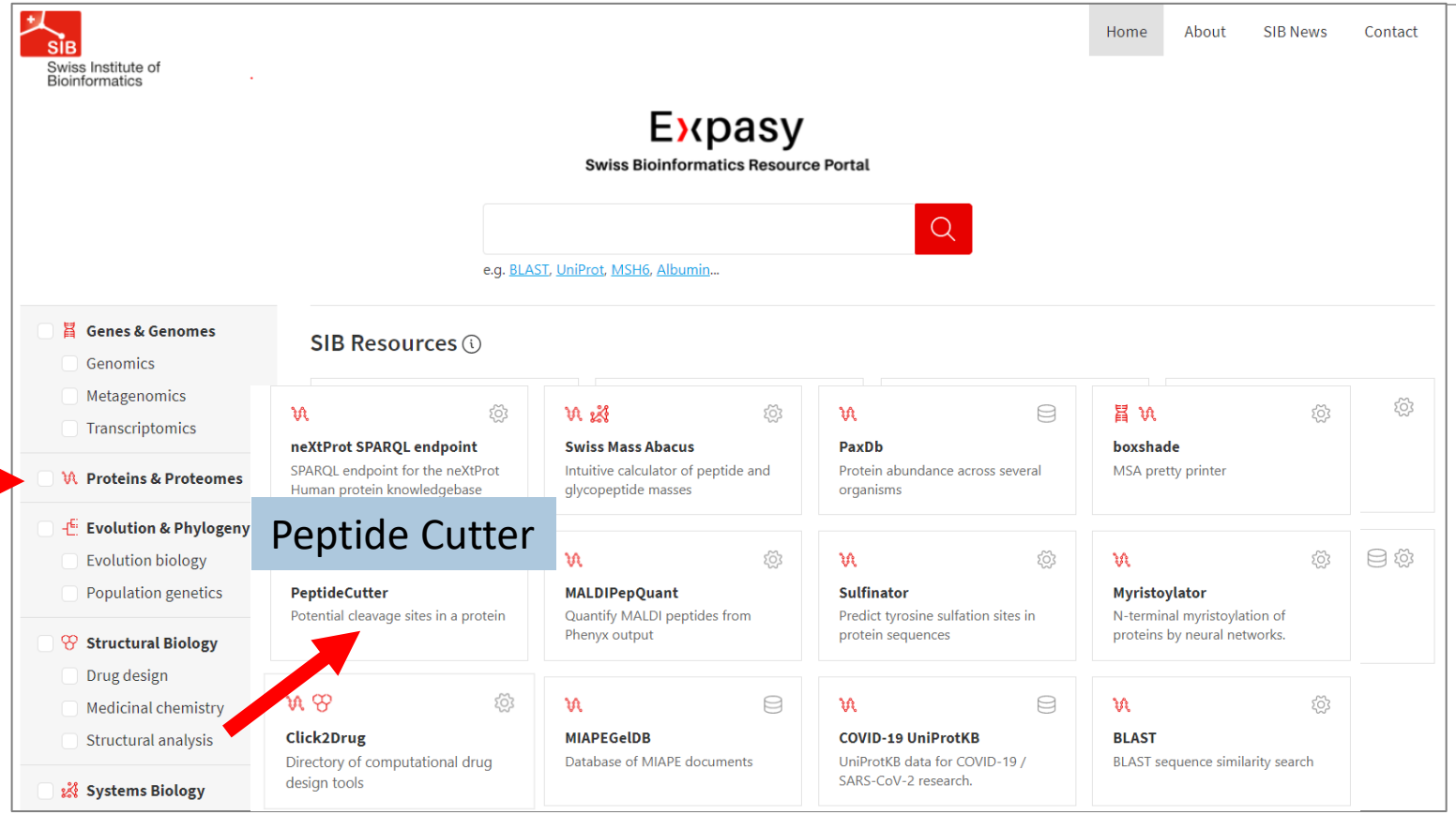
Základy praktické Bioinformatiky

PETRA MATOUŠKOVÁ

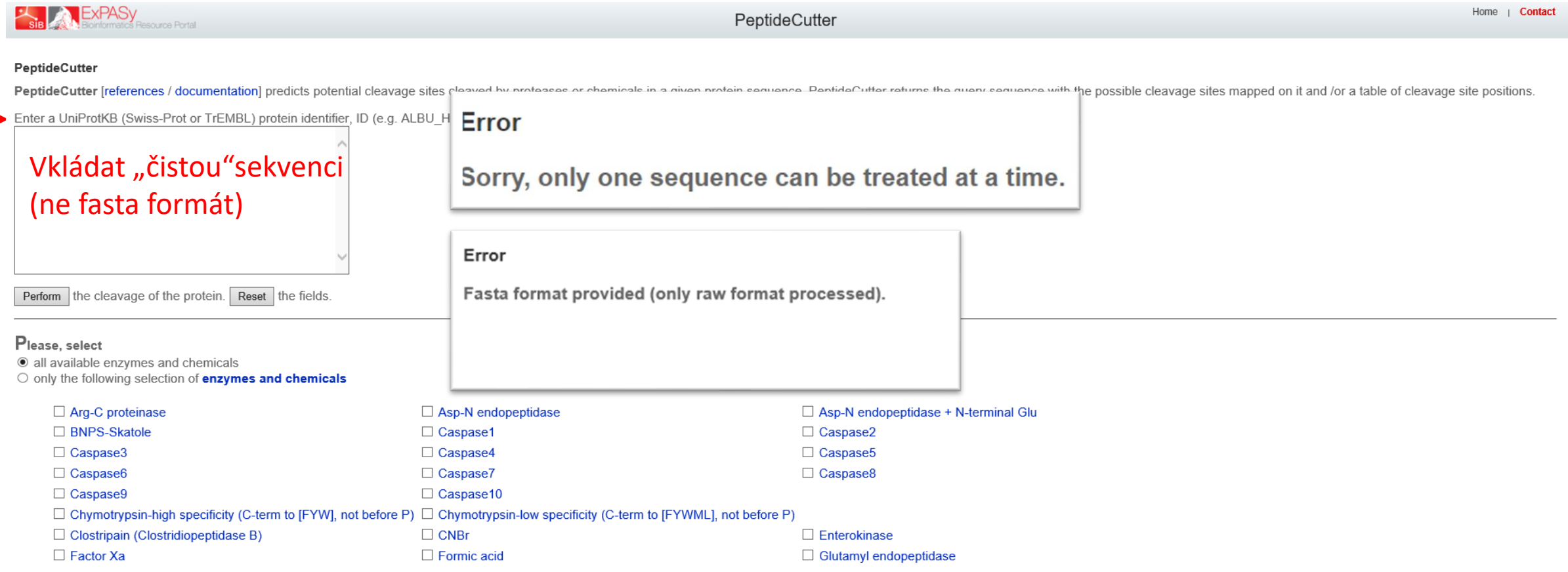
2025/2026

Štěpení proteasami

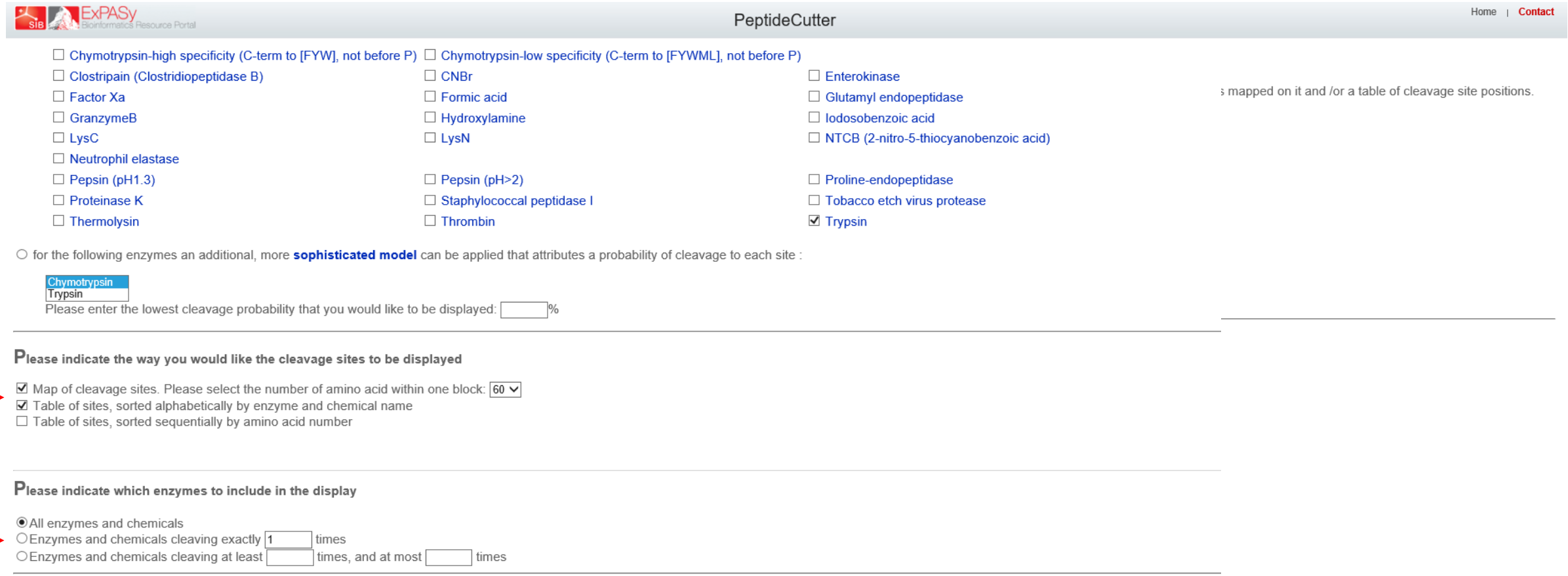


http://web.expasy.org/peptide_cutter/

Štěpení proteasami



Štěpení proteasami



Štěpení proteasami-PeptideCutter

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

These chosen enzymes do not cut:

Caspase1
Caspase10

„defaultní“ nastavení

Štěpení proteasami-PeptideCutter

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

Konkrétní enzym

[*] *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

Co štěpí přesně jednou?

Štěpení proteasami-PeptideCutter

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],	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
Clostripain		
Enterokinase		
Formic acid		
Glutamyl endopeptid		
Iodosobenzoic acid		
LysC		
LysN		
NTCB (2-nitro-5-thio acid)		
Pepsin (pH1.3)		
Pepsin (pH>2)		
Proteinase K		
Staphylococcal pepti		
Thermolysin		
Trypsin		

These chosen enzymes do not

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 th of the occurring amino acid re

[*] *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 , Try
