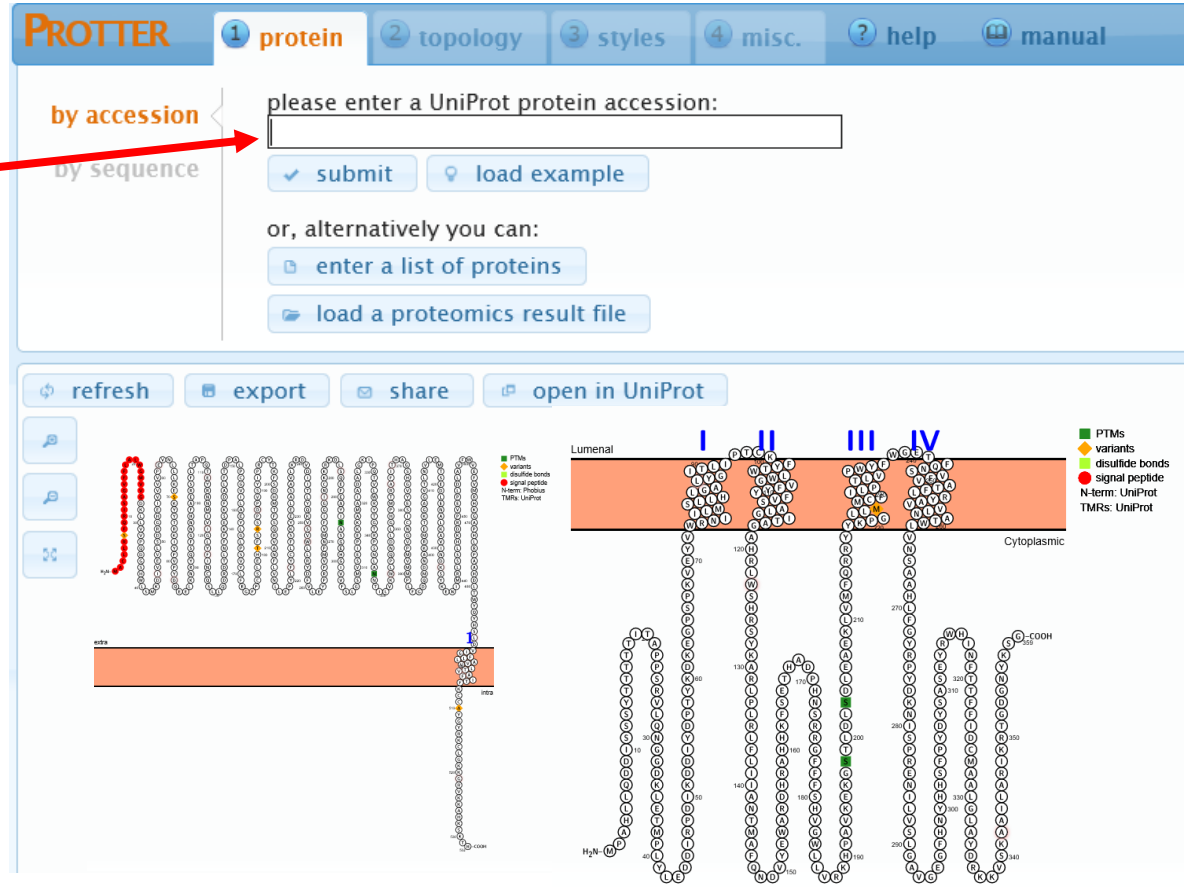
-uses UniProt ID:

UGT1A6 (P19224)

