Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2024/2025

3/10

„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)

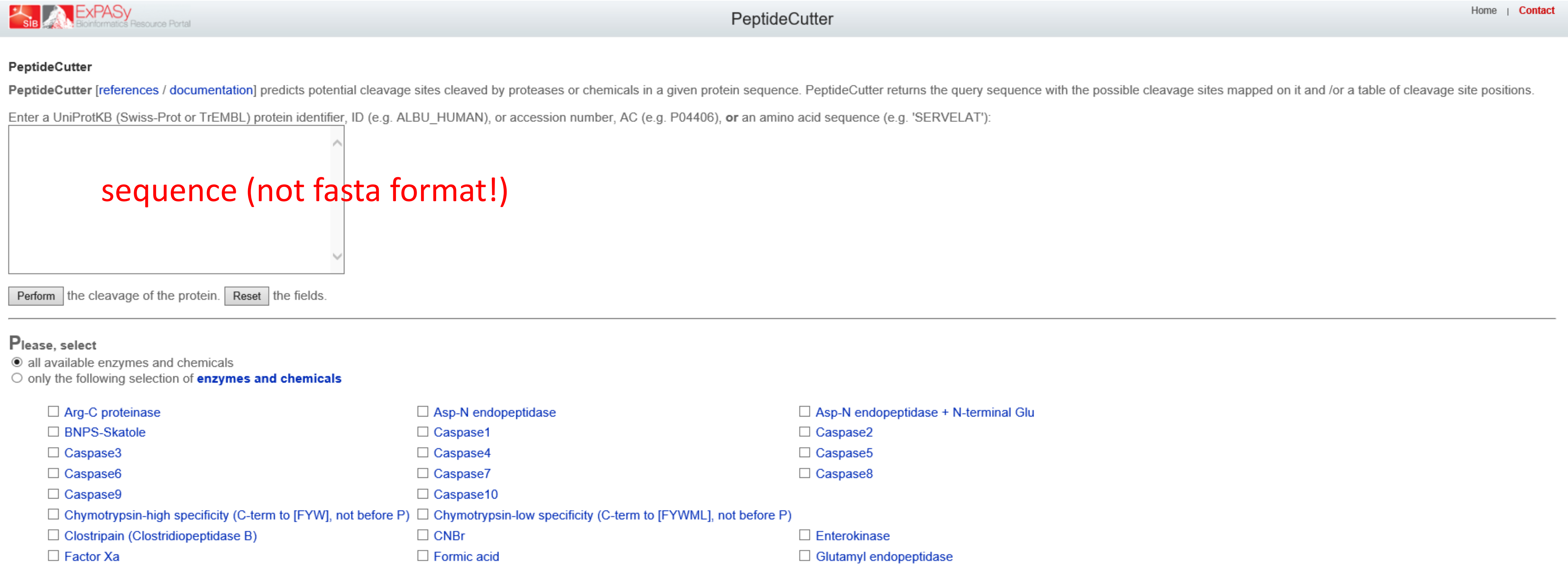
proteinase K - commonly used in molecular biology to digest protein and remove contamination from preparations of nucleic acid.

- cleaves ubiquitously

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

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

Prediction of proteases cutting



PeptideCutter

PeptideCutter [references / documentation] predicts potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. PeptideCutter returns the query sequence with the possible cleavage sites mapped on it and /or a table of cleavage site positions.

Enter a UniProtKB (Swiss-Prot or TrEMBL) protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SERVELAT'):

sequence (not fasta format!)

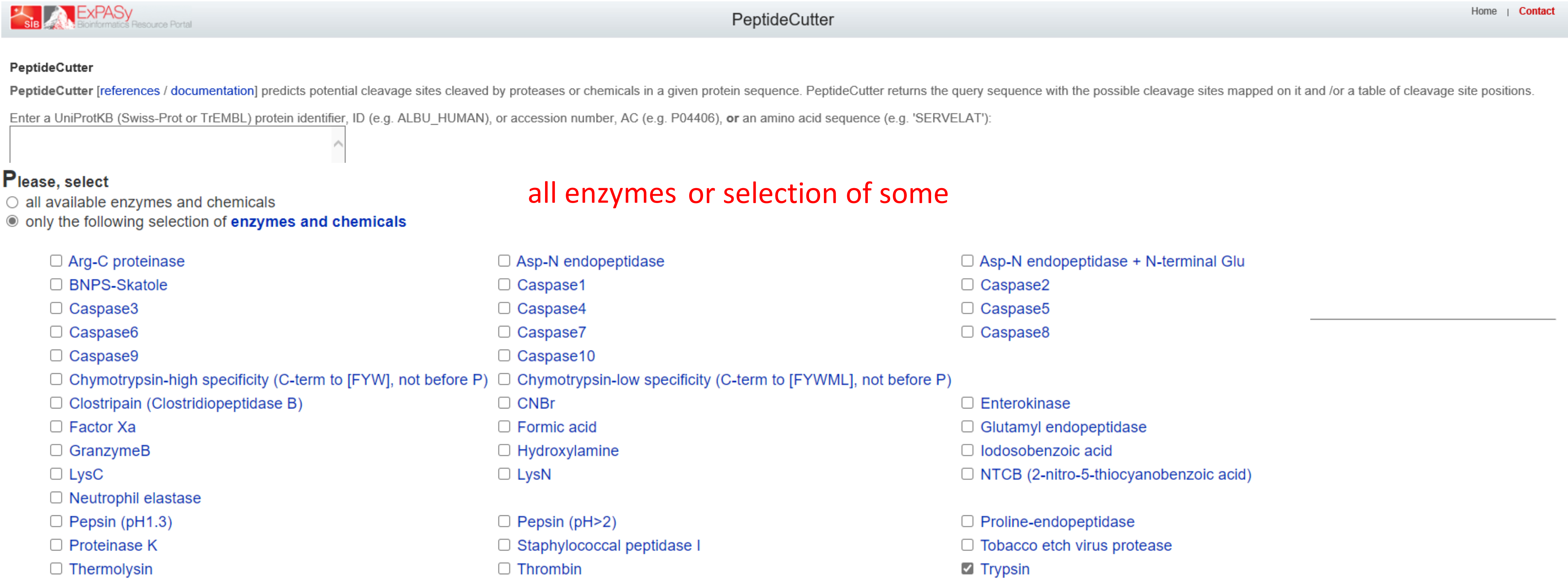
Perform the cleavage of the protein. Reset the fields.

Please, select

● all available enzymes and chemicals
○ only the following selection of **enzymes and chemicals**

- □ Arg-C proteinase
- □ Asp-N endopeptidase
- □ Asp-N endopeptidase + N-terminal Glu
- □ BNPS-Skatole
- □ Caspase1
- □ Caspase2
- □ Caspase3
- □ Caspase4
- □ Caspase5
- □ Caspase6
- □ Caspase7
- □ Caspase8
- □ Caspase9
- □ Caspase10
- □ Chymotrypsin-high specificity (C-term to [FYW], not before P)
- □ Chymotrypsin-low specificity (C-term to [FYWML], not before P)
- □ Clostripain (Clostridiopeptidase B)
- □ CNBr
- □ Enterokinase
- □ Factor Xa
- □ Formic acid
- □ Glutamyl endopeptidase

Prediction of proteases cutting



PeptideCutter

PeptideCutter

PeptideCutter [references / documentation] predicts potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. PeptideCutter returns the query sequence with the possible cleavage sites mapped on it and /or a table of cleavage site positions.

Enter a UniProtKB (Swiss-Prot or TrEMBL) protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SERVELAT'):

Please, select

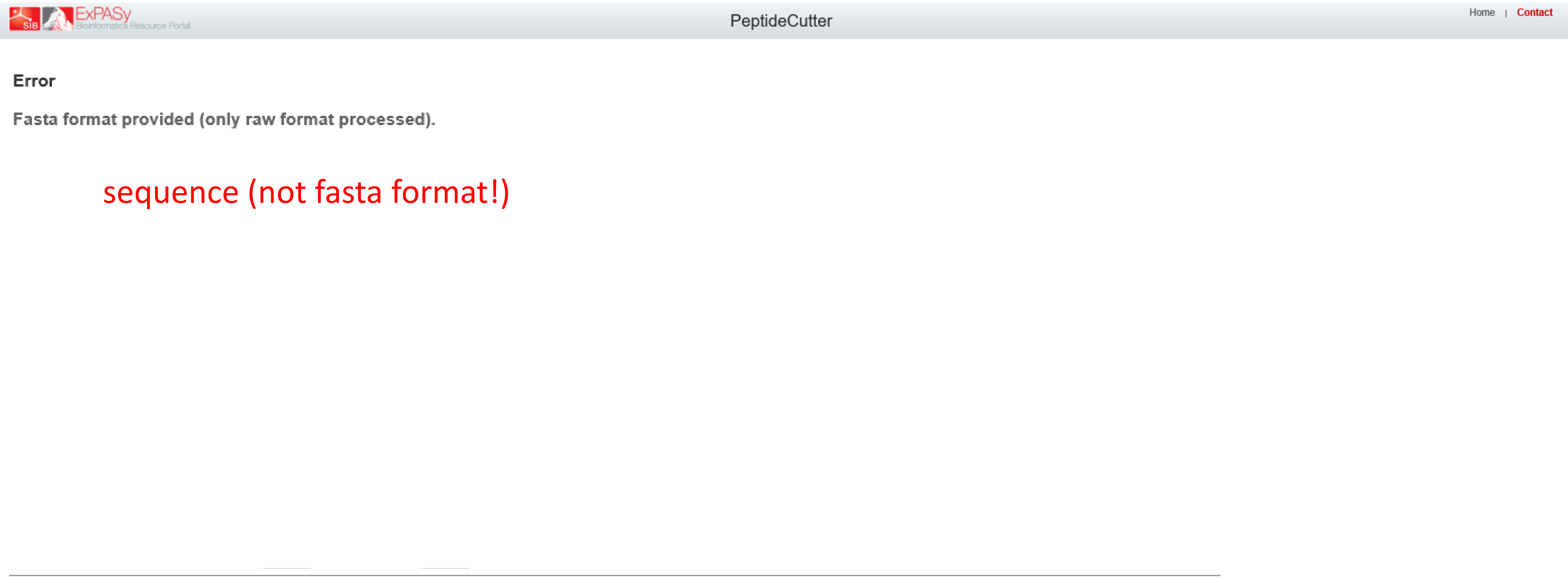
○ all available enzymes and chemicals

● only the following selection of **enzymes and chemicals**

all enzymes or selection of some

□ Arg-C proteinase □ Asp-N endopeptidase □ Asp-N endopeptidase + N-terminal Glu
□ BNPS-Skatole □ Caspase1 □ Caspase2
□ Caspase3 □ Caspase4 □ Caspase5
□ Caspase6 □ Caspase7 □ Caspase8
□ Caspase9 □ Caspase10
□ Chymotrypsin-high specificity (C-term to [FYW], not before P) □ Chymotrypsin-low specificity (C-term to [FYWML], not before P)
□ Clostripain (Clostridiopeptidase B) □ CNBr □ Enterokinase
□ Factor Xa □ Formic acid □ Glutamyl endopeptidase
□ GranzymeB □ Hydroxylamine □ Iodosobenzoic acid
□ LysC □ LysN □ NTCB (2-nitro-5-thiocyanobenzoic acid)
□ Neutrophil elastase
□ Pepsin (pH1.3) □ Pepsin (pH>2) □ Proline-endopeptidase
□ Proteinase K □ Staphylococcal peptidase I □ Tobacco etch virus protease
□ Thermolysin □ Thrombin ☑ Trypsin

Prediction of proteases cutting

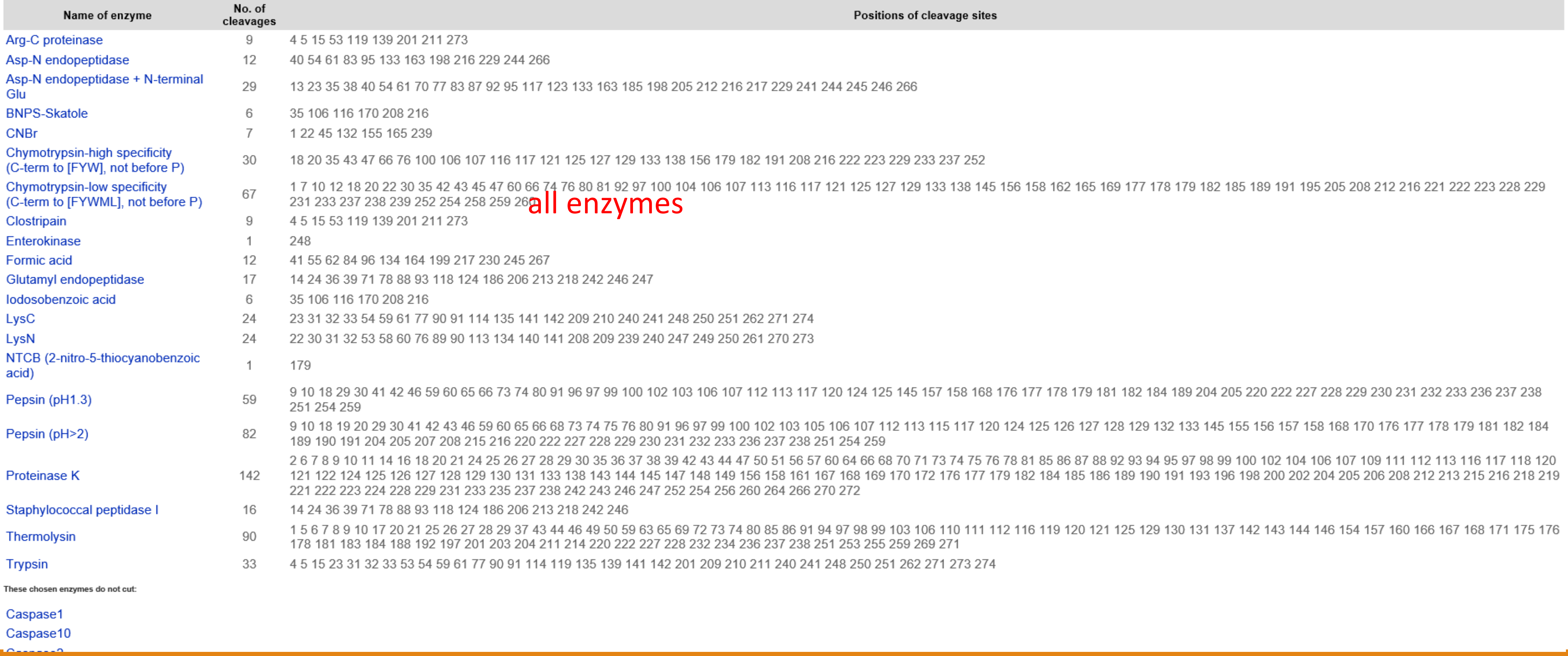


Error

Fasta format provided (only raw format processed).

sequence (not fasta format!)

Prediction of proteases cutting



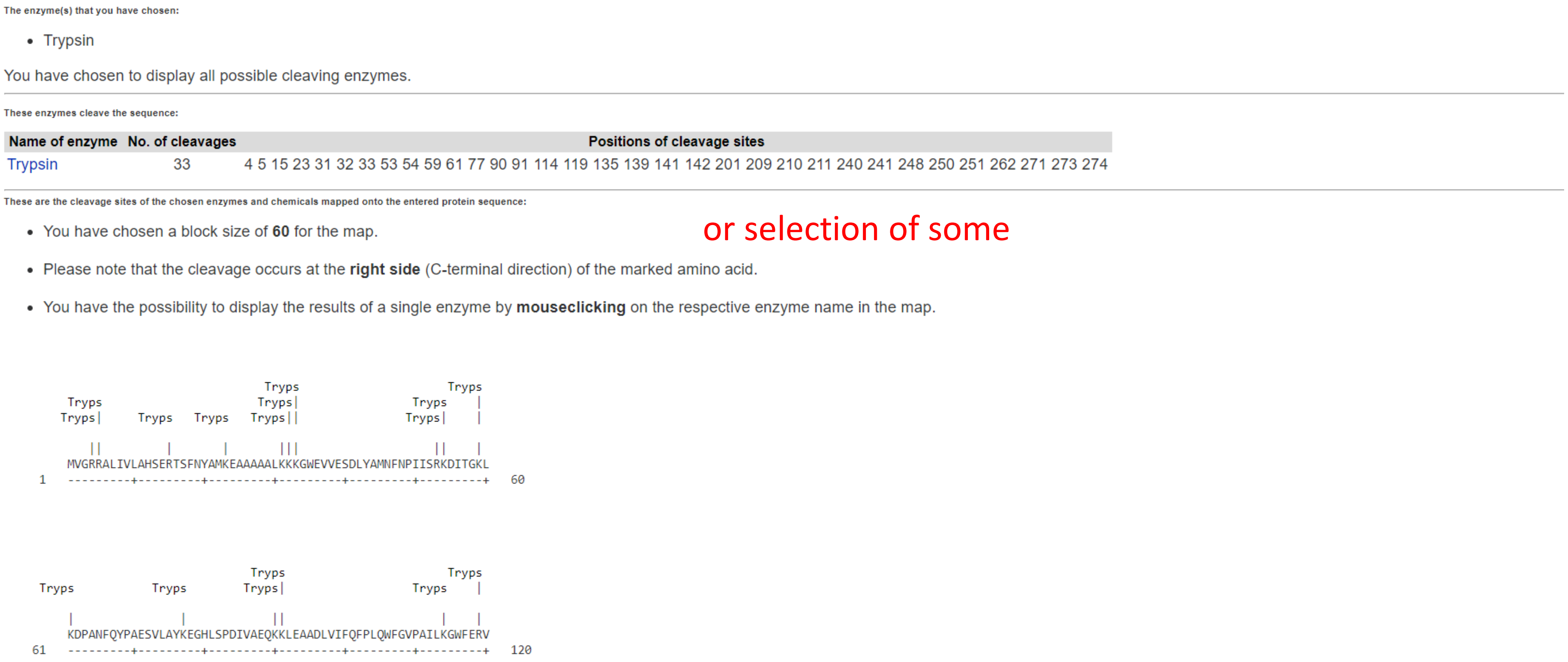
Name of enzyme	No. of cleavages	Positions of cleavage sites
Arg-C proteinase	9	4 5 15 53 119 139 201 211 273
Asp-N endopeptidase	12	40 54 61 83 95 133 163 198 216 229 244 266
Asp-N endopeptidase + N-terminal Glu	29	13 23 35 38 40 54 61 70 77 83 87 92 95 117 123 133 163 185 198 205 212 216 217 229 241 244 245 246 266
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 26
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