Chymotrypsin-high specificity (C-term to [FYW], not before P) , Chymotrypsin-low specificity (C-term to [FYWML], not before P) , LysC , BNPS-Skatole , LysN ,

Position of cleavage site	Name of cl
4	Trypsin
5	Trypsin
15	Trypsin
23	Trypsin
31	Trypsin
32	Trypsin
33	Trypsin
53	Trypsin
54	Trypsin
59	Trypsin
61	Trypsin
77	Trypsin
90	Trypsin
91	Trypsin
114	Trypsin
119	Trypsin
135	Trypsin
139	Trypsin

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:

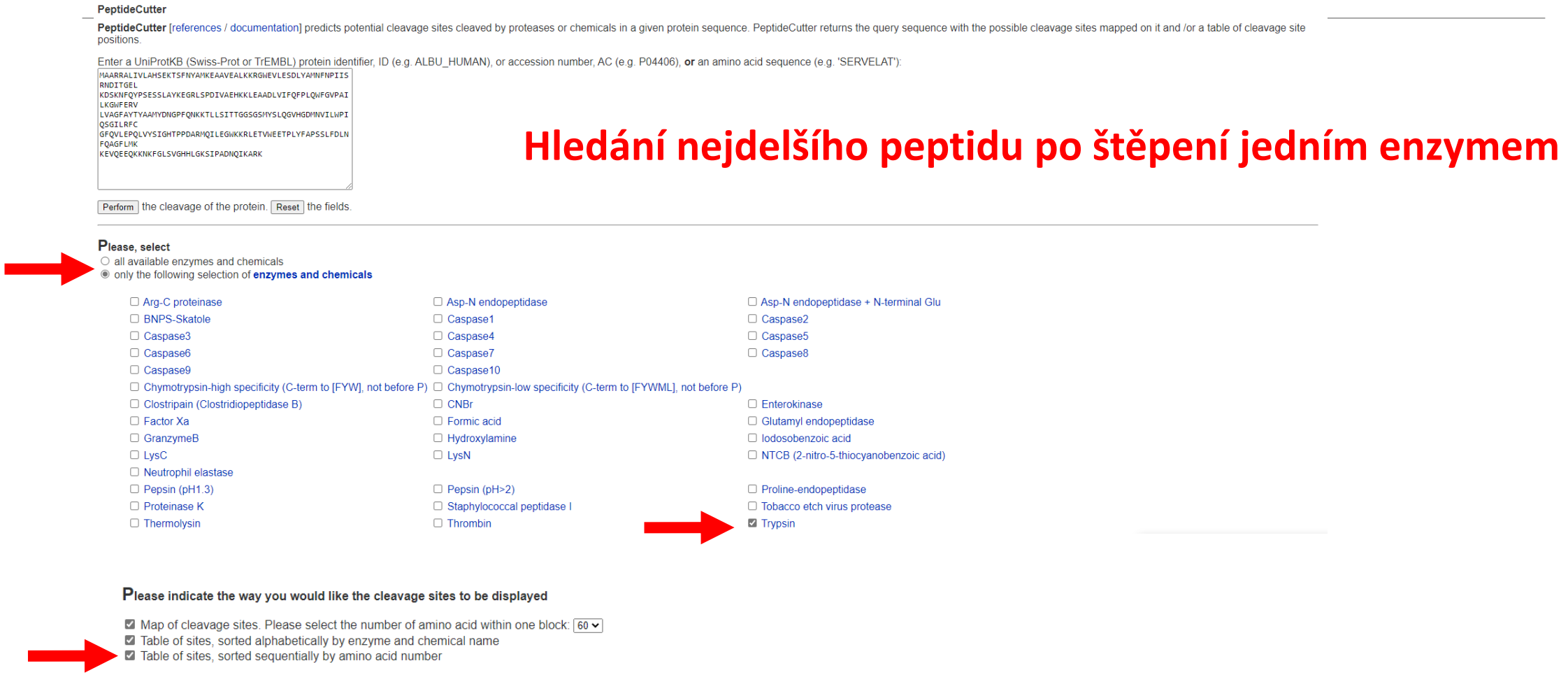
- Please note that the size of the peptides are calculated as if **all chosen enzymes were present** during digestion. If you want to obtain the size of the peptides resulting from the cleavage of only one enzyme, please, deselect th
- Please be aware of the fact that the present version of the PeptideCutter program does not take into consideration any kind of **modification** neither of the protein sequence nor of modifications evoked by the cleavage. Mass co 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)
179	NTCB (2-nitro-5-thiocyanobenzoic acid)	MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWEVVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRSKKAVLSITTGGSGSMYSLQGIHGDM
248	Enterokinase	CGFQVLEPQLTYSIGHTPADARIQILEGWKKRLENIWDETPLYFAPSSLFDLNFQAGFL MKKEVQDEEK
274	**end of sequence**	NKKFGLSVGHHLGKSIPTDNQIKARK