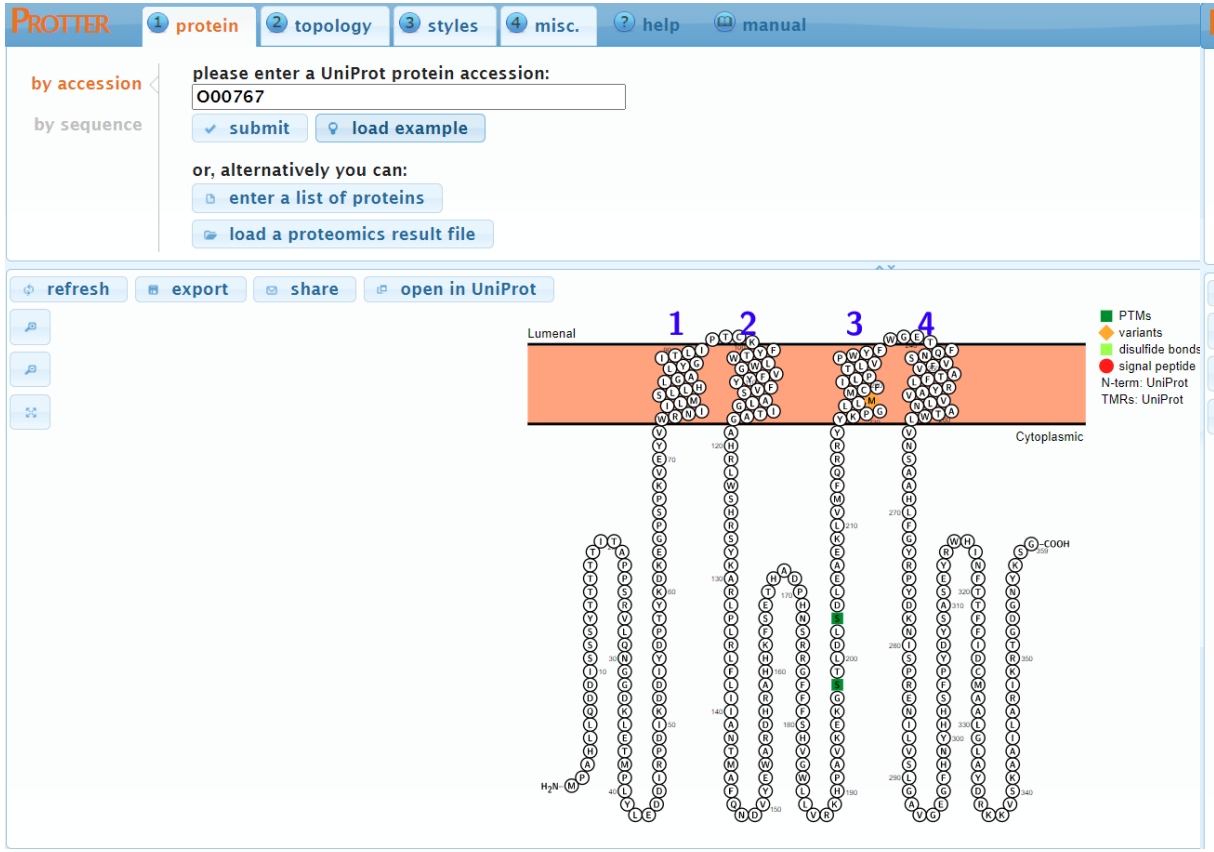
Desaturase (O00767)

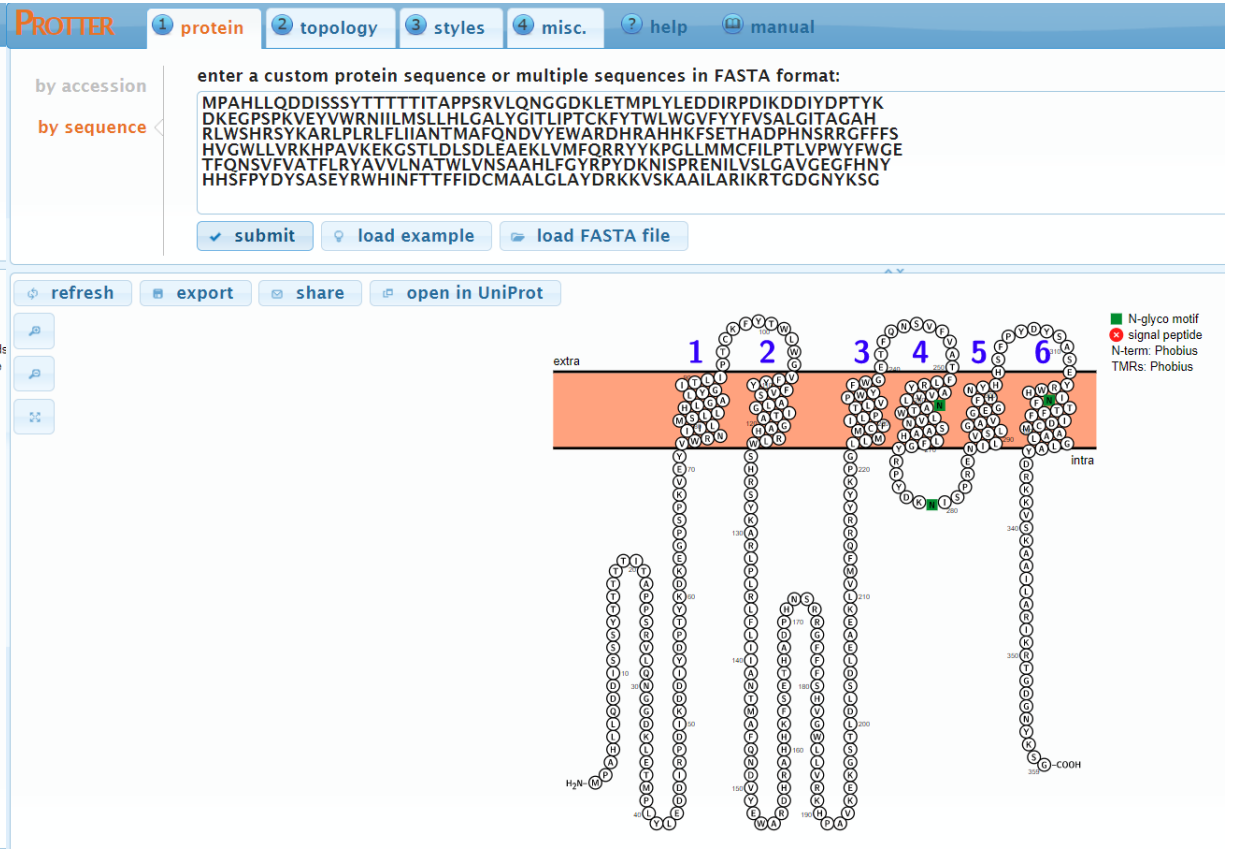
(prepro) insulin (P01308)