These chosen enzymes do not cut:

Caspase1
Caspase10
Caspase2

all enzymes

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

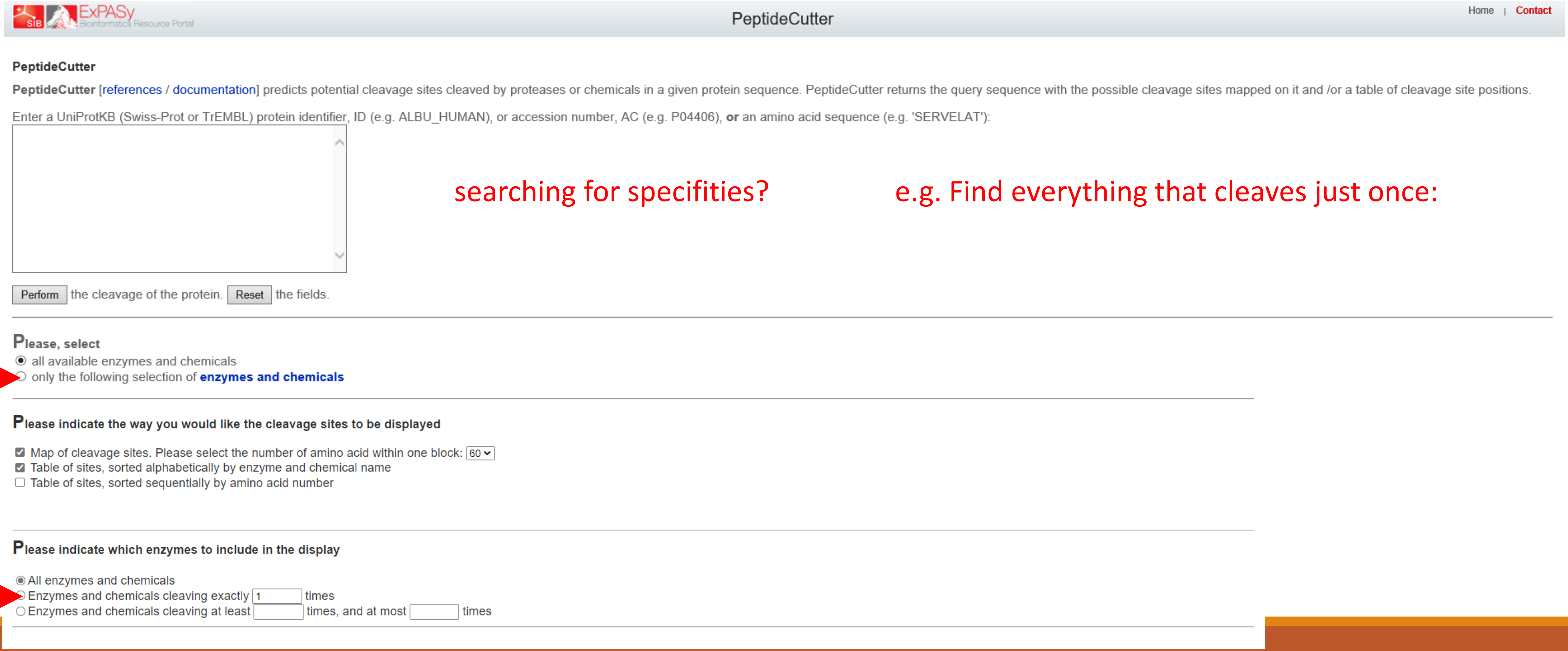
```
                                 Tryps                     Tryps
     Tryps                      Tryps|                Tryps    |
    Tryps|     Tryps   Tryps   Tryps||               Tryps|    |

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


                               Tryps                       Tryps
 Tryps           Tryps        Tryps|                  Tryps    |

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

Prediction of proteases cutting



PeptideCutter

PeptideCutter [references / documentation] predicts potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. PeptideCutter returns the query sequence with the possible cleavage sites mapped on it and /or a table of cleavage site positions.

Enter a UniProtKB (Swiss-Prot or TrEMBL) protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SERVELAT'):

searching for specifities?

e.g. Find everything that cleaves just once:

Perform the cleavage of the protein. Reset the fields.

Please, select

● all available enzymes and chemicals
○ only the following selection of **enzymes and chemicals**

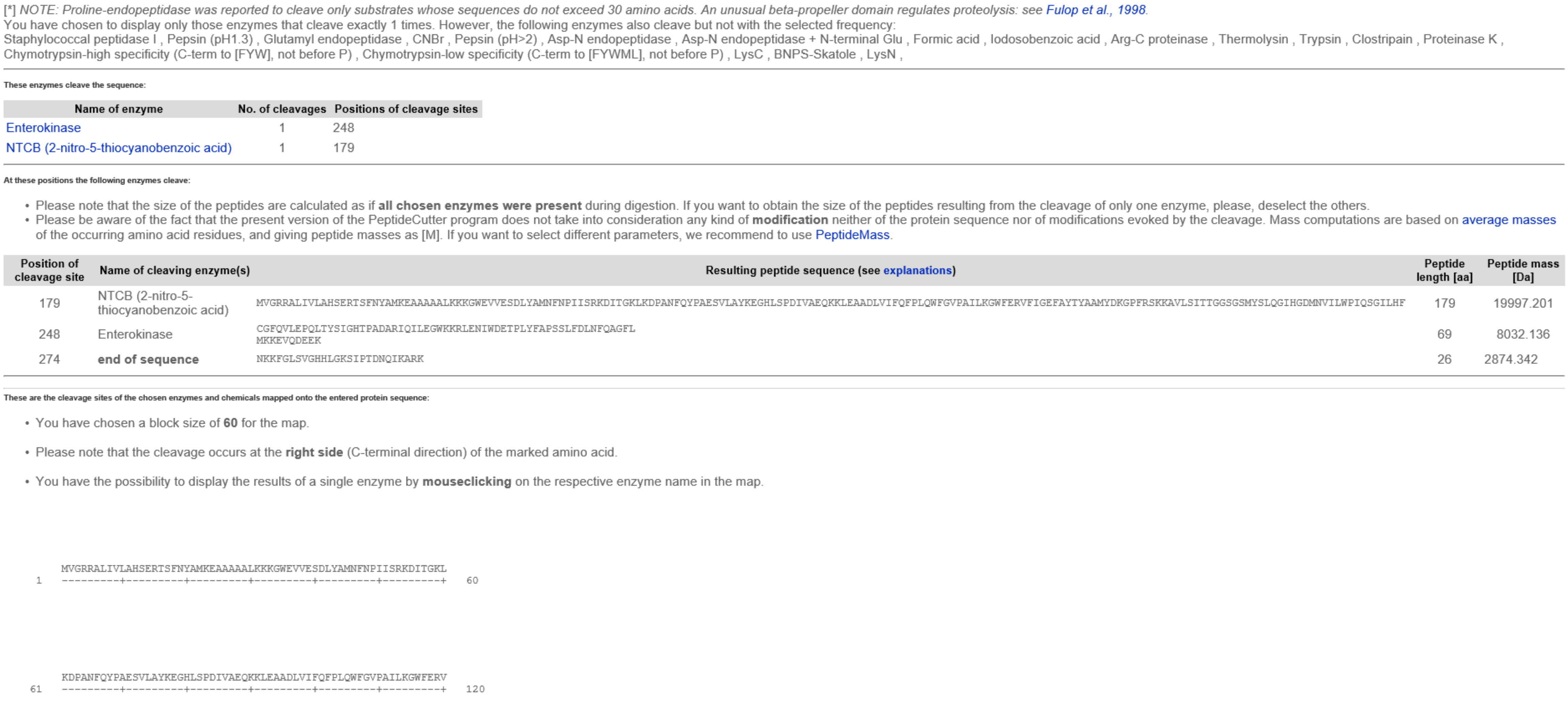
Please indicate the way you would like the cleavage sites to be displayed

☑ Map of cleavage sites. Please select the number of amino acid within one block: 60
☑ Table of sites, sorted alphabetically by enzyme and chemical name
☐ Table of sites, sorted sequentially by amino acid number

Please indicate which enzymes to include in the display

● All enzymes and chemicals
○ Enzymes and chemicals cleaving exactly 1 times
○ Enzymes and chemicals cleaving at least times, and at most times

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:

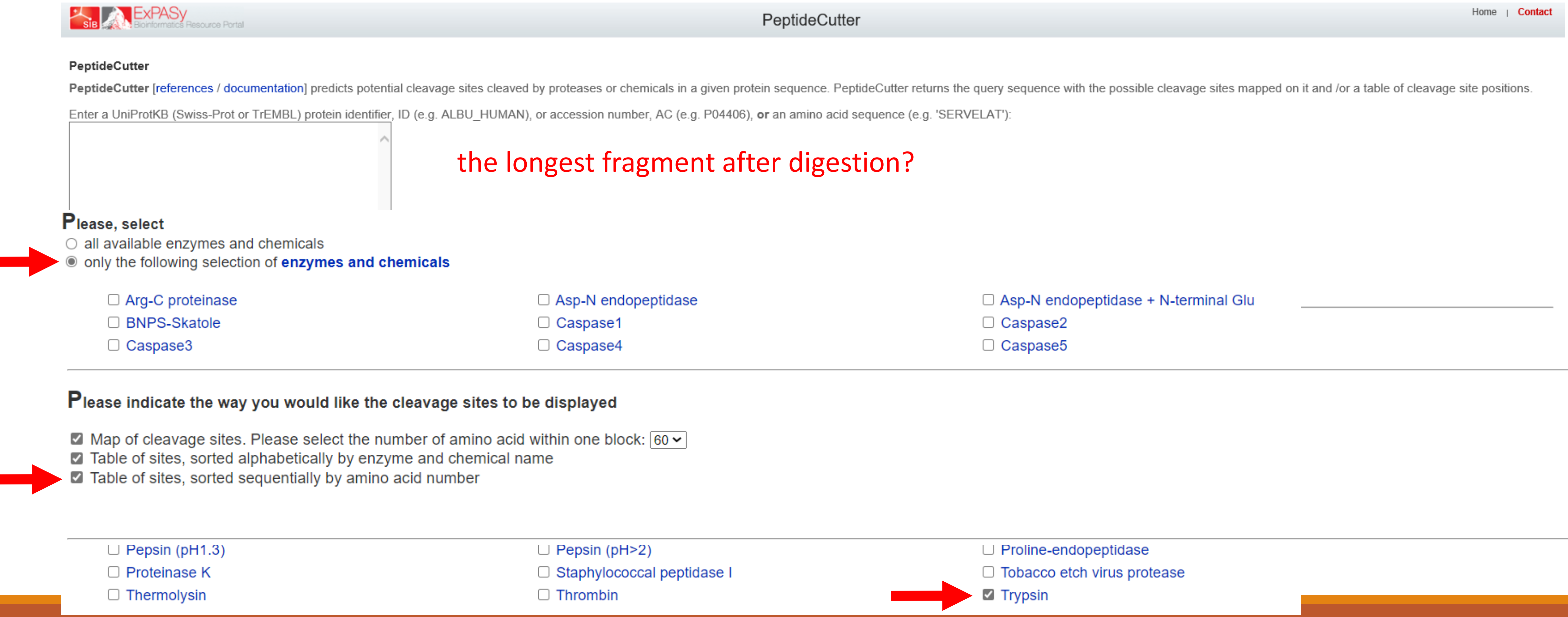
- You have chosen a block size of **60** for the map.
- Please note that the cleavage occurs at the **right side** (C-terminal direction) of the marked amino acid.
- You have the possibility to display the results of a single enzyme by **mouseclicking** on the respective enzyme name in the map.

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



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

Prediction of proteases cutting



PeptideCutter

PeptideCutter

PeptideCutter [references / documentation] predicts potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. PeptideCutter returns the query sequence with the possible cleavage sites mapped on it and /or a table of cleavage site positions.

Enter a UniProtKB (Swiss-Prot or TrEMBL) protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SERVELAT'):

the longest fragment after digestion?

Please, select

○ all available enzymes and chemicals

● only the following selection of **enzymes and chemicals**

□ Arg-C proteinase
□ BNPS-Skatole
□ Caspase3
□ Asp-N endopeptidase
□ Caspase1
□ Caspase4
□ Asp-N endopeptidase + N-terminal Glu
□ Caspase2
□ Caspase5

Please indicate the way you would like the cleavage sites to be displayed

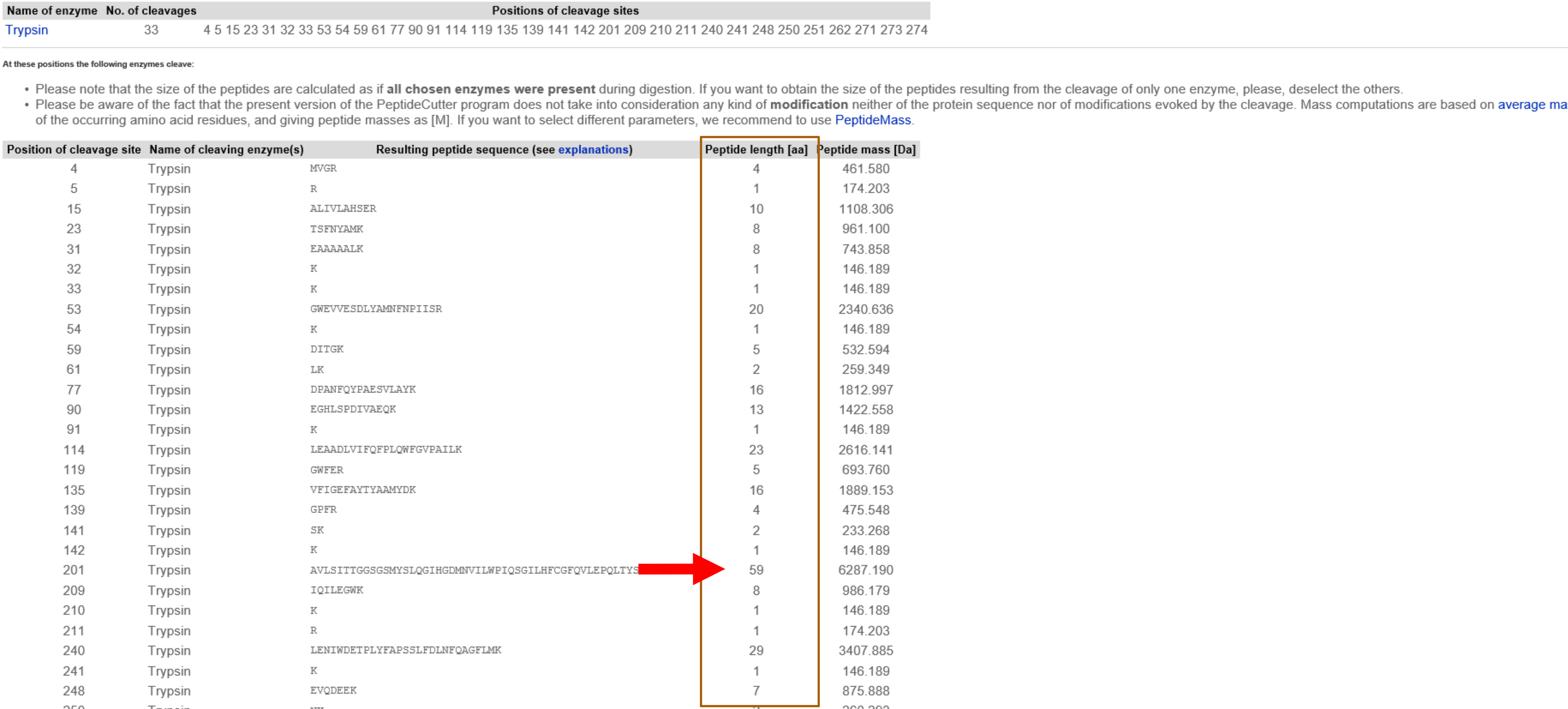
☑ Map of cleavage sites. Please select the number of amino acid within one block: 60

☑ Table of sites, sorted alphabetically by enzyme and chemical name

☑ Table of sites, sorted sequentially by amino acid number

□ Pepsin (pH1.3)
□ Proteinase K
□ Thermolysin
□ Pepsin (pH>2)
□ Staphylococcal peptidase I
□ Thrombin
□ Proline-endopeptidase
□ Tobacco etch virus protease
☑ Trypsin

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

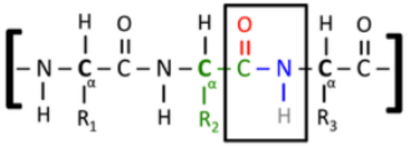
→ **Finding 3-D structure and the domain organization of proteins**

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

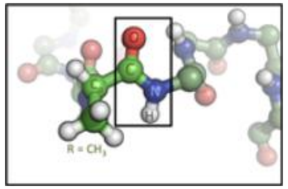
Finding evolutionary relationships between proteins, drawing proteins’ family trees

Computing the optimal alignment between two or more protein sequences

…



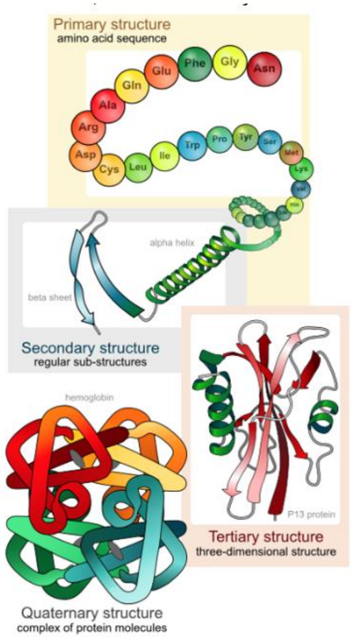
Proteins



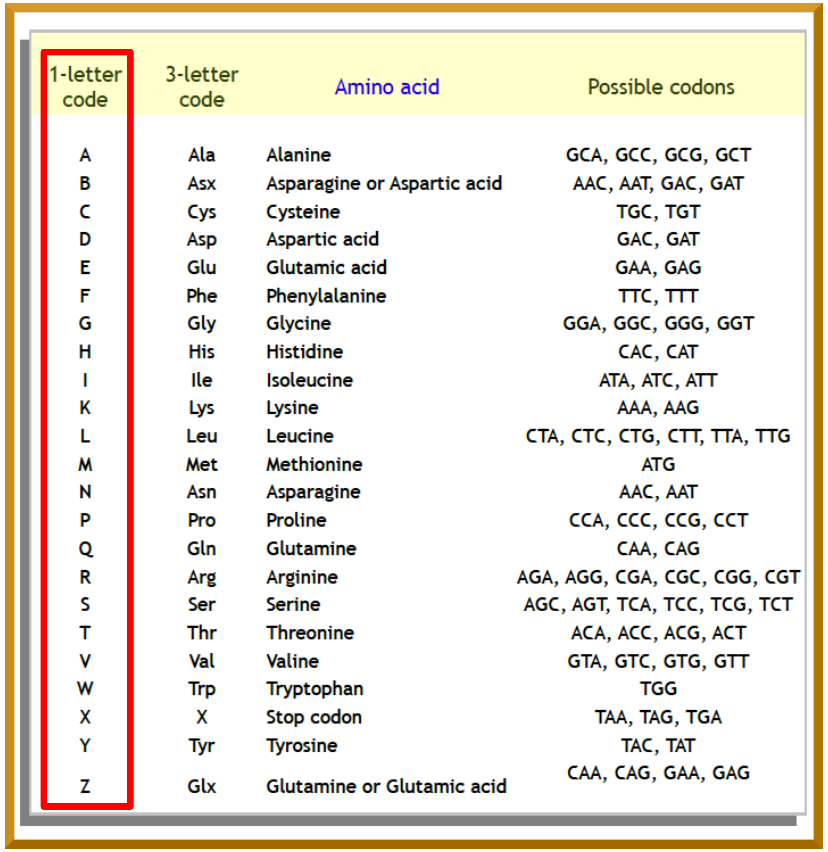
20 Aminoacids – primary structure:

(Frederick Sanger-1958 Nobel prize for insulin sequencing)

Secondary structure
Tertiary structure
Quaternary structure



SEQUENCE ⇨ STRUCTURE ⇨ FUNCTION

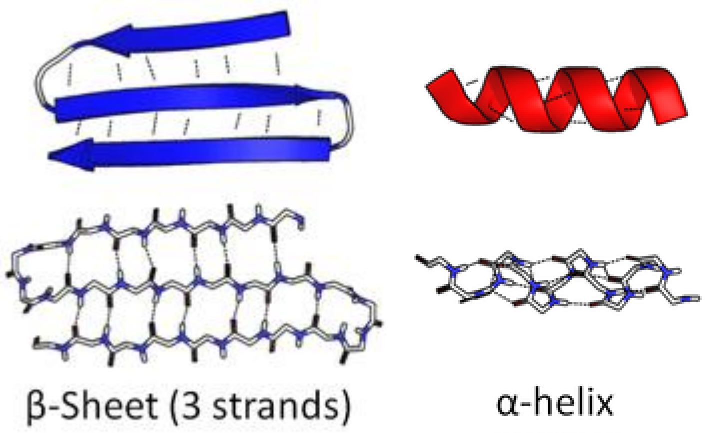


1-letter code	3-letter code	Amino acid	Possible codons
A	Ala	Alanine	GCA, GCC, GCG, GCT
B	Asx	Asparagine or Aspartic acid	AAC, AAT, GAC, GAT
C	Cys	Cysteine	TGC, TGT
D	Asp	Aspartic acid	GAC, GAT
E	Glu	Glutamic acid	GAA, GAG
F	Phe	Phenylalanine	TTC, TTT
G	Gly	Glycine	GGA, GGC, GGG, GGT
H	His	Histidine	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

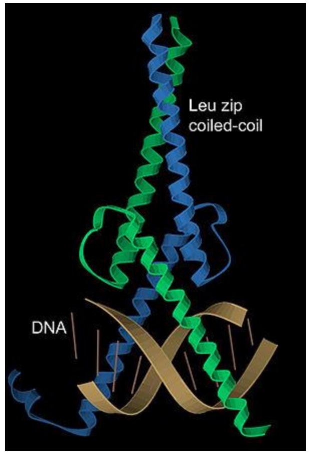
Protein domain

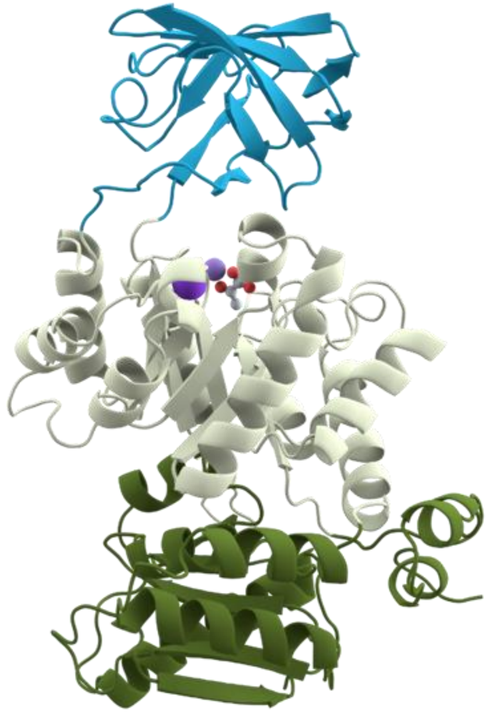
- region of a protein's polypeptide chain that folds independently from the rest
- forms a compact folded three-dimensional structure
- many proteins consist of several domains

Secondary



Tertiary



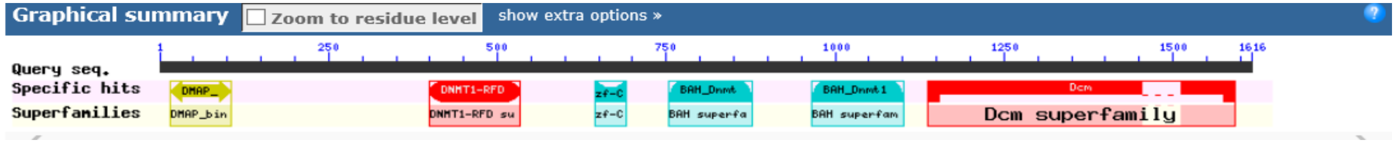


Conserved domain search

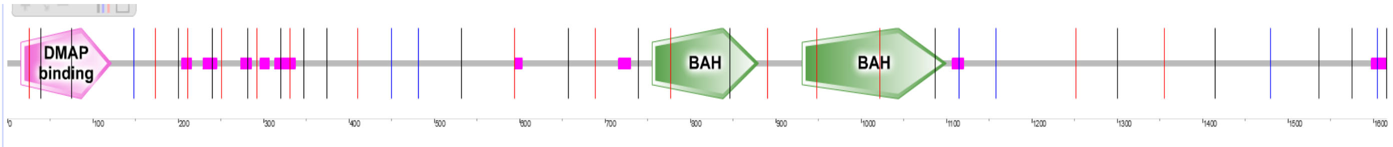
SEQUENCE ⇨ STRUCTURE ⇨ FUNCTION

- Conserved domain databases:

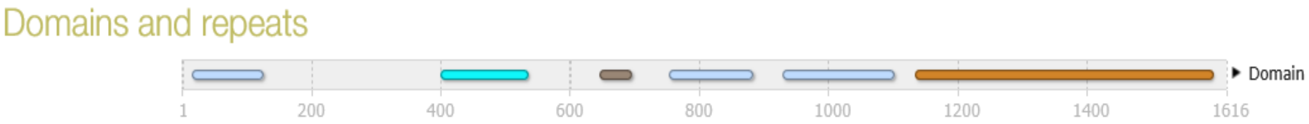
NCBI/CDD



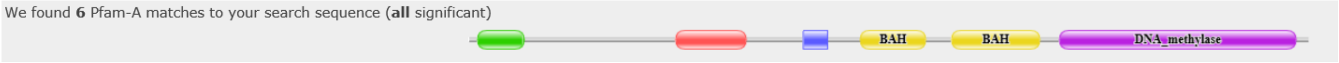
SMART



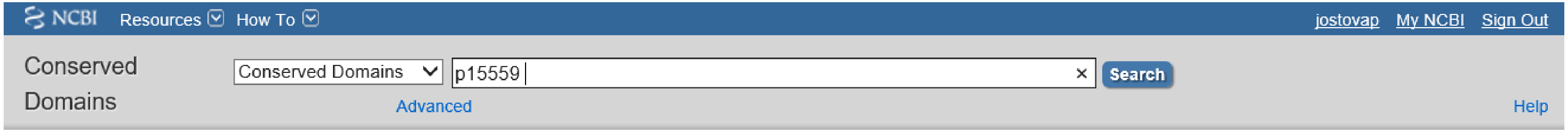
EMBL/InterPro

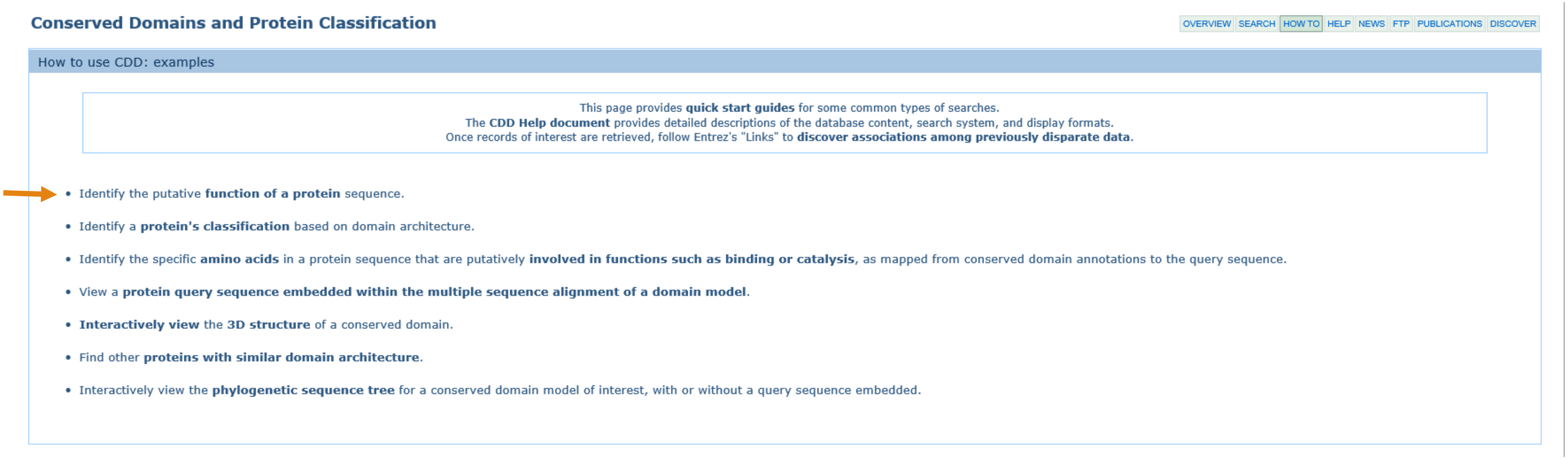


Pfam


We found **6** Pfam-A matches to your search sequence (**all** significant)

Conserved domain search - CD (NCBI)





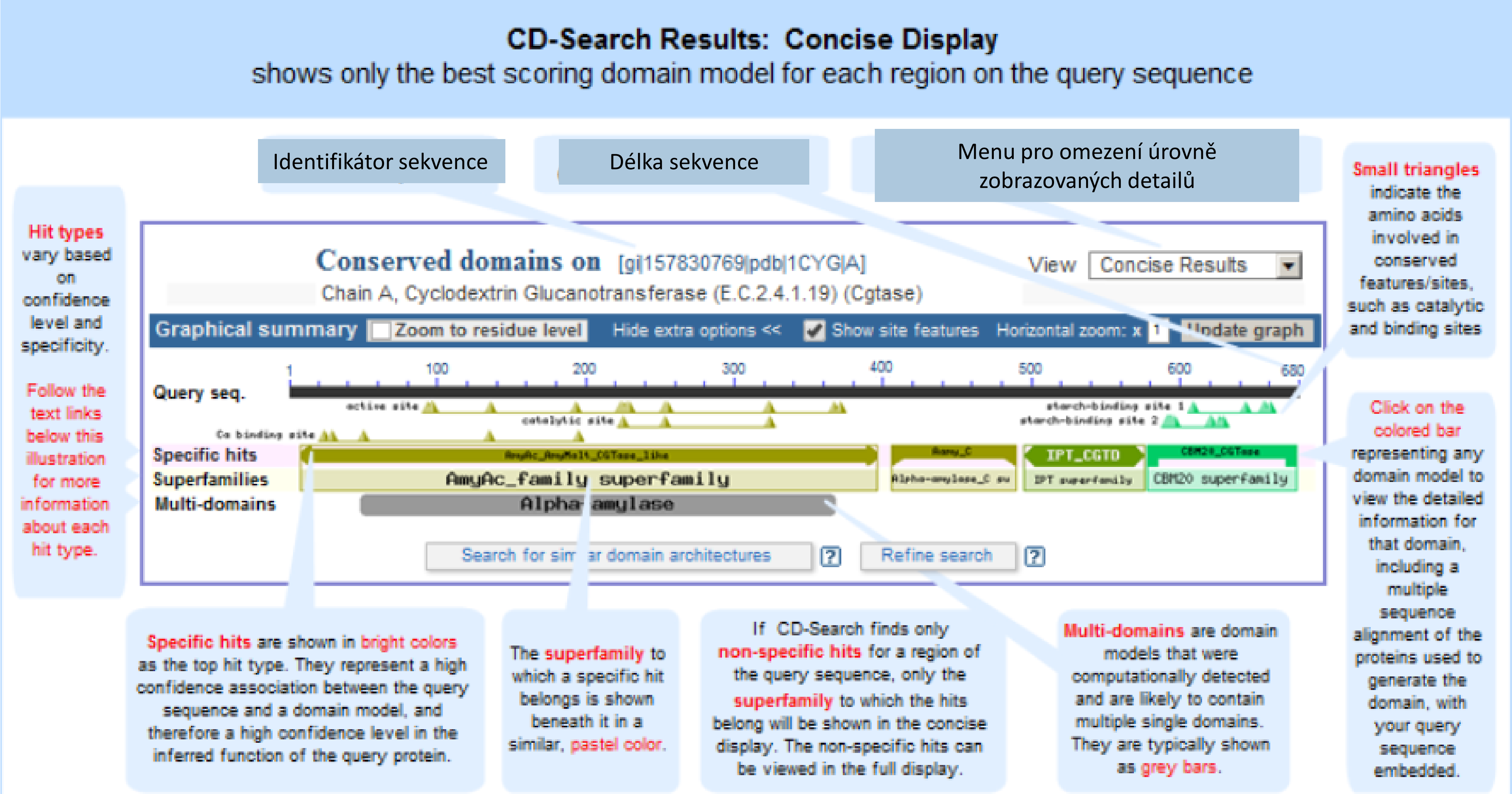
Conserved Domains and Protein Classification

How to use CDD: examples

This page provides **quick start guides** for some common types of searches.
The **CDD Help document** provides detailed descriptions of the database content, search system, and display formats.
Once records of interest are retrieved, follow Entrez's "Links" to **discover associations among previously disparate data**.

- Identify the putative **function of a protein** sequence.
- Identify a **protein's classification** based on domain architecture.
- Identify the specific **amino acids** in a protein sequence that are putatively **involved in functions such as binding or catalysis**, as mapped from conserved domain annotations to the query sequence.
- View a **protein query sequence embedded within the multiple sequence alignment of a domain model**.
- **Interactively view** the **3D structure** of a conserved domain.
- Find other **proteins with similar domain architecture**.
- Interactively view the **phylogenetic sequence tree** for a conserved domain model of interest, with or without a query sequence embedded.

Conserved domain search - CD (NCBI)



CD-Search Results: Concise Display

shows only the best scoring domain model for each region on the query sequence

Identifikátor sekvence

Délka sekvence

Menu pro omezení úrovně zobrazovaných detailů

Hit types vary based on confidence level and specificity.

Follow the text links below this illustration for more information about each hit type.

Small triangles indicate the amino acids involved in conserved features/sites, such as catalytic and binding sites

Click on the colored bar representing any domain model to view the detailed information for that domain, including a multiple sequence alignment of the proteins used to generate the domain, with your query sequence embedded.

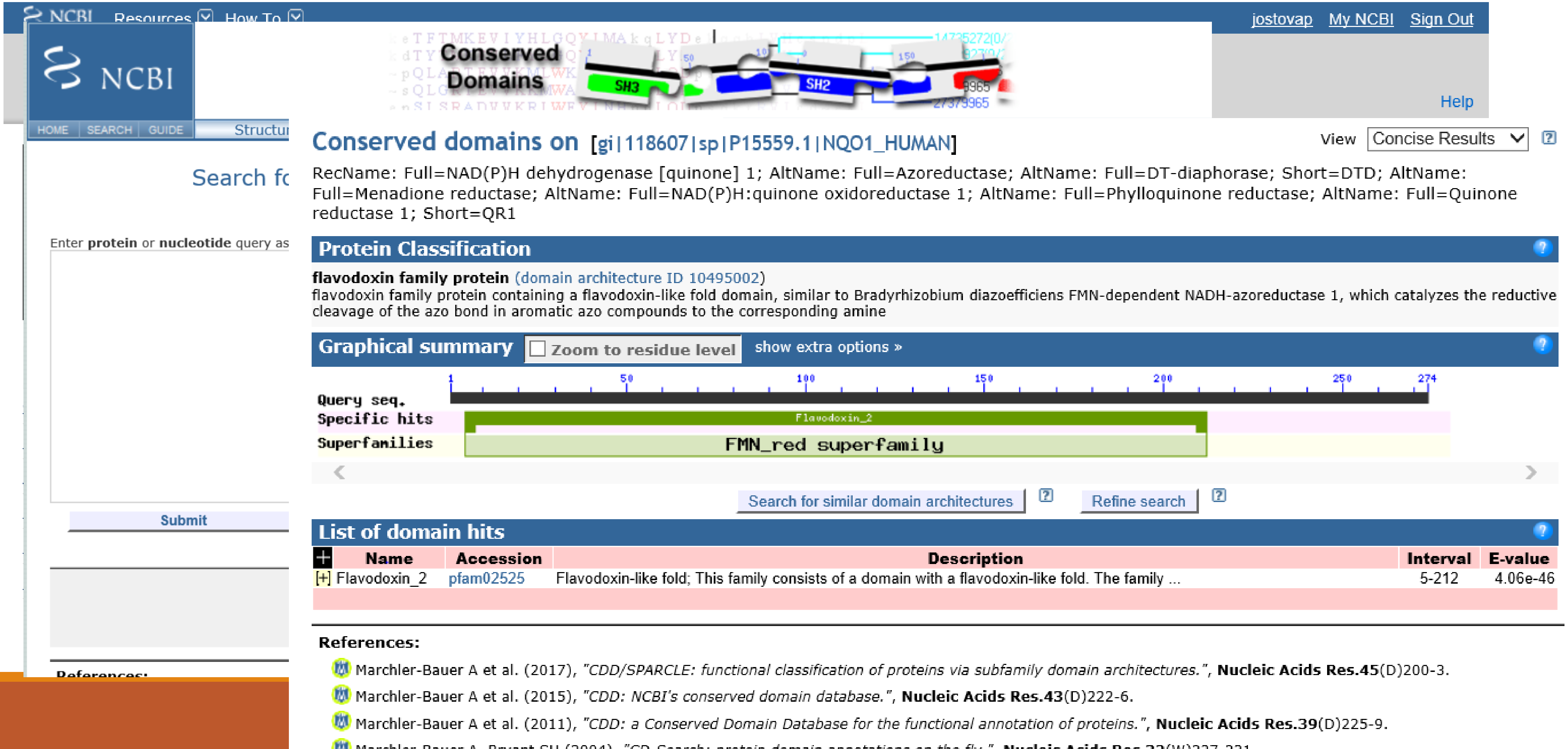
Specific hits are shown in bright colors as the top hit type. They represent a high confidence association between the query sequence and a domain model, and therefore a high confidence level in the inferred function of the query protein.

The **superfamily** to which a specific hit belongs is shown beneath it in a similar, pastel color.

If CD-Search finds only **non-specific hits** for a region of the query sequence, only the **superfamily** to which the hits belong will be shown in the concise display. The non-specific hits can be viewed in the full display.

Multi-domains are domain models that were computationally detected and are likely to contain multiple single domains. They are typically shown as grey bars.

Conserved domain search / NQO1