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

Štěpení proteasami-PeptideCutter



Hledání nejdelšího peptidu po štěpení jedním enzymem

ptideCutter

Name of enzyme	No. of cleavages	Positions of cleavage sites
Trypsin	32	4 5 15 23 31 32 33 53 61 64 77 80 90 91 114 119 141 142 178 201 209 210 211 240 241 248 249 251 262 271 273 274

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 peptic
- 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 p
amino acid residues, and giving peptide masses as [M]. If you want to select different parameters, we recommend to use PeptideMass.

y sequence with the possible cleavage sites mapped on it and /or a table of cleavage site

"):

Position of cleavage site	Name of cleaving enzyme(s)	Resulting peptide sequence (see explanations)	Peptide length [aa]	Peptide mass [Da]
4	Trypsin	MAAR	4	447.553
5	Trypsin	R	1	174.203
15	Trypsin	ALIVLAHSEK	10	1080.293
23	Trypsin	TSFNYAMK	8	[illegible]
31	Trypsin	EAAVEALK	8	829.949
32	Trypsin	K	1	146.189
33	Trypsin	R	1	174.203
53	Trypsin	GWEVLESDLYAMNFNPIISR	20	2354.662
61	Trypsin	NDITGELK	8	888.973
64	Trypsin	DSK	3	348.356
77	Trypsin	NFQYPSESSLAYK	13	1533.657
80	Trypsin	EGR	3	360.370
90	Trypsin	LSPDIVAEHK	10	1108.260
91	Trypsin	K	1	146.189
114	Trypsin	LEAADLVIFQFPLQWFGVPAILK	23	2616.141
119	Trypsin	GWFER	5	693.760
141	Trypsin	VLVAGFAYTYAAMYDNGPFQNK	22	2440.755
142	Trypsin	K	1	146.189
178	Trypsin	TLLSITTGGSGSMYSLQGVHGDMNVILWPIQSGILR	36	3803.406
201	Trypsin	FCGFQVLEPQLVYSIGHTPPDAR	23	2574.937
209	Trypsin	MQILEGWK	8	[illegible]
210	Trypsin	K	1	[illegible]
211	Trypsin	R	1	174.203
240	Trypsin	LETVWEETPLYFAPSSLFDLNFQAGFLMK	29	3394.886
241	Trypsin	K	1	146.189
248	Trypsin	EVQEEQK	7	888.930
249	Trypsin	K	1	146.189
251	Trypsin	NK	2	260.293
262	Trypsin	FGLSVGHHLGK	11	1151.333
271	Trypsin	SIPADNQIK	9	985.105
273	Trypsin	AR	2	245.282
274	**end of sequence**	K	1	146.189

rminal Glu

enzoic acid)

Hledání nejdelšího peptidu po štěpení

největší číslo

Vyzkoušejte si PeptideCutter

Analyzujte si „svou vlastní“ sekvenci

- Štěpí nějaký enzym pouze jedenkrát? (-vyzkoušejte štěpení všemi (default nastavení)
  (nebo: Vyberte pouze enzym, který štěpí jedenkrát.)
- Jak dlouhý je nejdelší vzniklý peptid po štěpení trypsinem? (omezte štěpení pouze na Trypsin, zaškrtněte si tabulku s „počtem aminokyselin“)

Základy praktické bioinformatiky

Téma 3/10

Proteinová bioinformatika II

Cíle:

Student bude schopen odhalit typické domény a transmembránové úseky v zadané sekvenci proteinu. Bude schopen vyhledat podobné sekvence a na základě podobnosti bude schopen odhadnout funkci neznámého proteinu.

shrnutí - Rešeršní projekt: NQO1

Vyhledejte dostupné informace o NQO1, získejte co nejvíce literárních, sekvenčních, případně i strukturních a dalších údajů o tomto genu/proteinu.

Nalezněte podobné sekvence a porovnejte je na proteinové úrovni.

- **NAD(P)H:chinonoxidoreduktasa / reference**
- **Protein: NP_000894, P15559 / 274 AMK; 30,8 kDa; trypsin štěpí 33x (59AMK nejdelší peptid)…**

„Proteinová bioinformatika“

Vyhledávání AMK sekvencí

Analýza vlastností sekvencí (aminokyselinové složení, molekulová hmotnost, isoelektrický bod…)

Štěpení proteasami

Analýza hydrofobních segmentů, transmembránových úseků

Predikce funkce (sekundární + terciární struktury → Hledání známých motivů

3D-struktura, vizualizace

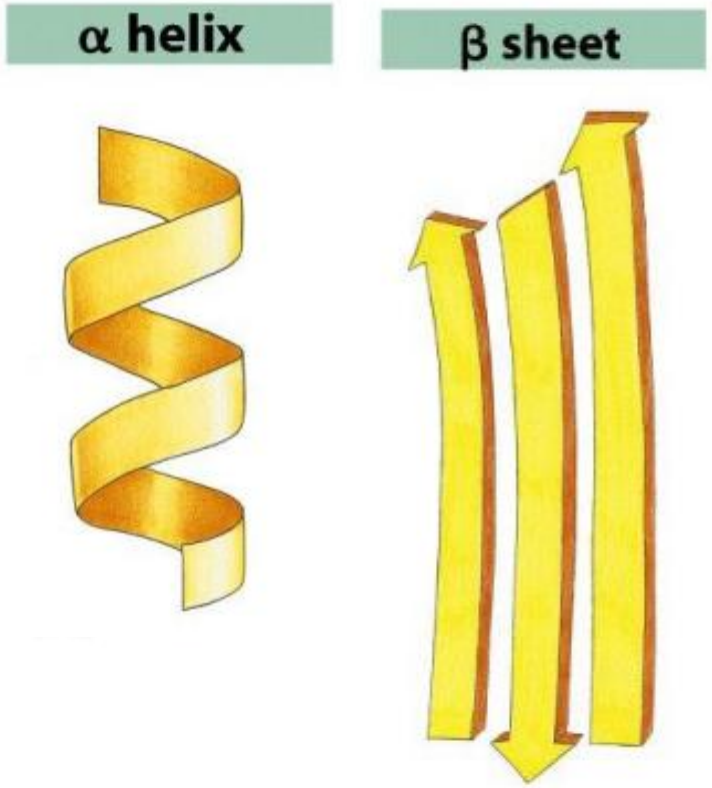
Vyhledání a porovnání podobných sekvencí