PROTTER-figure!



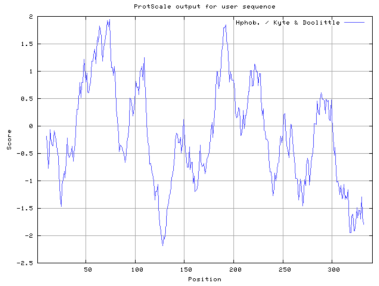


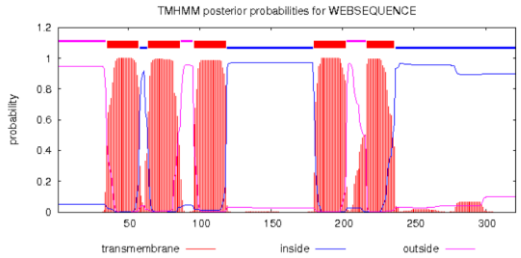
Prediction of transmembrane helices

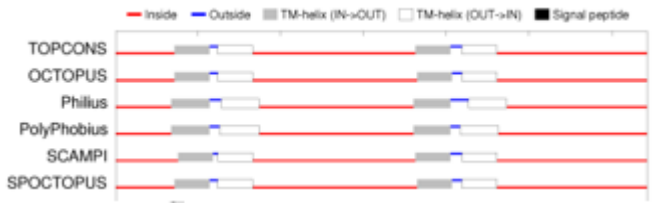
ProtScale

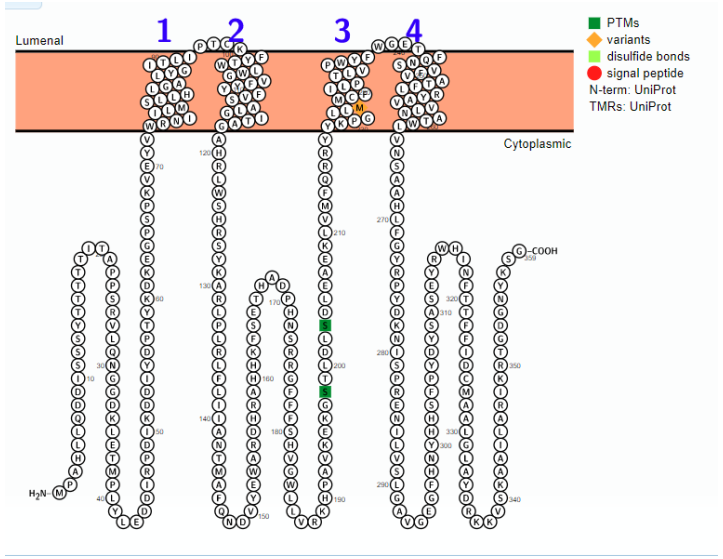
TMHMM

TOPCONS









- **Always try more programs!**

Try more programs.

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

„Protein bioinformatics I“ next lecture

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

…

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 a signal peptide (ER retention signal)?

E.g use „výstřižky“  „snipping tool“

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

Homework 3 - example

1) Trypsin 33

2) 1)

3) 2)

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

4)