Enter **protein** or **nucleotide** query as

References:

Conserved domains on [gi|118607|sp|P15559.1|NQO1_HUMAN]

RecName: Full=NAD(P)H dehydrogenase [quinone] 1; AltName: Full=Azoreductase; AltName: Full=DT-diaphorase; Short=DTD; AltName: Full=Menadione reductase; AltName: Full=NAD(P)H:quinone oxidoreductase 1; AltName: Full=Phylloquinone reductase; AltName: Full=Quinone reductase 1; Short=QR1

Protein Classification

flavodoxin family protein (domain architecture ID 10495002)

flavodoxin family protein containing a flavodoxin-like fold domain, similar to Bradyrhizobium diazoefficiens FMN-dependent NADH-azoreductase 1, which catalyzes the reductive cleavage of the azo bond in aromatic azo compounds to the corresponding amine

Graphical summary

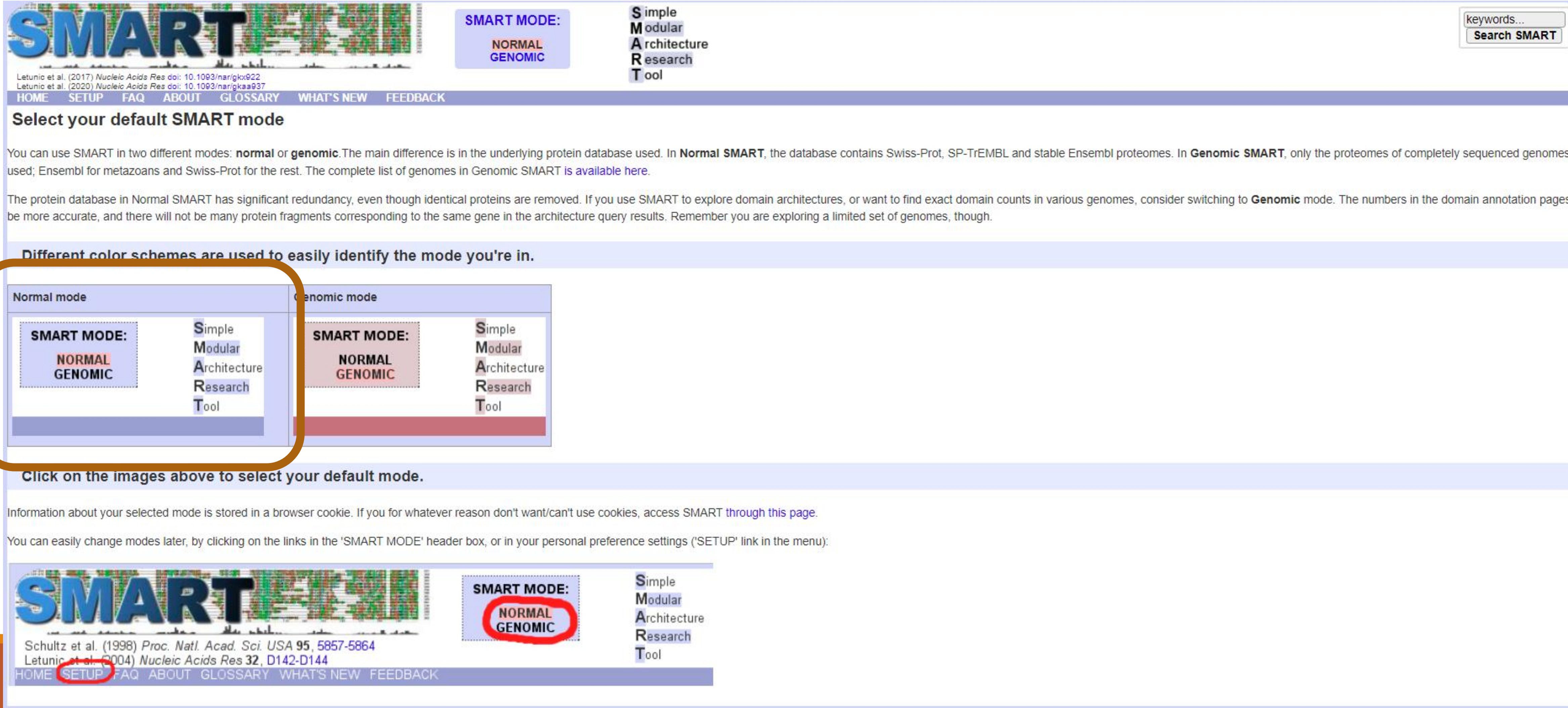
List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	Flavodoxin_2	pfam02525	Flavodoxin-like fold; This family consists of a domain with a flavodoxin-like fold. The family ...	5-212	4.06e-46

References:

- Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.
- Marchler-Bauer A et al. (2015), *"CDD: NCBI's conserved domain database."*, **Nucleic Acids Res.43**(D)222-6.
- Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.
- Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331

Conserved domain search - SMART



Letunic et al. (2017) *Nucleic Acids Res* doi: 10.1093/nar/gkx922
Letunic et al. (2020) *Nucleic Acids Res* doi: 10.1093/nar/gkaa937

Select your default SMART mode

You can use SMART in two different modes: **normal** or **genomic**.The main difference is in the underlying protein database used. In **Normal SMART**, the database contains Swiss-Prot, SP-TrEMBL and stable Ensembl proteomes. In **Genomic SMART**, only the proteomes of completely sequenced genomes used; Ensembl for metazoans and Swiss-Prot for the rest. The complete list of genomes in Genomic SMART is available here.

The protein database in Normal SMART has significant redundancy, even though identical proteins are removed. If you use SMART to explore domain architectures, or want to find exact domain counts in various genomes, consider switching to **Genomic** mode. The numbers in the domain annotation pages be more accurate, and there will not be many protein fragments corresponding to the same gene in the architecture query results. Remember you are exploring a limited set of genomes, though.

Different color schemes are used to easily identify the mode you're in.

Normal mode	Genomic mode
SMART MODE: NORMAL GENOMIC Simple Modular Architecture Research Tool	SMART MODE: NORMAL GENOMIC Simple Modular Architecture Research Tool

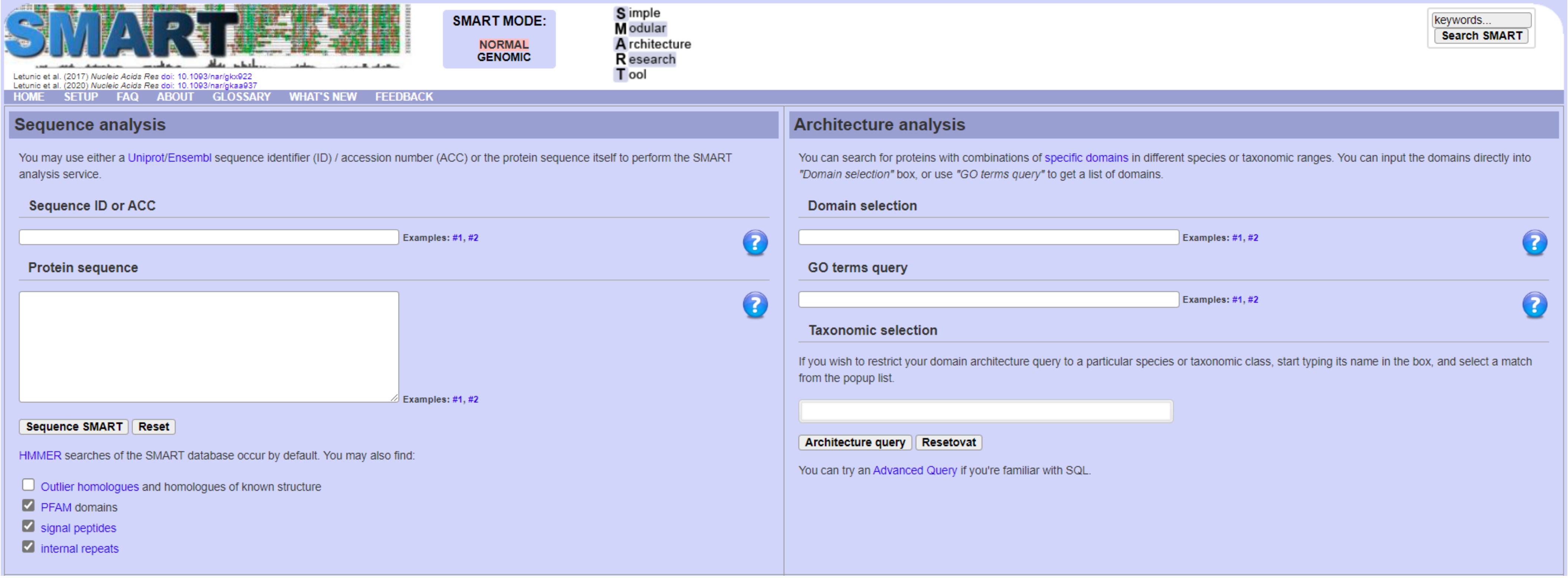
Click on the images above to select your default mode.

Information about your selected mode is stored in a browser cookie. If you for whatever reason don't want/can't use cookies, access SMART through this page.

You can easily change modes later, by clicking on the links in the 'SMART MODE' header box, or in your personal preference settings ('SETUP' link in the menu):

Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5857-5864
Letunic et al. (2004) *Nucleic Acids Res* **32**, D142-D144

Conserved domain search - SMART



Letunic et al. (2017) *Nucleic Acids Res* doi: 10.1093/nar/gkx922
Letunic et al. (2020) *Nucleic Acids Res* doi: 10.1093/nar/gkaa937

Sequence analysis

You may use either a Uniprot/Ensembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to perform the SMART analysis service.

Sequence ID or ACC

Examples: #1, #2

Protein sequence

Examples: #1, #2

HMMER searches of the SMART database occur by default. You may also find:

- □ Outlier homologues and homologues of known structure
- ☑ PFAM domains
- ☑ signal peptides
- ☑ internal repeats

Architecture analysis

You can search for proteins with combinations of specific domains in different species or taxonomic ranges. You can input the domains directly into *"Domain selection"* box, or use *"GO terms query"* to get a list of domains.

Domain selection

Examples: #1, #2

GO terms query

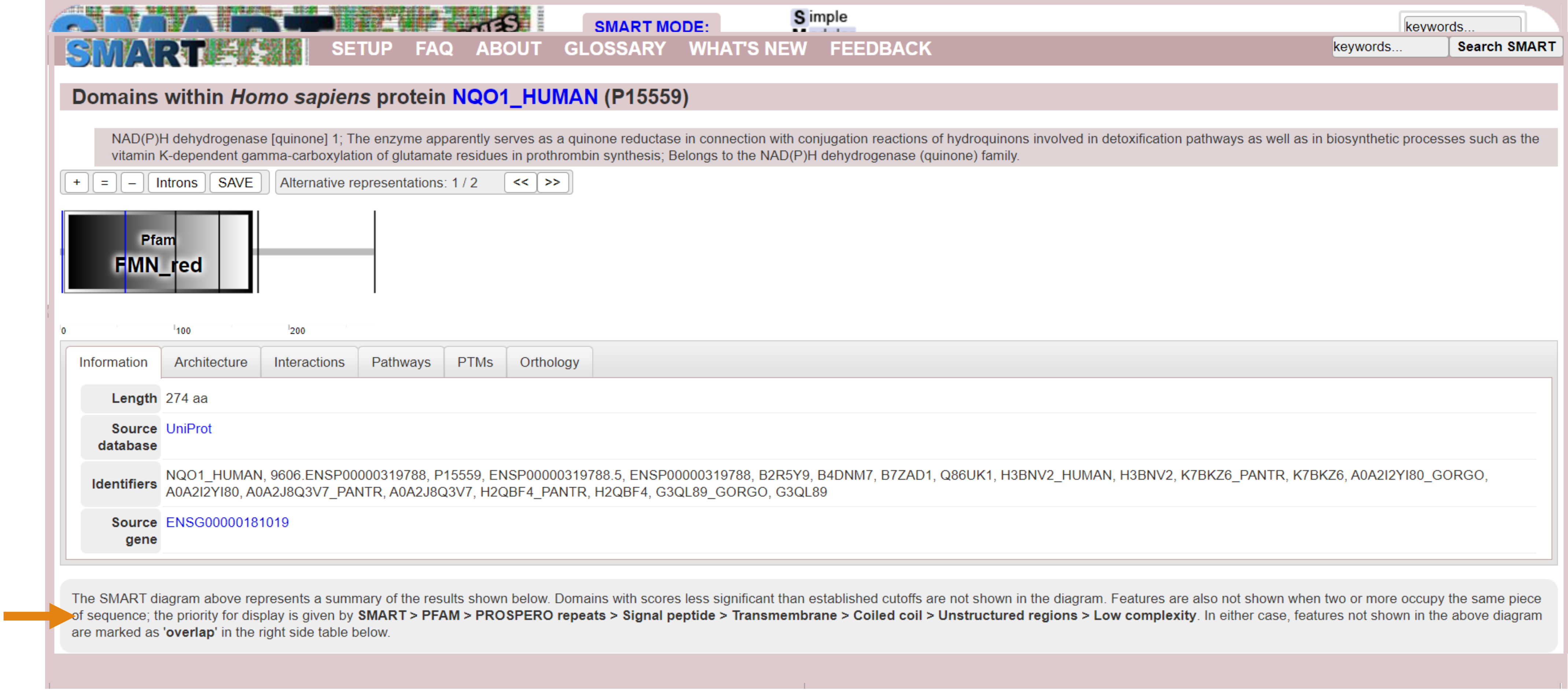