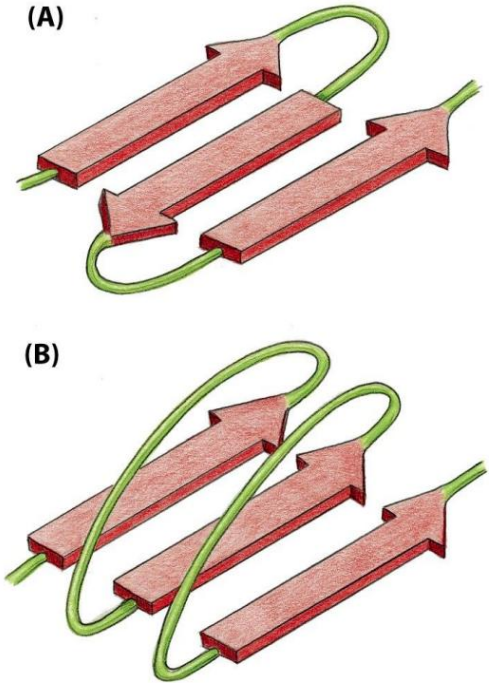
Evoluční příbuznost sekvencí

…

Proteiny sekundární struktura

- díky nekovalentním interakcím (lokálně mezi blízkými atomy)



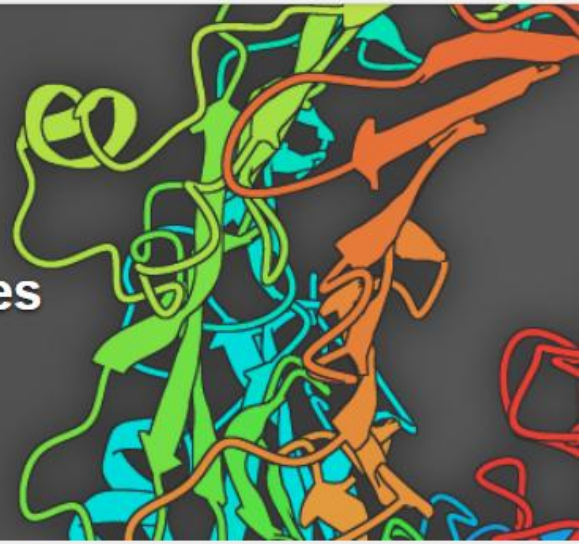


antiparalelní β-list

paralelní β-list

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

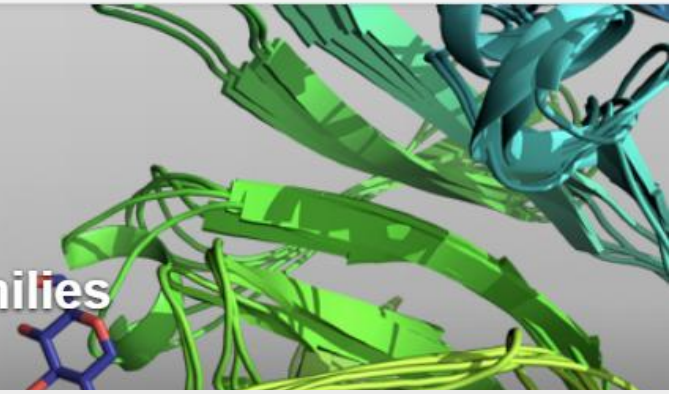
Find out more Go

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)

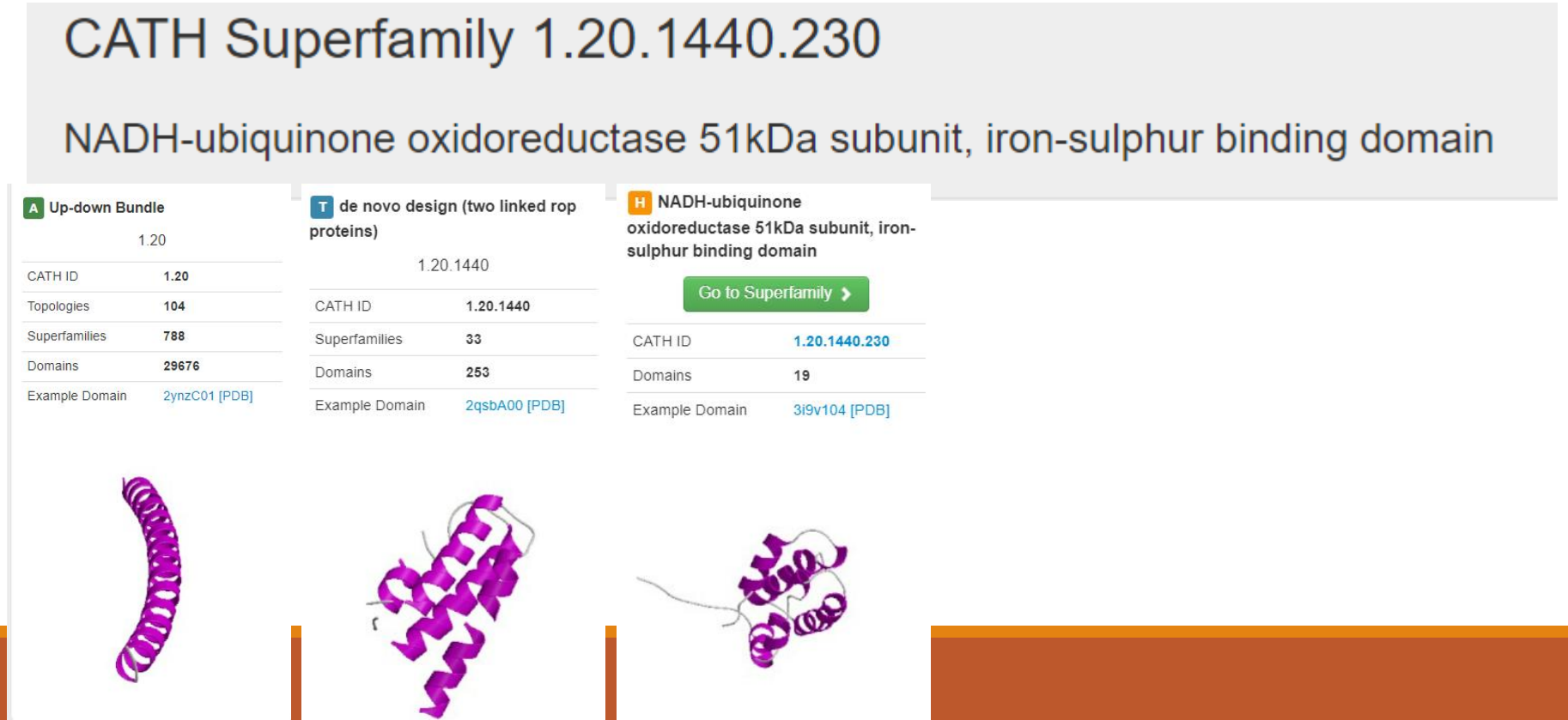
C	1	Mainly Alpha	5 Architectures, 404 Folds, 2033 Superfamilies, 103788 Domains
C	2	Mainly Beta	21 Architectures, 244 Folds, 1290 Superfamilies, 124032 Domains
C	3	Alpha Beta	14 Architectures, 634 Folds, 2337 Superfamilies, 262275 Domains
C	4	Few Secondary Structures	1 Architectures, 108 Folds, 181 Superfamilies,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

CATH - vyzkoušejte

- podívejte se jaké číslo „nadrodiny“ (superfamily) váš protein – hledání pomocí zkratky (nebo ID)

- podívejte se do jaké skupiny patří váš protein – hledání ve „stromě“ - browse

NQO1:



Proteiny terciární struktura

- kombinace sekundárních struktur – trojrozměrná konformace

- nekovalentní interakce mezi vzdálenými atomy → **proteinové domény – typické motivy**

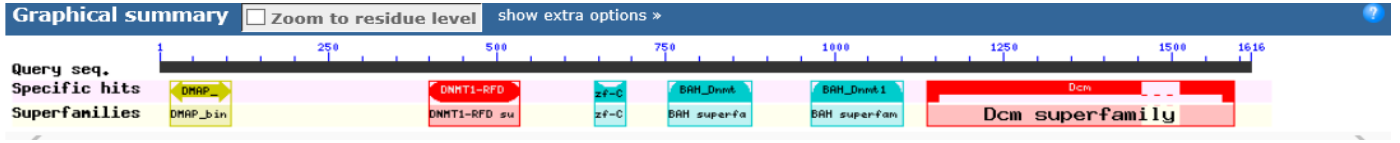
Hledání známých motivů

SEKVENCE➪STRUKTURA➪FUNKCE

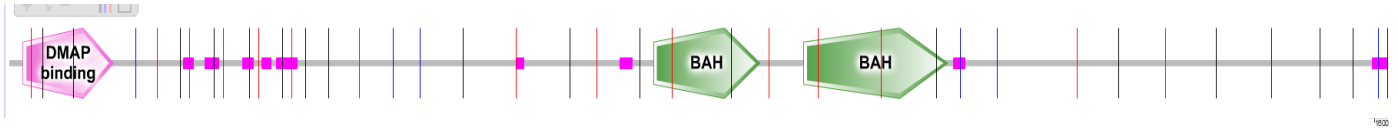
Sekvenční motivy, signální peptidy, místa posttranslačních modifikací ..

- Databáze známých motivů

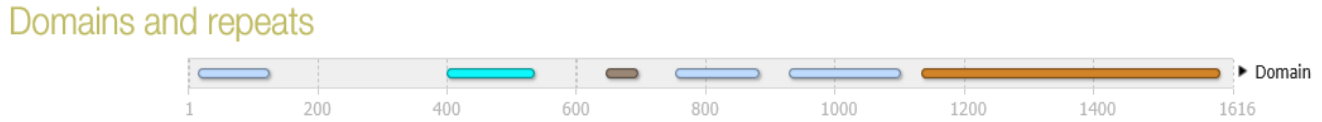
NCBI/CDD



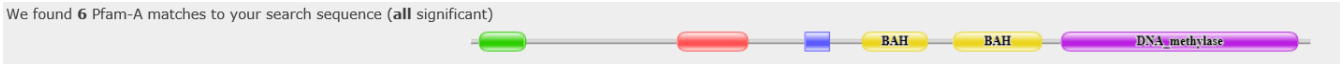
SMART



EMBL/InterPro

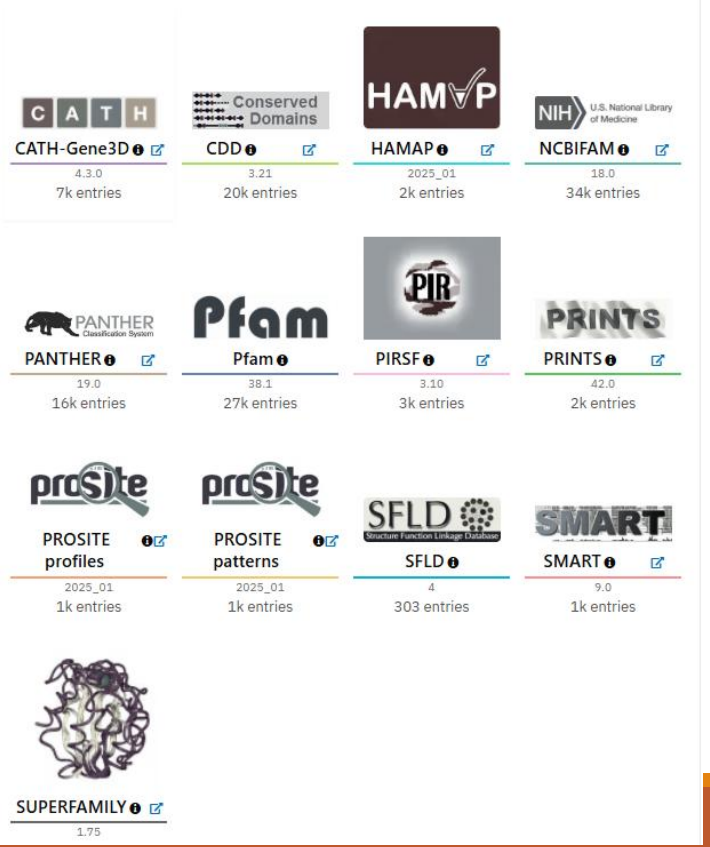


Pfam



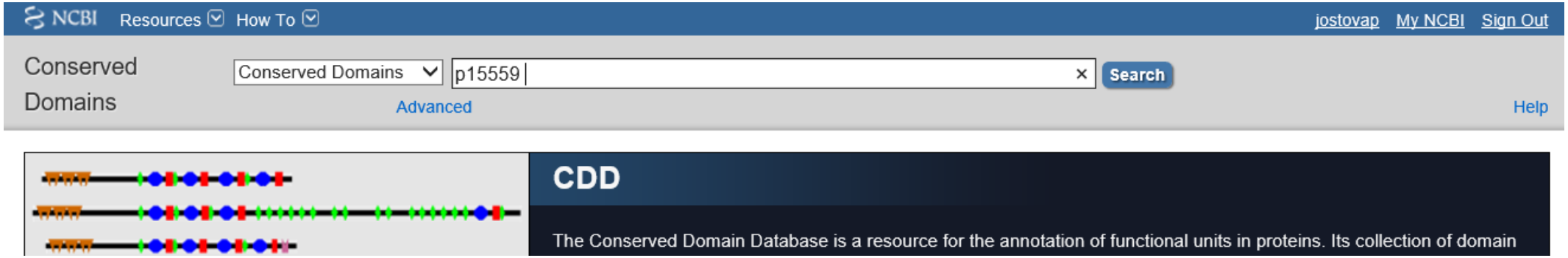
Pfam (bývalá databáze známých motivů - čísla **PF…**)