Examples: #1, #2

Taxonomic selection

If you wish to restrict your domain architecture query to a particular species or taxonomic class, start typing its name in the box, and select a match from the popup list.

You can try an Advanced Query if you're familiar with SQL.

Conserved domain search - SMART



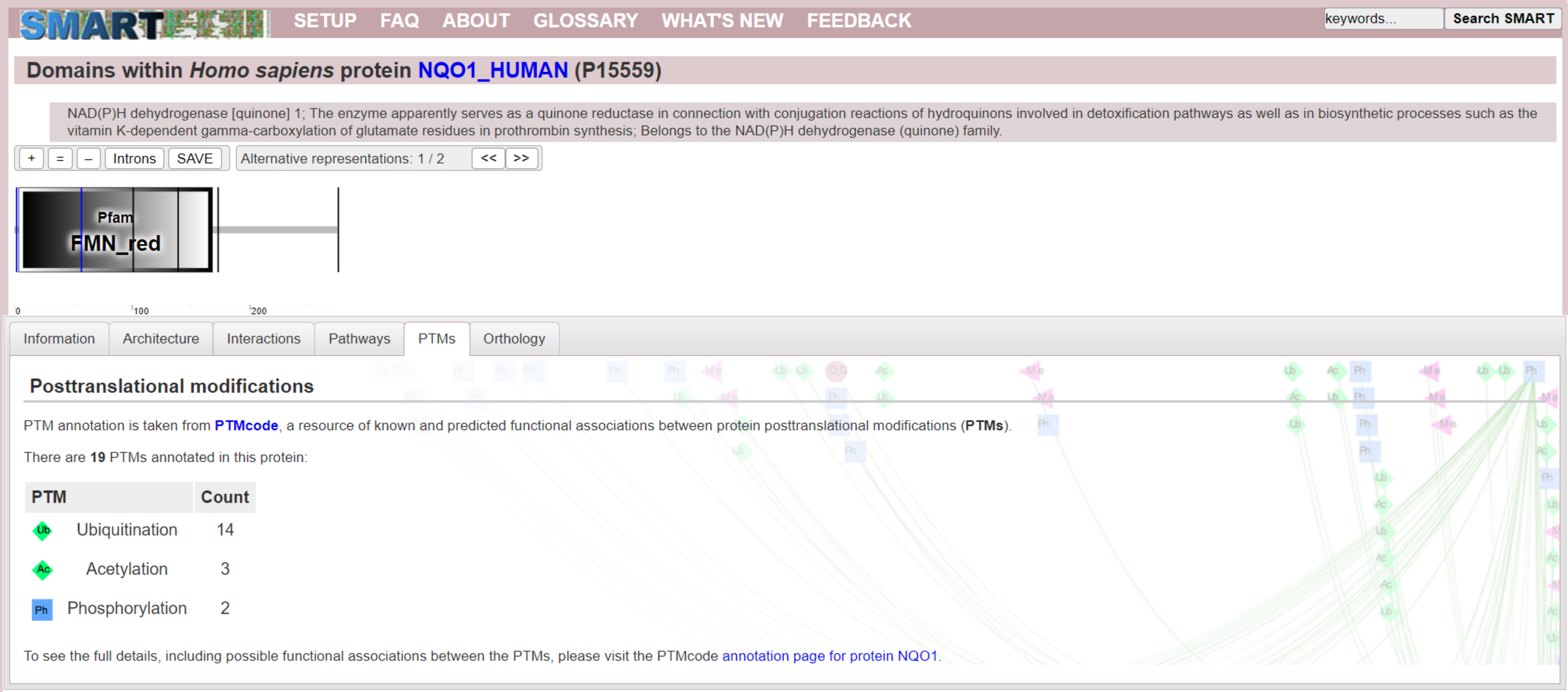
Domains within *Homo sapiens* protein NQO1_HUMAN (P15559)

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.

Length	274 aa
Source database	UniProt
Identifiers	NQO1_HUMAN, 9606.ENSP00000319788, P15559, ENSP00000319788.5, ENSP00000319788, B2R5Y9, B4DNM7, B7ZAD1, Q86UK1, H3BNV2_HUMAN, H3BNV2, K7BKZ6_PANTR, K7BKZ6, A0A2I2YI80_GORGO, A0A2I2YI80, A0A2J8Q3V7_PANTR, A0A2J8Q3V7, H2QBF4_PANTR, H2QBF4, G3QL89_GORGO, G3QL89
Source gene	ENSG00000181019

The SMART diagram above represents a summary of the results shown below. Domains with scores less significant than established cutoffs are not shown in the diagram. Features are also not shown when two or more occupy the same piece of sequence; the priority for display is given by **SMART > PFAM > PROSPERO repeats > Signal peptide > Transmembrane > Coiled coil > Unstructured regions > Low complexity**. In either case, features not shown in the above diagram are marked as '**overlap**' in the right side table below.

Conserved domain search - SMART



Domains within *Homo sapiens* protein NQO1_HUMAN (P15559)

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.

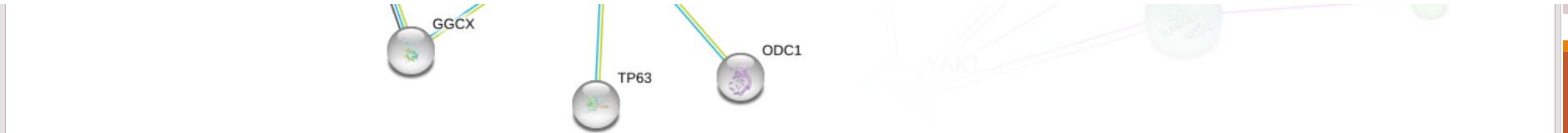
Posttranslational modifications

PTM annotation is taken from PTMcode, a resource of known and predicted functional associations between protein posttranslational modifications (**PTMs**).

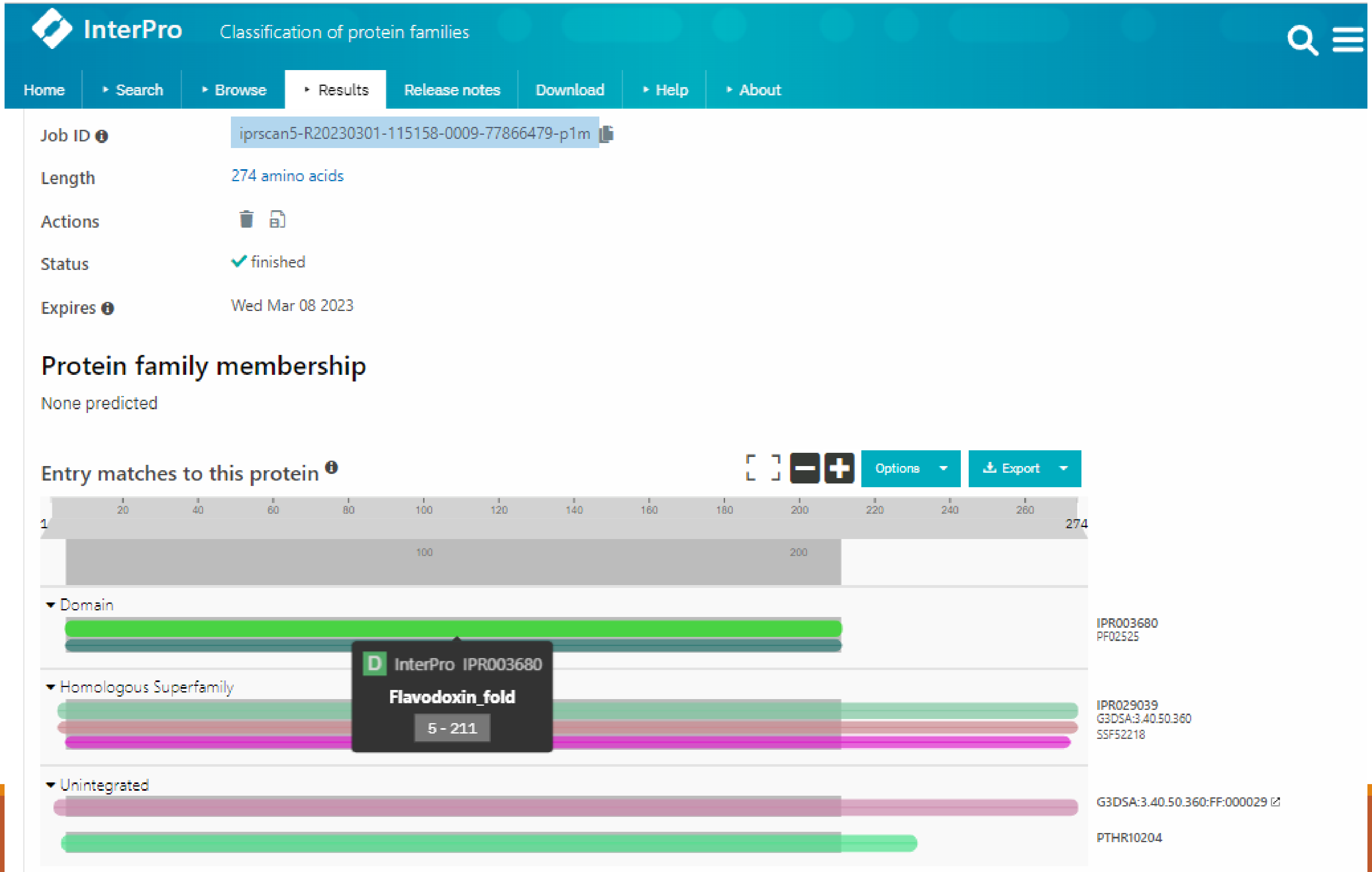
There are **19** PTMs annotated in this protein:

PTM	Count
Ub Ubiquitination	14
Ac Acetylation	3
Ph Phosphorylation	2

To see the full details, including possible functional associations between the PTMs, please visit the PTMcode annotation page for protein NQO1.



Conserved domain search - InterPro



Job ID iprscan5-R20230301-115158-0009-77866479-p1m

Length 274 amino acids

Actions

Status ✓ finished

Expires Wed Mar 08 2023

Protein family membership

None predicted

Entry matches to this protein

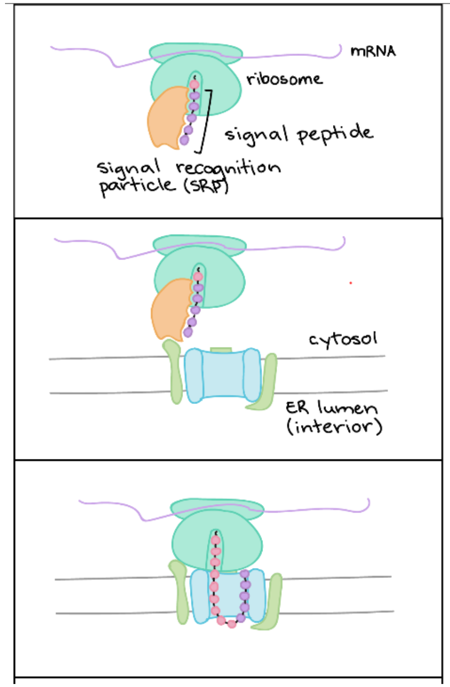
Practical part

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



Signal recognition particle (SRP) binds to the signal peptide as it emerges from the ribosome.

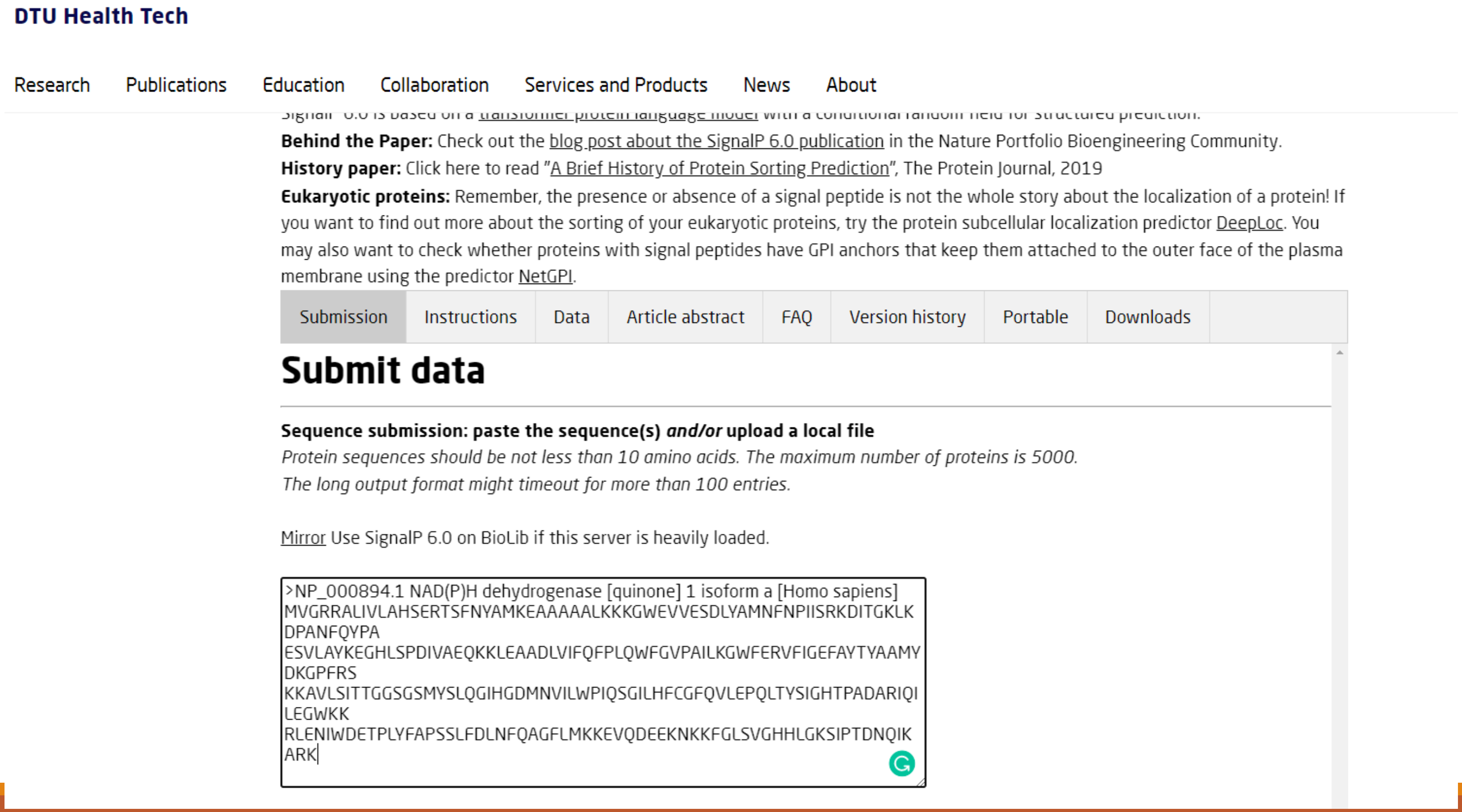
SRP brings the ribosome to the ER by binding to a receptor on the ER surface. The receptor is associated with other proteins that make a pore.

The ribosome resumes translating, feeding the polypeptide though the pore and into the ER lumen (interior).

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

Signal peptides

SignalP



SignalP 6.0 is based on a transformer protein language model with a conditional random field for structured prediction.

Behind the Paper: Check out the blog post about the SignalP 6.0 publication in the Nature Portfolio Bioengineering Community.

History paper: Click here to read "A Brief History of Protein Sorting Prediction", The Protein Journal, 2019

Eukaryotic proteins: Remember, the presence or absence of a signal peptide is not the whole story about the localization of a protein! If you want to find out more about the sorting of your eukaryotic proteins, try the protein subcellular localization predictor DeepLoc. You may also want to check whether proteins with signal peptides have GPI anchors that keep them attached to the outer face of the plasma membrane using the predictor NetGPI.

Submit data

Sequence submission: paste the sequence(s) *and/or* upload a local file

Protein sequences should be not less than 10 amino acids. The maximum number of proteins is 5000.

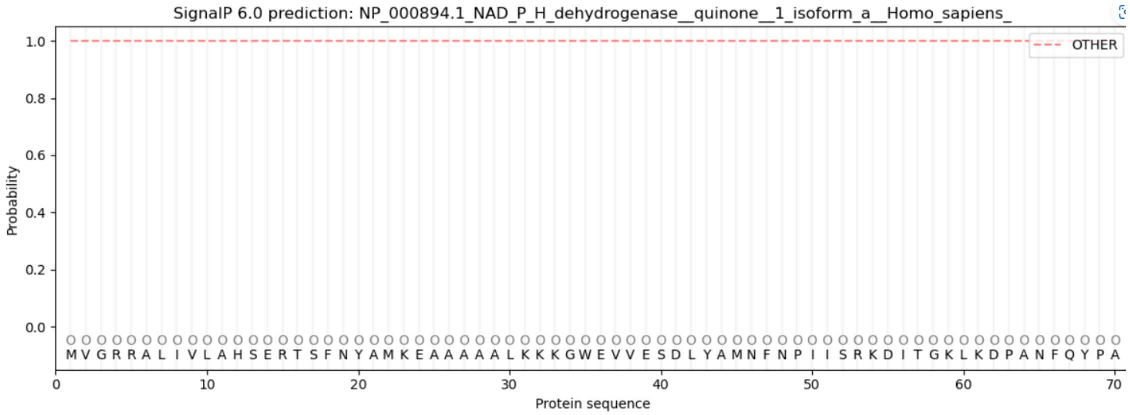
The long output format might timeout for more than 100 entries.

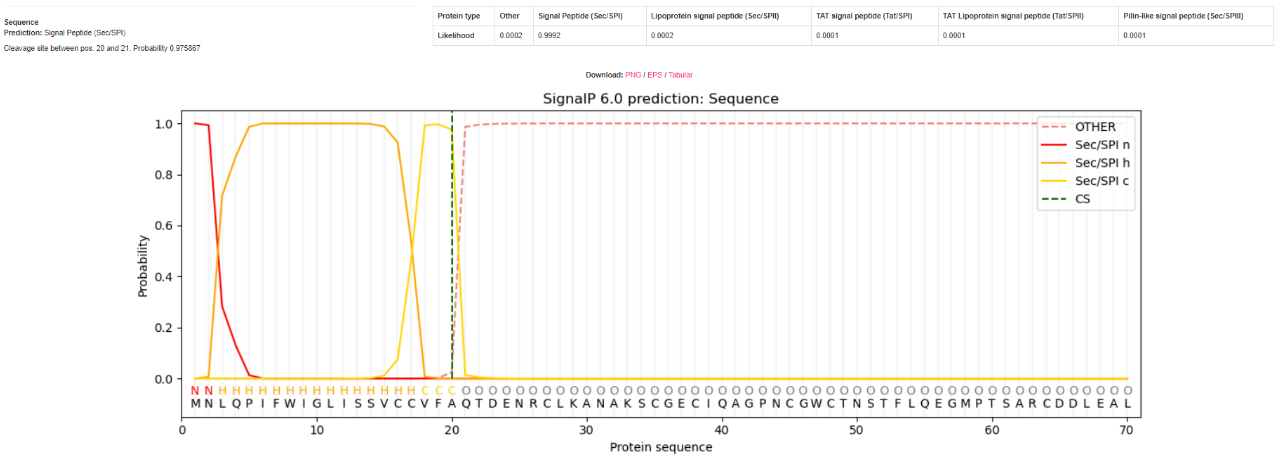
Mirror Use SignalP 6.0 on BioLib if this server is heavily loaded.