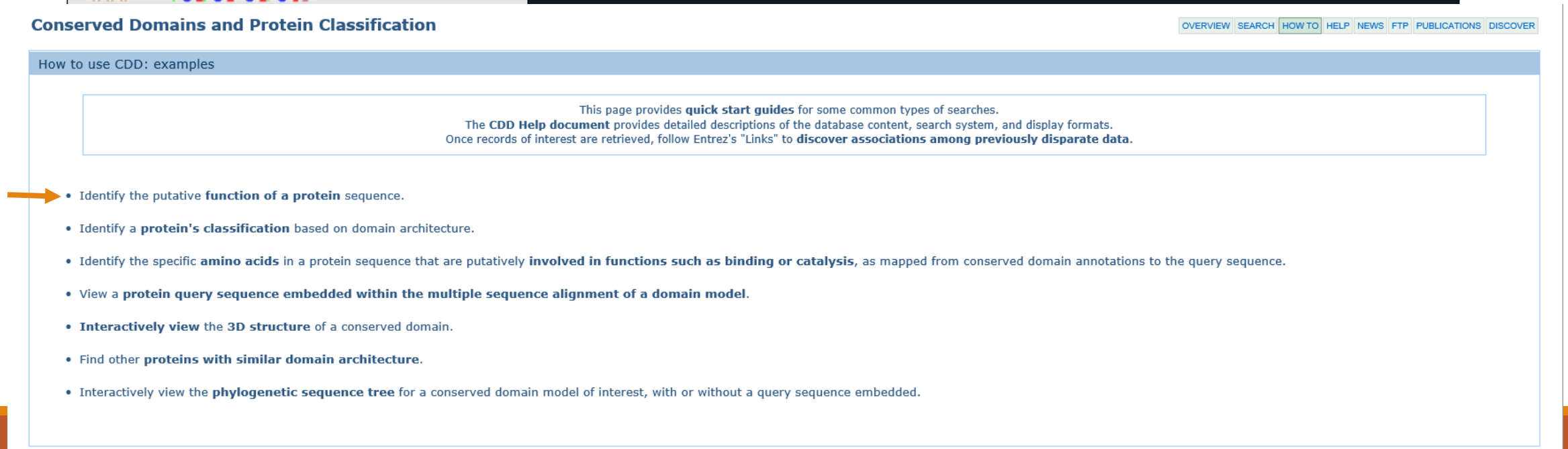
InterPro


CATH-Gene3D
4.3.0
7k entries
Conserved Domains
CDD
3.21
20k entries
HAMAP
2025_01
2k entries
NIH U.S. National Library of Medicine
NCBIFAM
18.0
34k entries
PANTHER
19.0
16k entries
Pfam
38.1
27k entries
PIR
PIRSF
3.10
3k entries
PRINTS
42.0
2k entries
prosite
PROSITE profiles
2025_01
1k entries
prosite
PROSITE patterns
2025_01
1k entries
SFLD
Structure Function Linkage Database
SFLD
4
303 entries
SMART
9.0
1k entries
SUPERFAMILY
1.75

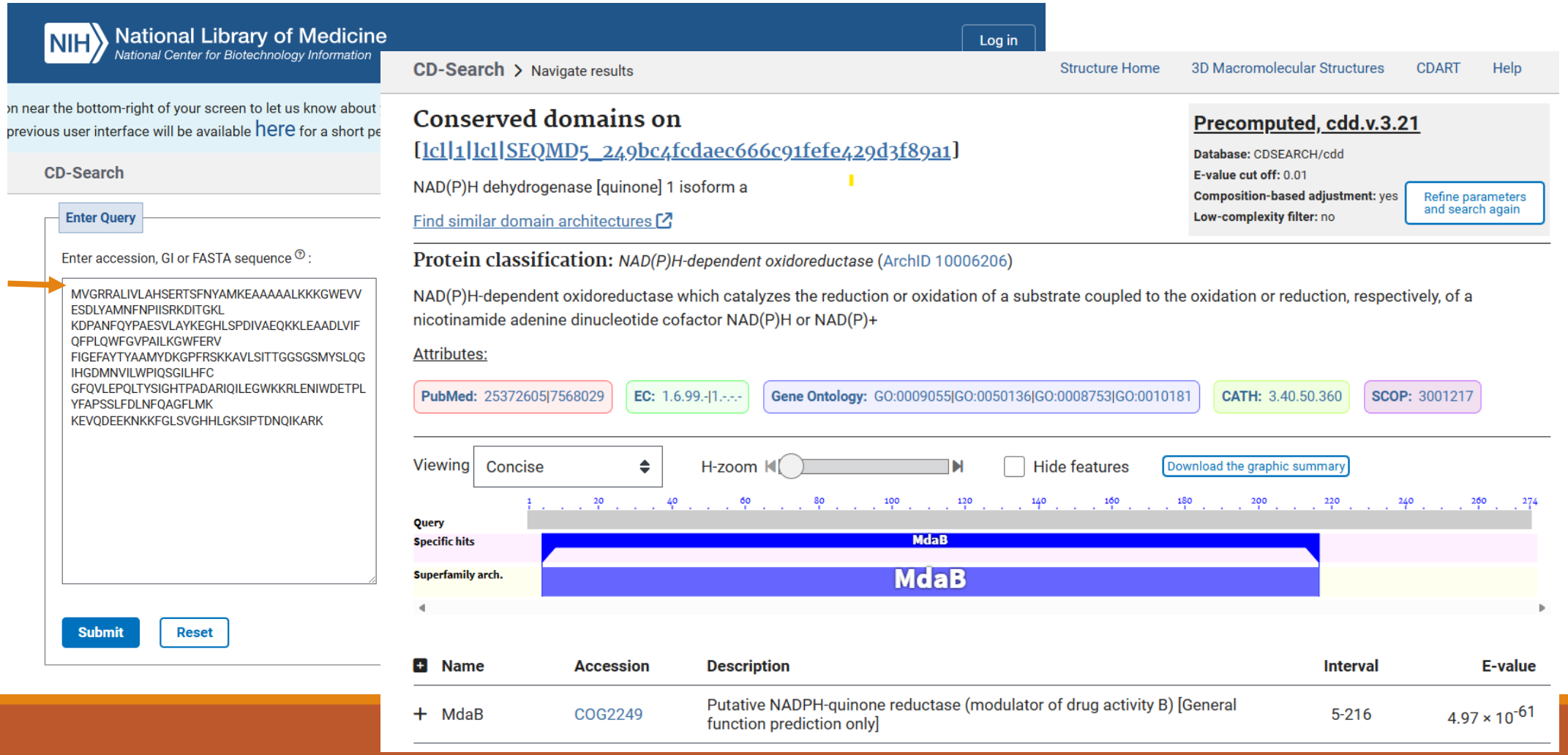
InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
InterPro 108.0
29 January 2026
Classification of protein families
InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.
If you use InterPro and/or Pfam, please cite our latest publications:
Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinares-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A.
InterPro: the protein sequence classification resource in 2025
Nucleic Acids Research. 2025, doi: 10.1093/nar/gkae1082
Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, Salazar GA, Bileschi ML, Llinares-López F, Meng-Papaxanthos L, Colwell LJ, Grishin NV, Schaeffer RD, Clementel D, Tosatto SCE, Sonnhammer E, Wood V, Bateman A.
The Pfam protein families database: embracing AI/ML
Nucleic Acids Research. 2025, doi: 10.1093/nar/gkae997
Search by sequence
Search by text
Search by Domain Architecture

Hledání známých motivů-CD (NCBI)

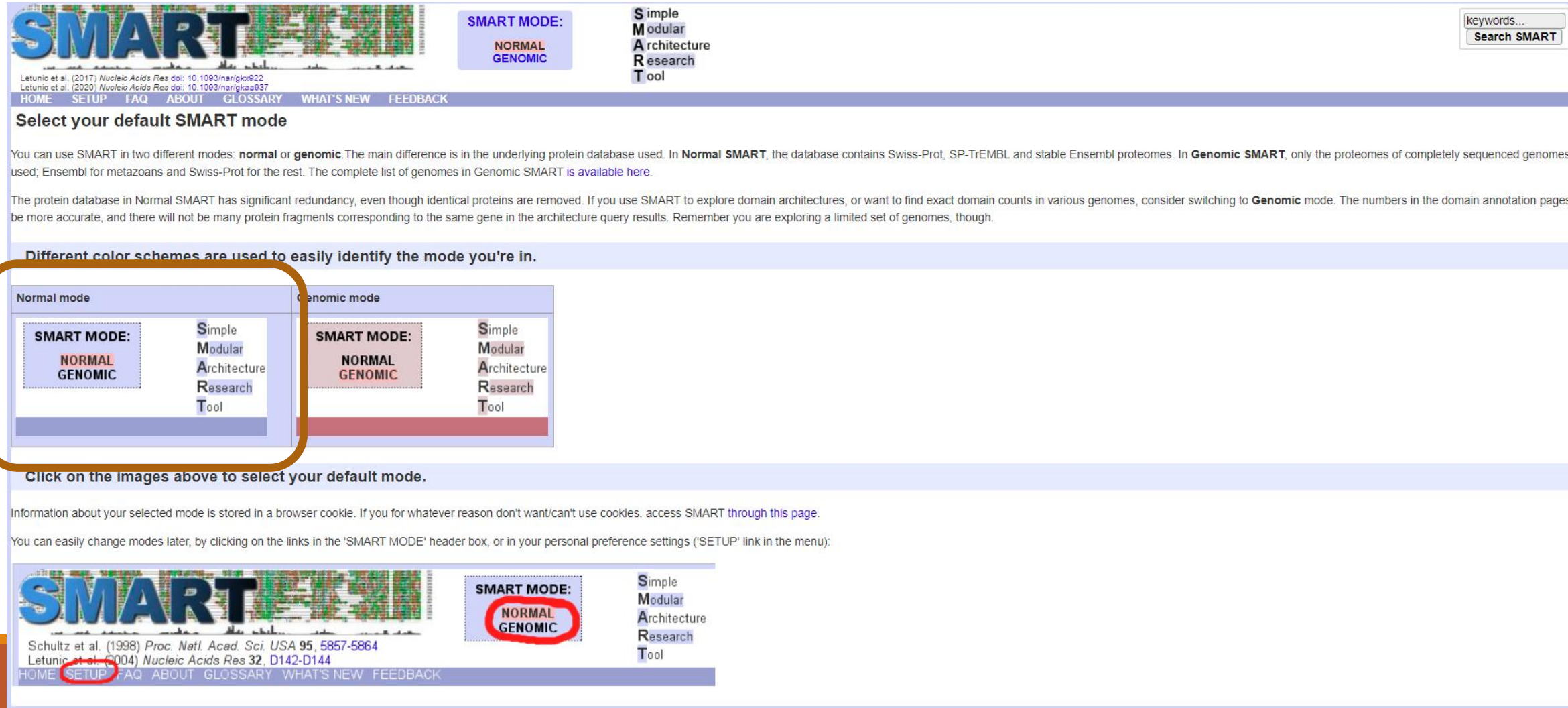




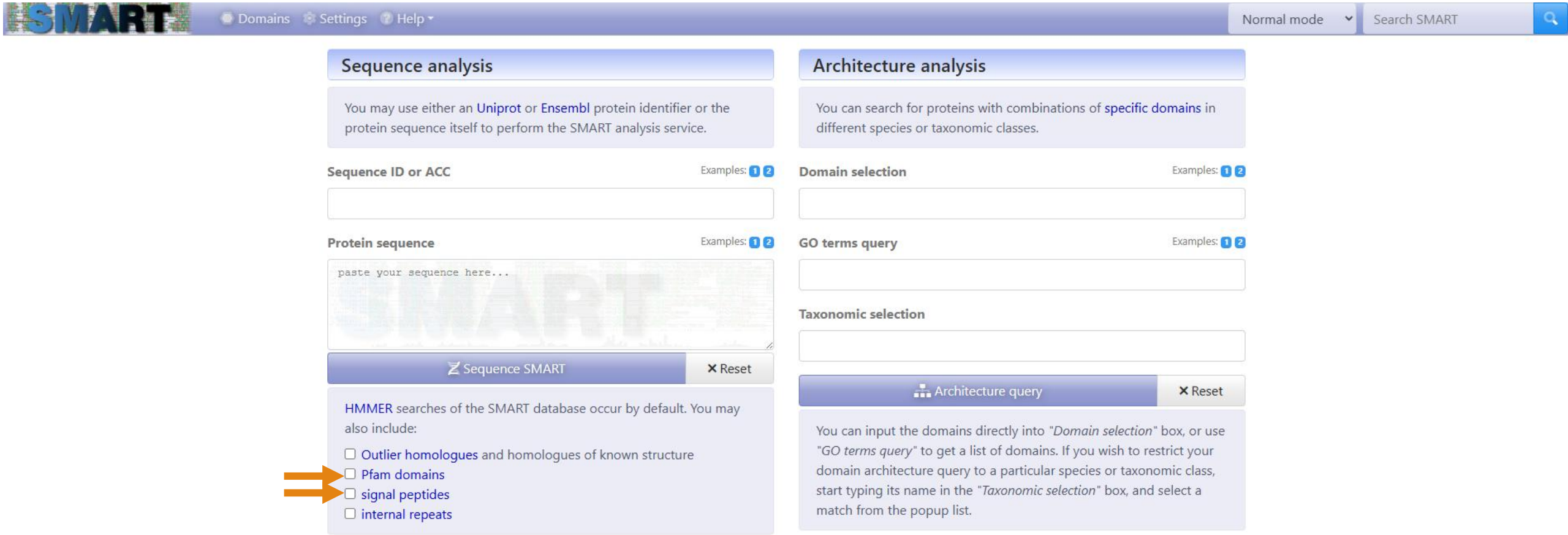
Hledání známých motivů-CD (NCBI)