```
>NP_000894.1 NAD(P)H dehydrogenase [quinone] 1 isoform a [Homo sapiens]
MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLK
DPANFQYPA
ESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMY
DKGPFRS
KKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHFCGFQVLEPQLTYSIGHTPADARIQI
LEGWKK
RLENIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEKNKKFGLSVGHHLGKSIPTDNQIK
ARK
```

Signal peptides

SignalP





Sequence
Prediction: Signal Peptide (Sec/SPI)

Cleavage site between pos. 20 and 21. Probability 0.975867

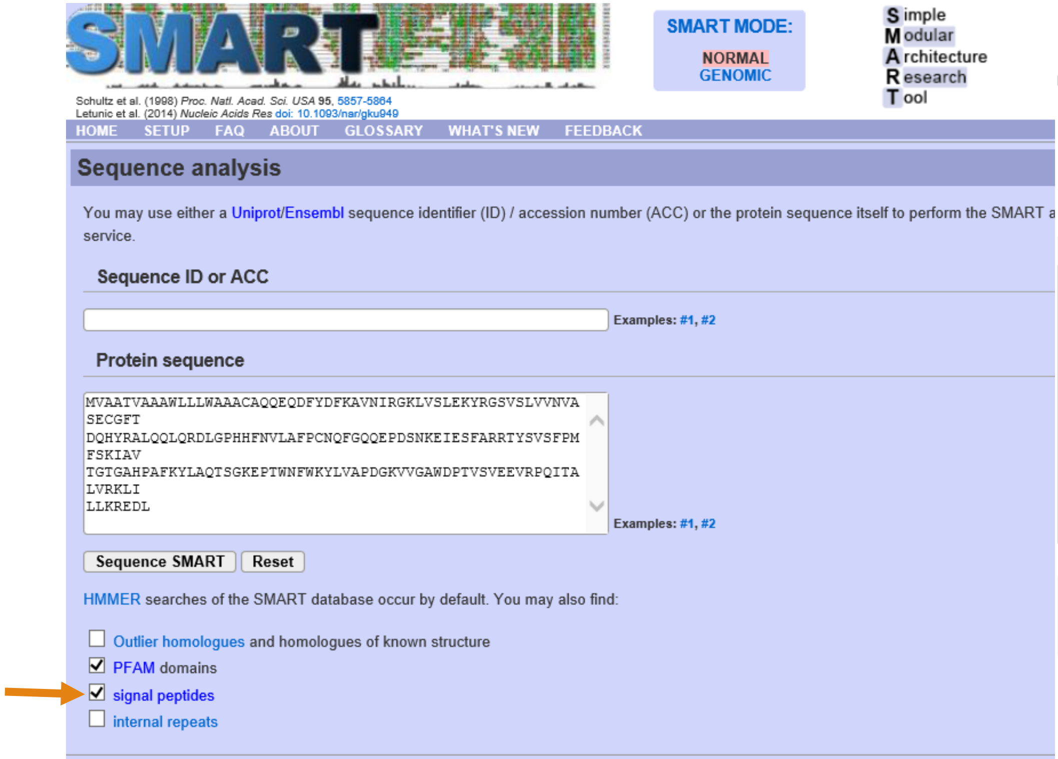
Protein type	Other	Signal Peptide (Sec/SPI)	Lipoprotein signal peptide (Sec/SPII)	TAT signal peptide (Tat/SPI)	TAT Lipoprotein signal peptide (Tat/SPII)	Pilin-like signal peptide (Sec/SPIII)
Likelihood	0.0002	0.9992	0.0002	0.0001	0.0001	0.0001

Download: PNG / EPS / Tabular
SignalP 6.0 prediction: Sequence
OTHER
Sec/SPI n
Sec/SPI h
Sec/SPI c
CS
Probability
1.0
0.8
0.6
0.4
0.2
0.0
N N H H H H H H H H H H H H H H H C C C O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O O
M N L Q P I F W I G L I S S V C C V F A Q T D E N R C L K A N A K S C G E C I Q A G P N C G W C T N S T F L Q E G M P T S A R C D D L E A L
0
10
20
30
40
50
60
70
Protein sequence

protein does not have signal peptide

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

Signal peptides



Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5857-5864
Letunic et al. (2014) *Nucleic Acids Res* doi: 10.1093/nar/gku949

Sequence analysis

You may use either a Uniprot/Ensembl sequence identifier (ID) / accession number (ACC) or the protein sequence itself to perform the SMART a
service.

Sequence ID or ACC

Examples: #1, #2

Protein sequence

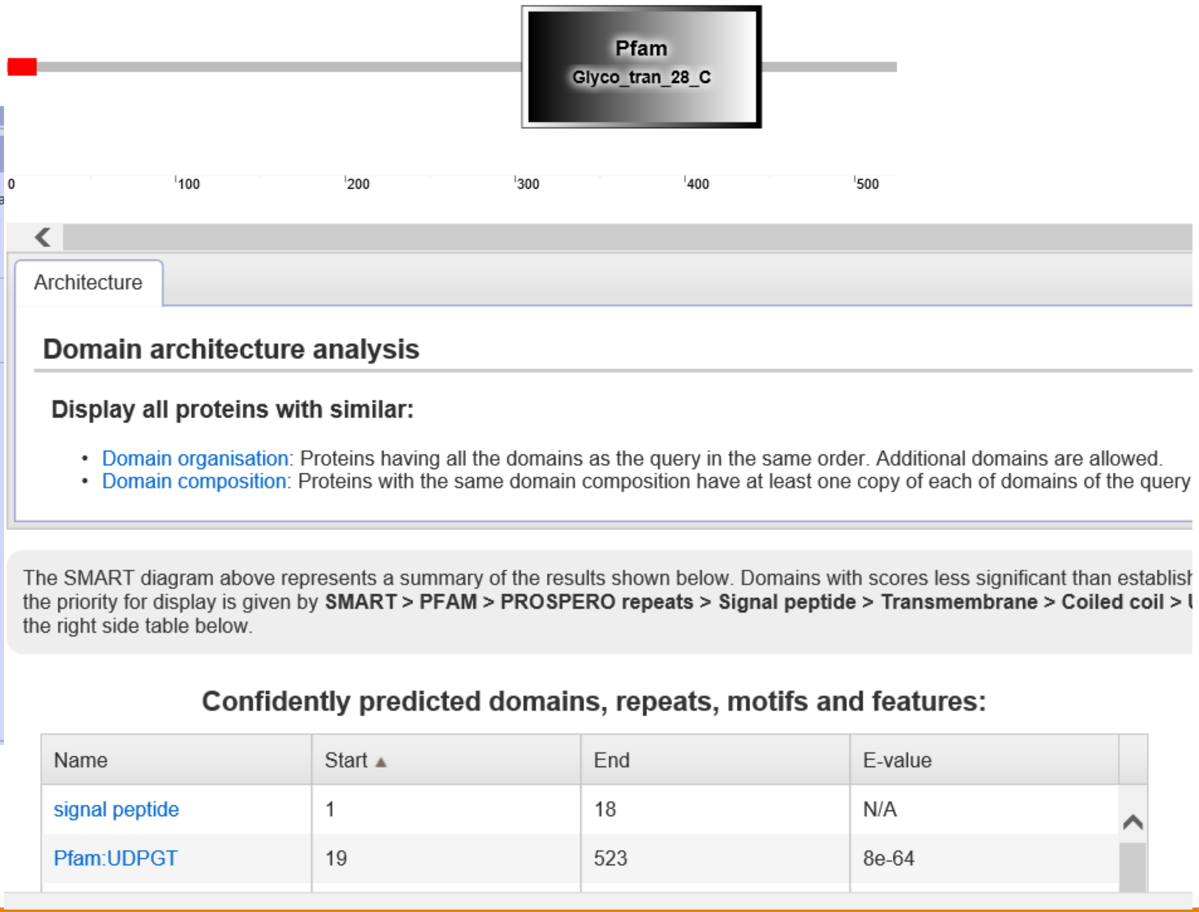
```
MVAATVAAAWLLLWAAACAQQEQDFYDFKAVNIRGKLVSLEKYRGSVSLVVNVA
SECGFT
DQHYRALQQLQRDLGPHHFNVLAFPCNQFGQQEPDSNKEIESFARRTYSVSFPM
FSKIAV
TGTGAHPAFKYLAQTSGKEPTWNFWKYLVAPDGKVVGAWDPTVSVEEVRPQITA
LVRKLI
LLKREDL
```

Examples: #1, #2

Sequence SMART | Reset

HMMER searches of the SMART database occur by default. You may also find:

- ☐ Outlier homologues and homologues of known structure
- ☑ PFAM domains
- ☑ signal peptides
- ☐ internal repeats



Domain architecture analysis

Display all proteins with similar:

- Domain organisation: Proteins having all the domains as the query in the same order. Additional domains are allowed.
- Domain composition: Proteins with the same domain composition have at least one copy of each of domains of the query

The SMART diagram above represents a summary of the results shown below. Domains with scores less significant than establish
the priority for display is given by **SMART > PFAM > PROSPERO repeats > Signal peptide > Transmembrane > Coiled coil > U**
the right side table below.

Confidently predicted domains, repeats, motifs and features:

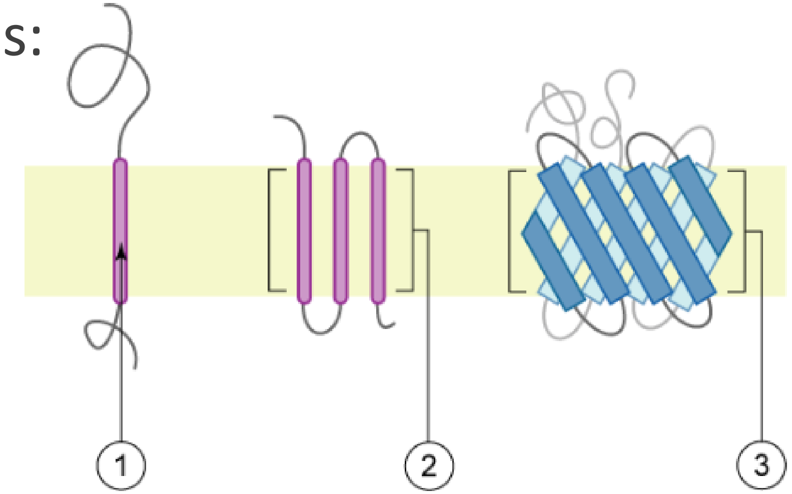
Name	Start ▲	End	E-value
signal peptide	1	18	N/A
Pfam:UDPGT	19	523	8e-64

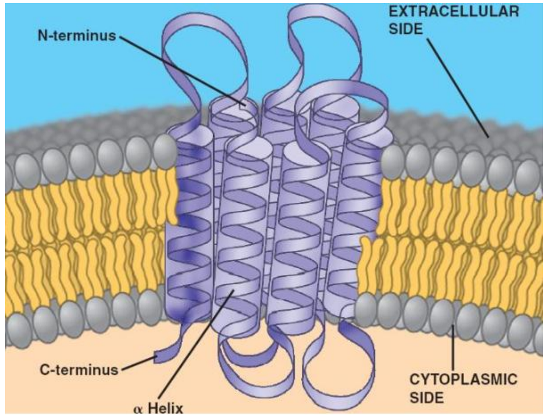
Practical part

search for signal peptide in your sequnce

Prediction of transmembrane helices

Transmembrane proteins:



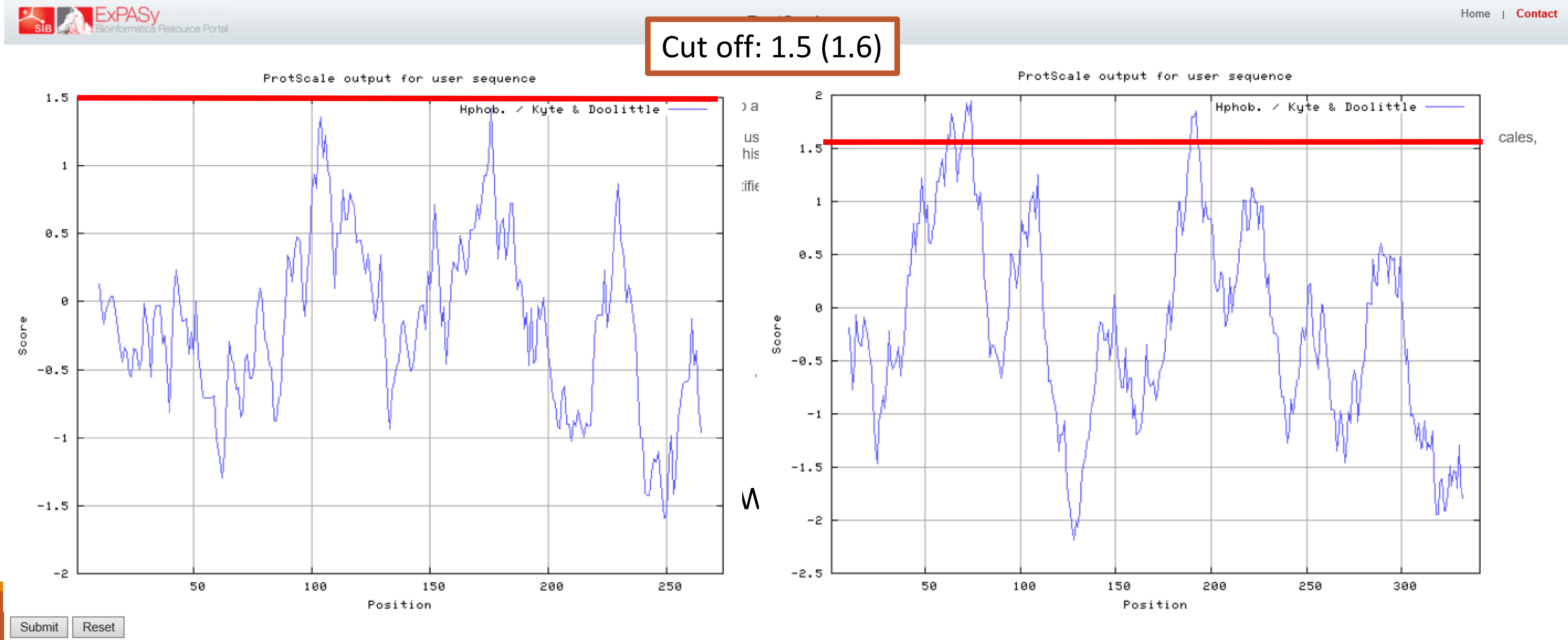


Amino acid Hydrofobicity

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

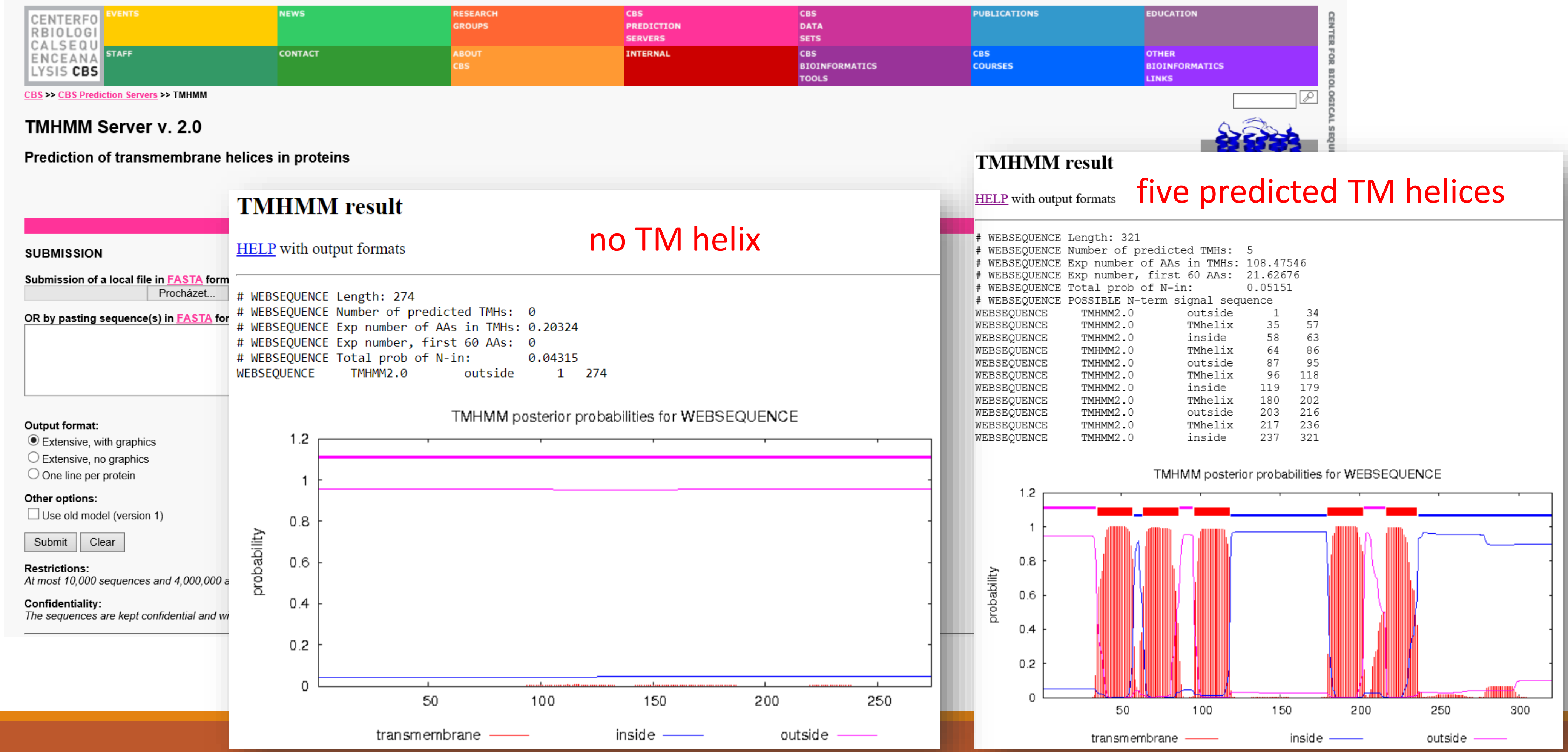
Prediction of transmembrane helices

Profile of amino acids hydrofobicity



Cut off: 1.5 (1.6)

TMHMM



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
```

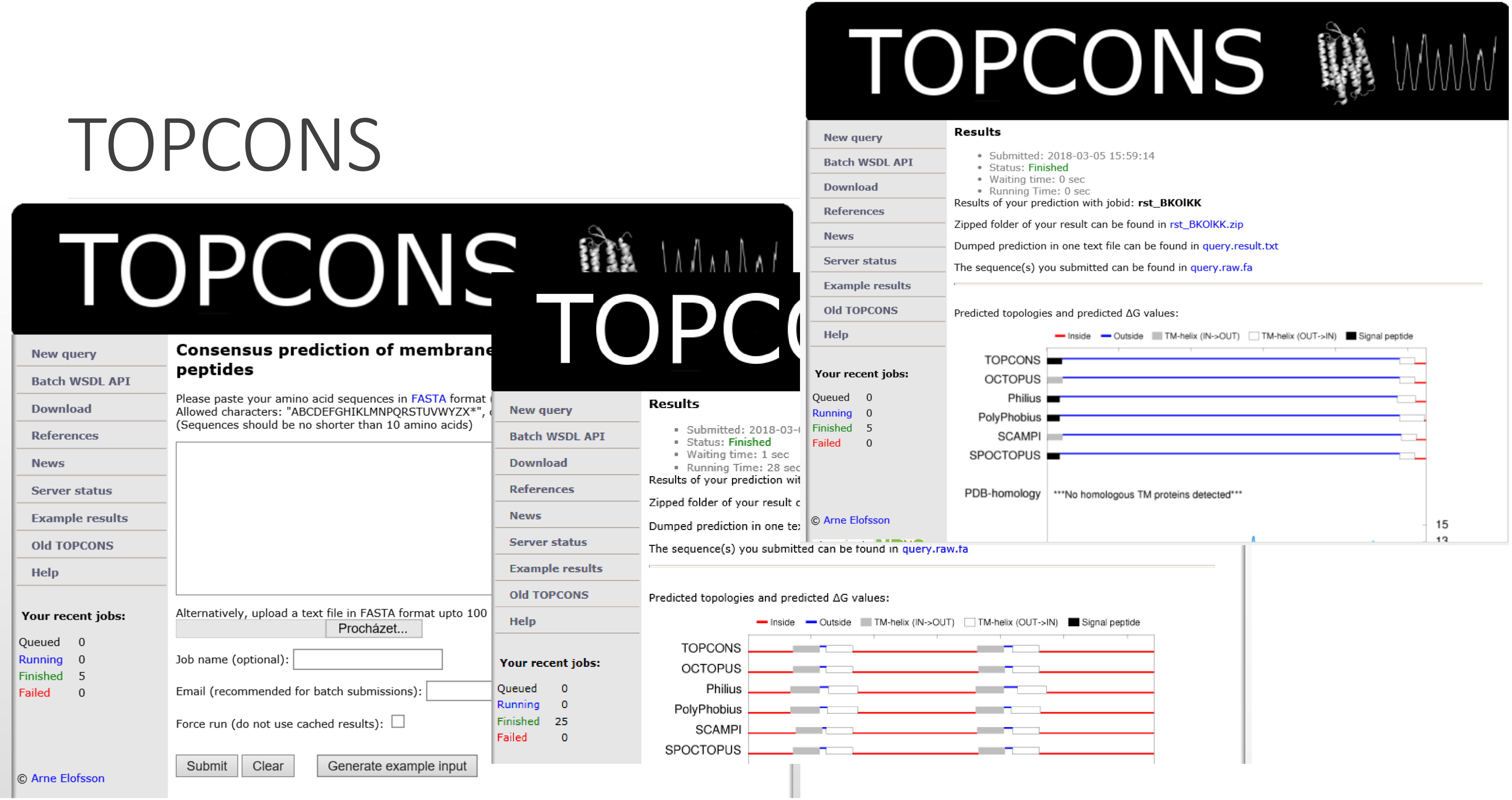
TMHMM result

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
```

TOPCONS



TOPCONS

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

Job name (optional):

Email (recommended for batch submissions):

Force run (do not use cached results): □

Your recent jobs:

Queued 0
Running 0
Finished 5
Failed 0

© Arne Elofsson

TOPCO

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:

Your recent jobs:

Queued 0
Running 0
Finished 25
Failed 0

TOPCONS

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:

Your recent jobs:

Queued 0
Running 0
Finished 5
Failed 0

© Arne Elofsson

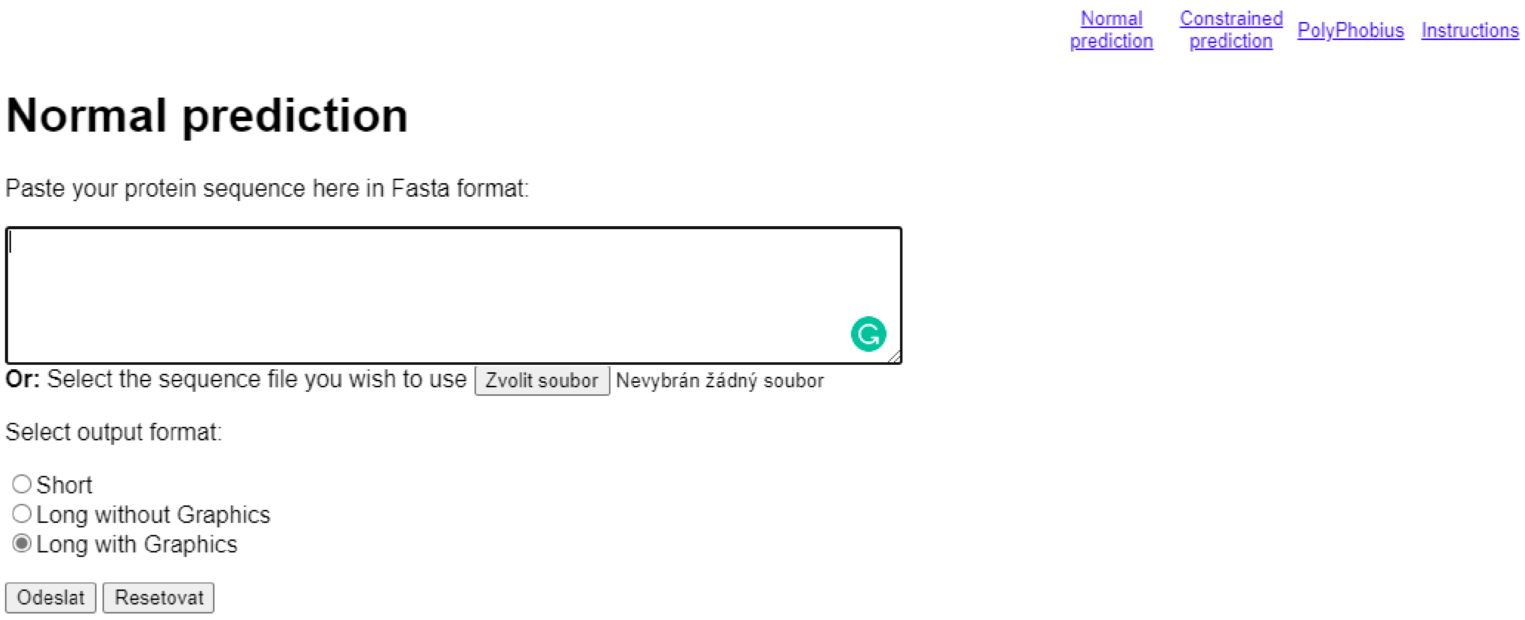
Phobius

Stockholm Bioinformatics Centre

SBC

Phobius

A combined transmembrane topology and signal peptide predictor



Normal prediction Constrained prediction PolyPhobius Instructions

Normal prediction

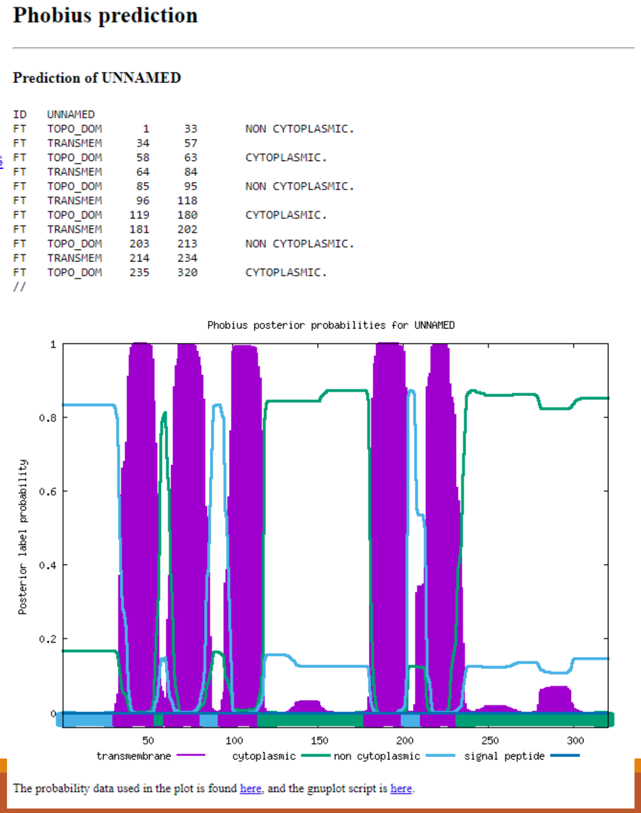
Paste your protein sequence here in Fasta format:

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



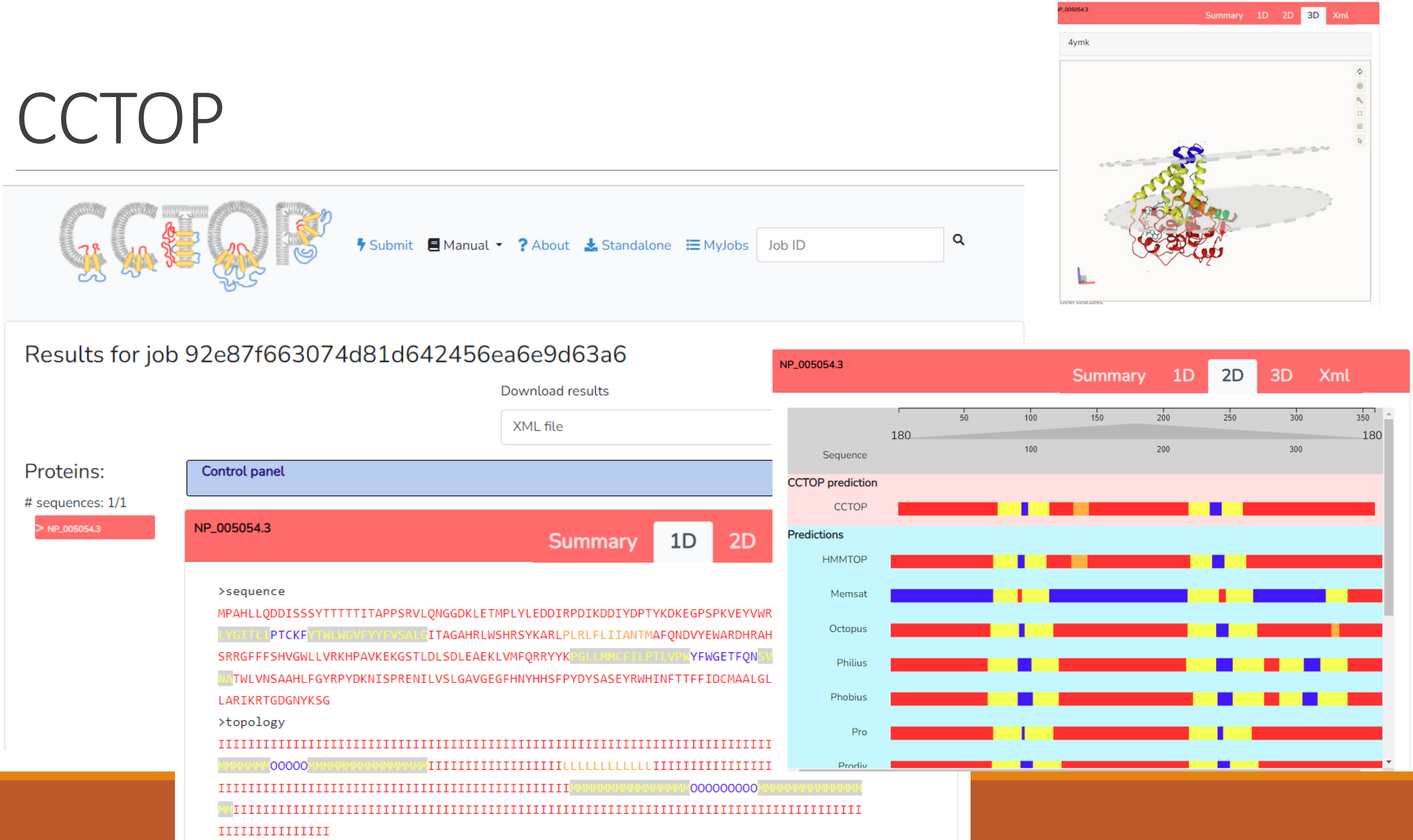
Phobius prediction

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



Results for job 92e87f663074d81d642456ea6e9d63a6

Download results

XML file

Proteins:

sequences: 1/1

> NP_005054.3

Control panel

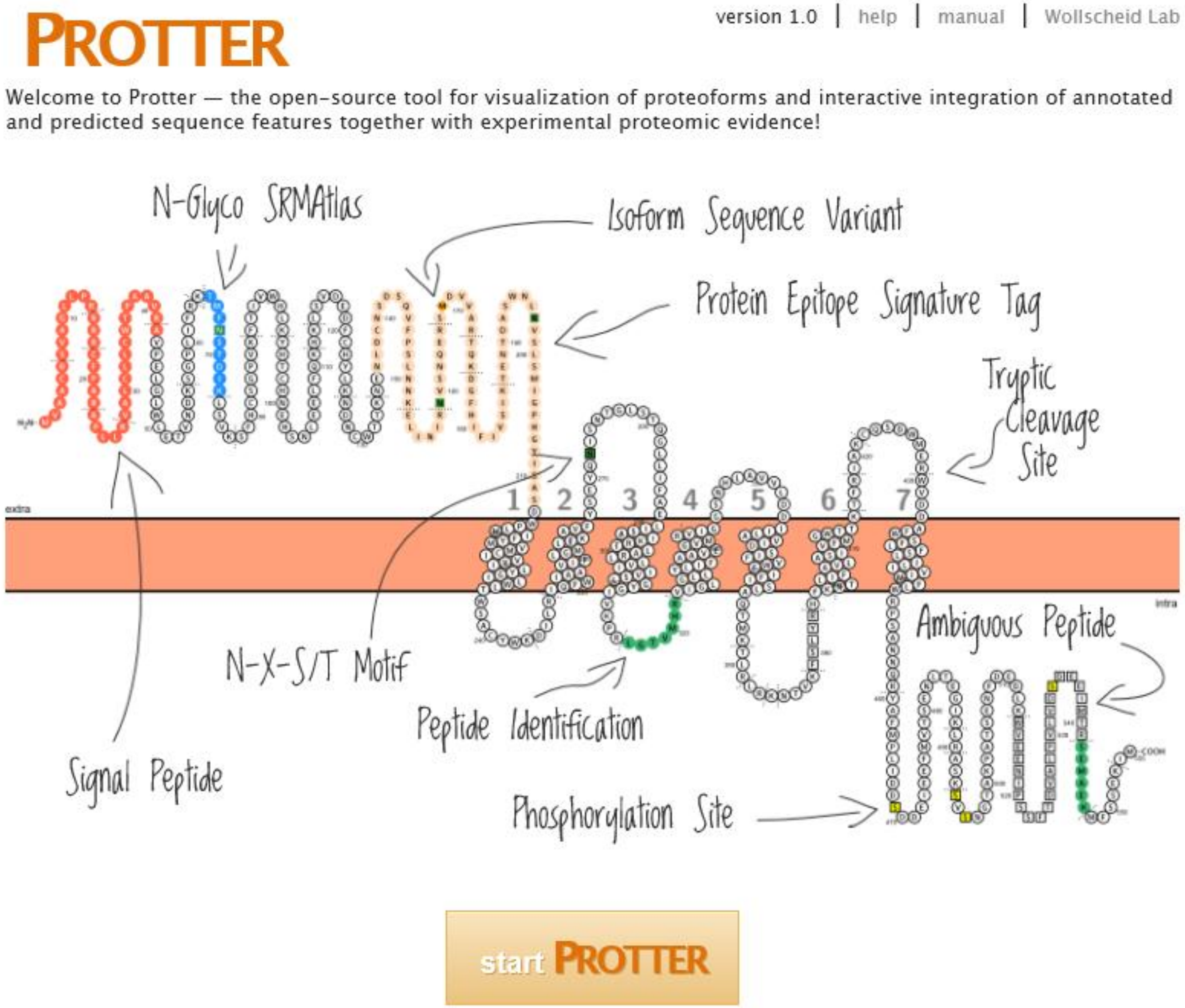
NP_005054.3

Summary 1D 2D

```
>sequence
MPAHLLQDDISSSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSPKVEYVWR
LYGITLIPTCKFYTWLWGVFYYFVSALGITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAH
SRRGFFFSHVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGETFQNSV
NATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINFTTFFIDCMAALGL
LARIKRTGDGNYKSG
>topology
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
MMMMMMMOOOOOMMMMMMMMMMMMMMMMIIIIIIIIIIIIIIIIIILLLLLLLLLLLLIIIIIIIIIIIIIIII
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIMMMMMMMMMMMMMMMMOOOOOOOOOMMMMMMMMMMMMMM
MMIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
IIIIIIIIIIIIIII
```

PROTTER-figure!

-creates figure from the UniProt data



Welcome to Protter — the open-source tool for visualization of proteoforms and interactive integration of annotated and predicted sequence features together with experimental proteomic evidence!

PROTTER-figure!

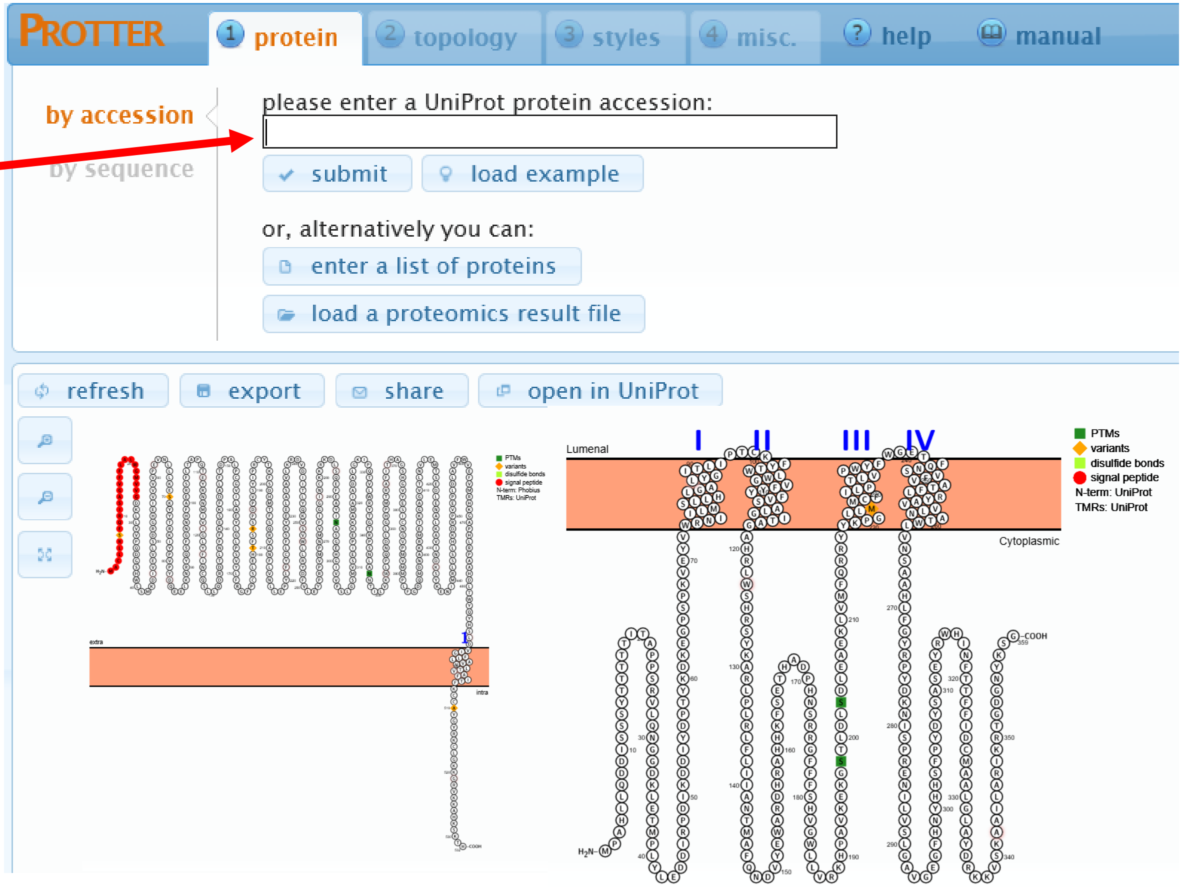
-creates figure from the UniProt data

-uses UniProt ID:

UGT1A6 (P19224)

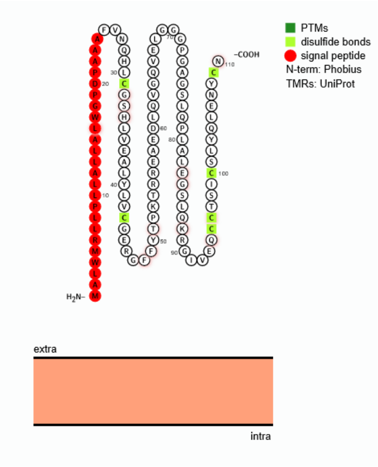
Desaturase (O00767)

(prepro) insulin (P01308)

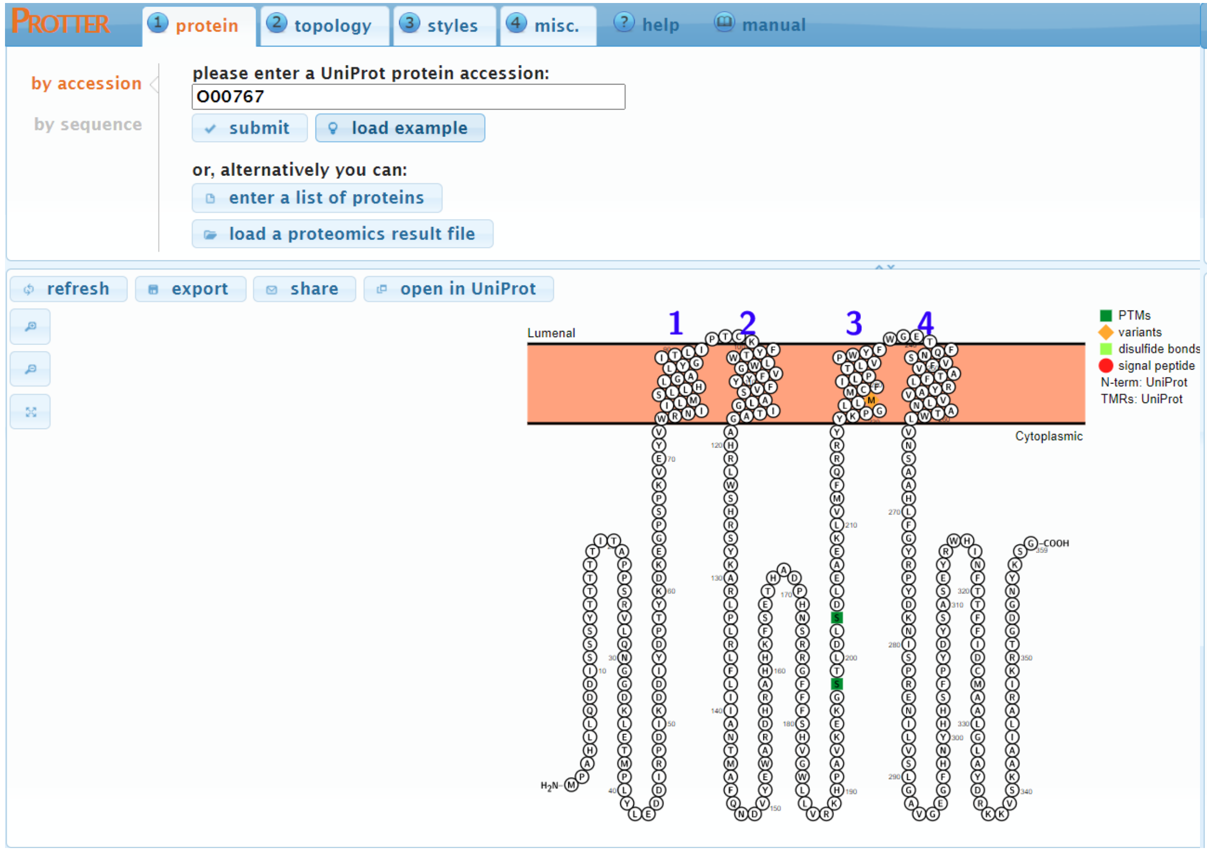


please enter a UniProt protein accession:

or, alternatively you can:



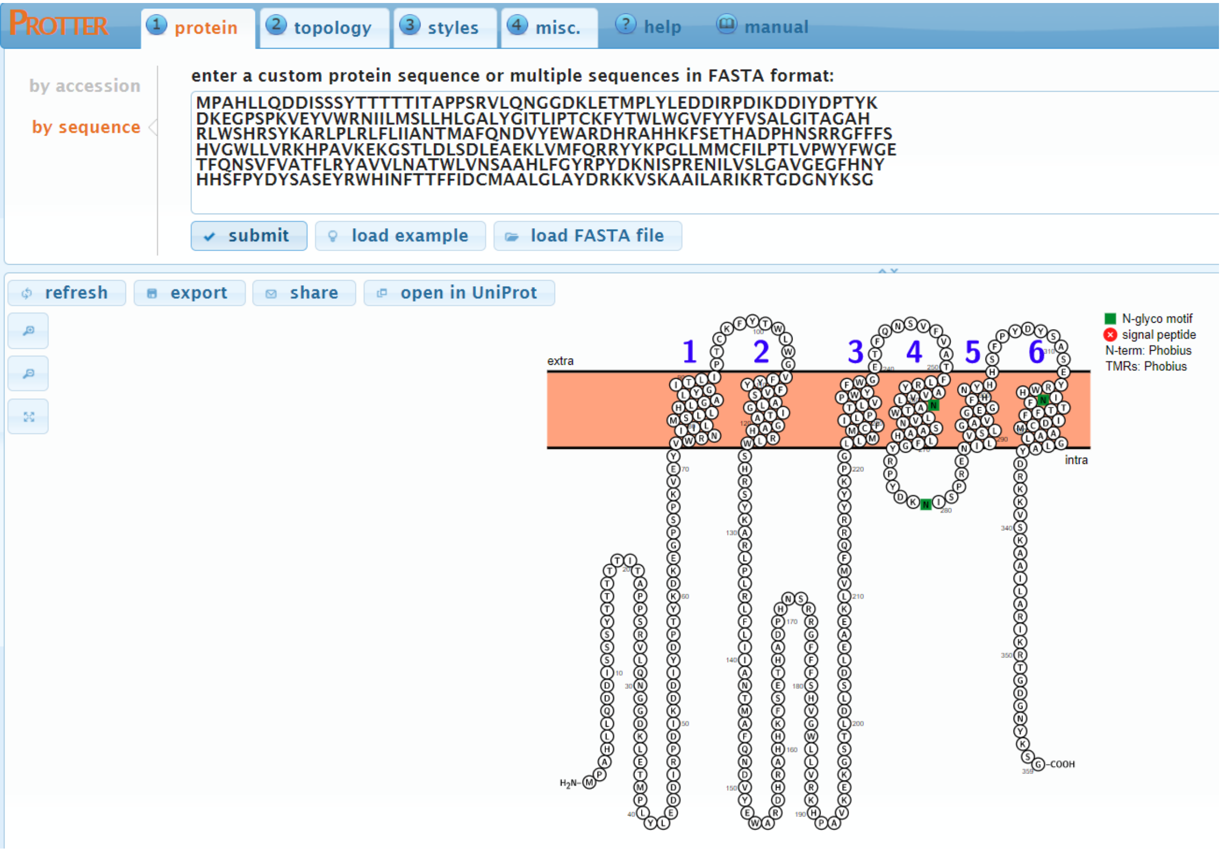
PROTTER-figure!



please enter a UniProt protein accession:

O00767

or, alternatively you can:



enter a custom protein sequence or multiple sequences in FASTA format:

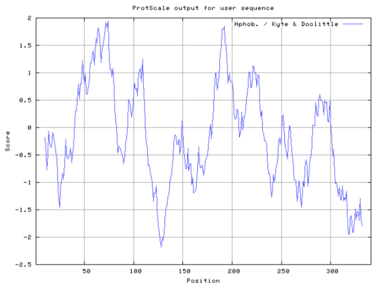
```
MPAHLLQDDISSSYTTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYK
DKEGPSPKVEYVWRNIILMSLLHLGALYGITLIPTCKFYTWLWGVFYYFVSALGITAGAH
RLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARDHRAHHKFSETHADPHNSRRGFFFS
HVGWLLVRKHPAVKEKGSTLDLSDLEAEKLVMFQRRYYKPGLLMMCFILPTLVPWYFWGE
TFQNSVFVATFLRYAVVLNATWLVNSAAHLFGYRPYDKNISPRENILVSLGAVGEGFHNY
HHSFPYDYSASEYRWHINFTTFFIDCMAALGLAYDRKKVSKAAILARIKRTGDGNYKSG
```

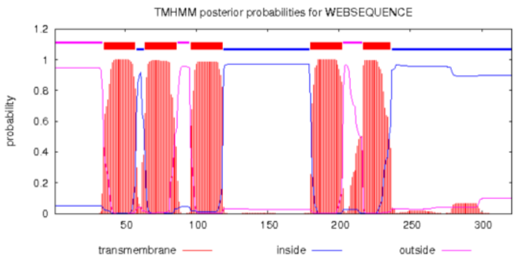
Prediction of transmembrane helices

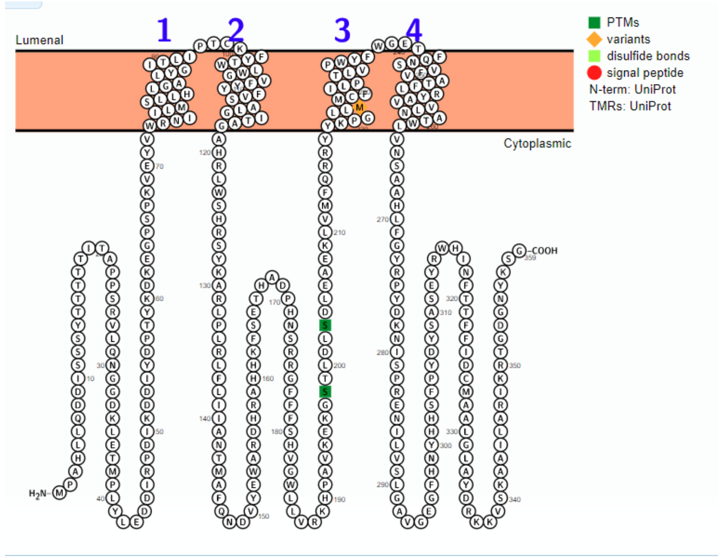
ProtScale

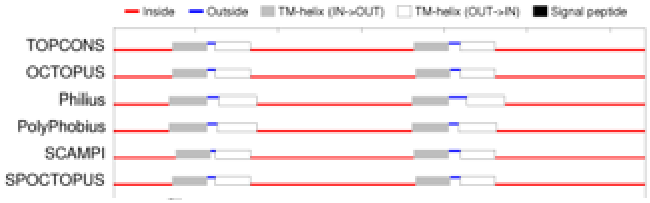
TMHMM

TOPCONS









➢ **Always try more programs!**

Practical part

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

…

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

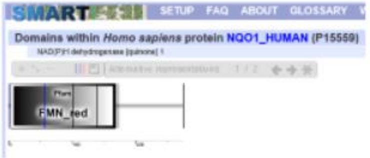
Homework 3 - example

1) Trypsin 33

D

2)

1)

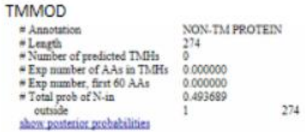


Domains within *Homo sapiens* protein NQO1_HUMAN (P15559)

NAD(P)H dehydrogenase [quinone] 1

2)

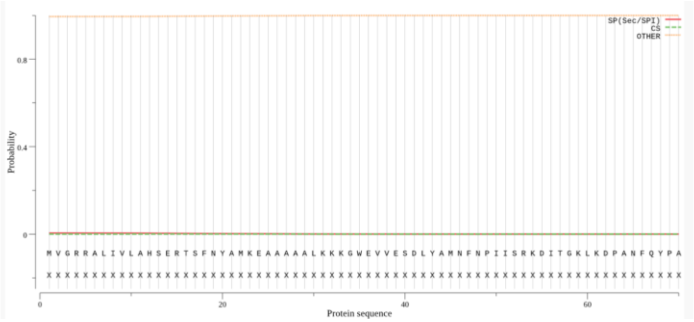
3)



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
```

4)



SP(Sec/SPI)
CS
OTHER
Probability
M V G R R A L I V L A H S E R T S F N Y A M K E A A A A A L K K K G W E V V E S D L Y A M N F N P I I S R K D I T G K L K D P A N F Q Y P A
X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
Protein sequence