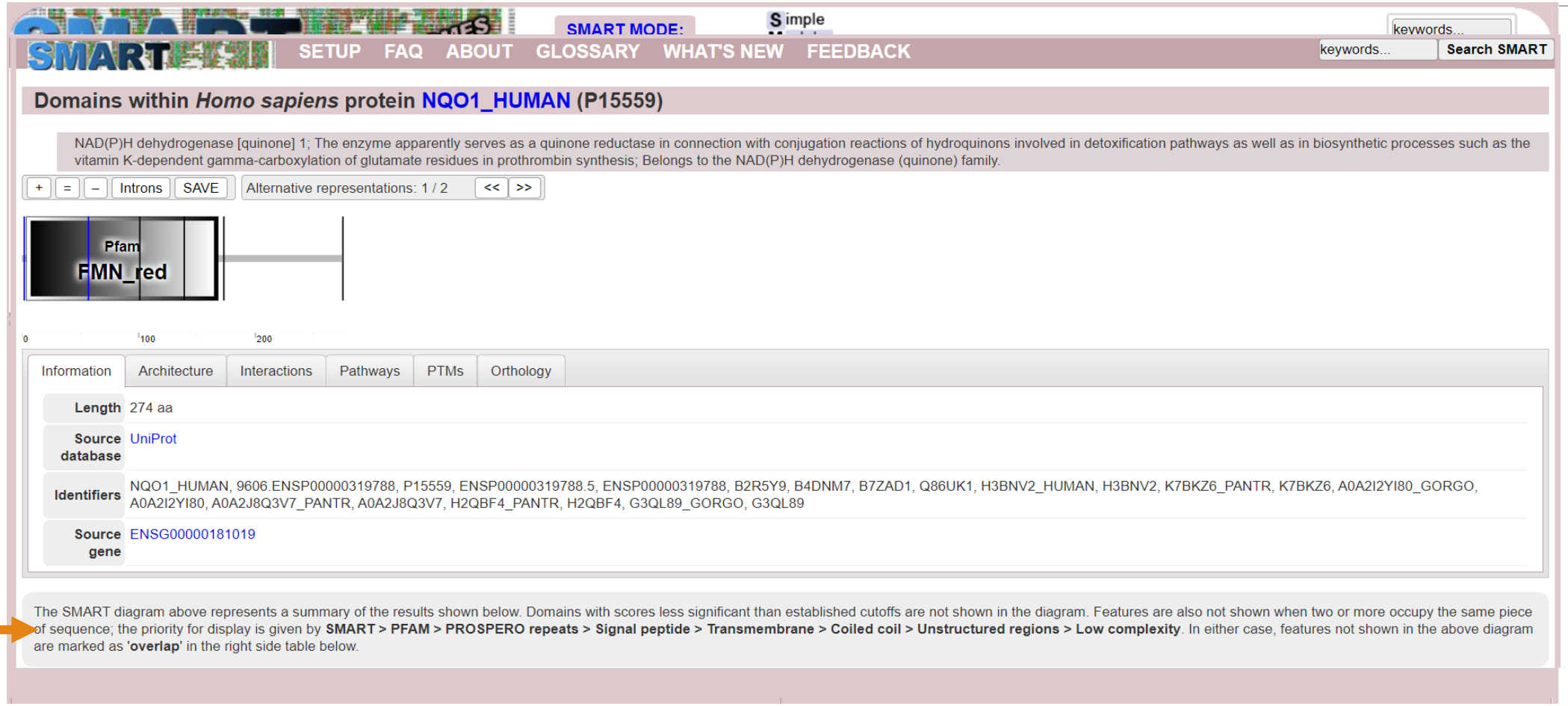
Hledání známých motivů-SMART



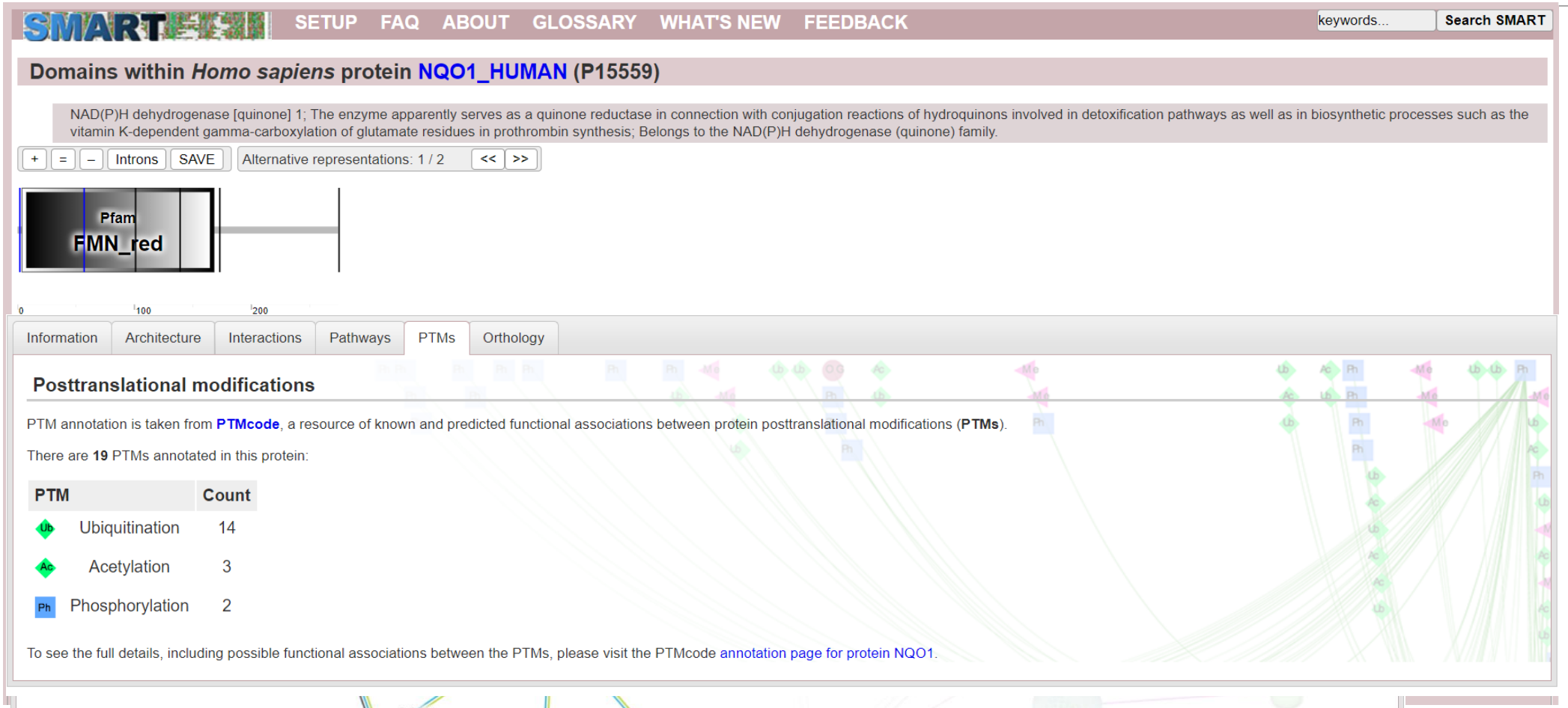
Hledání známých motivů-SMART



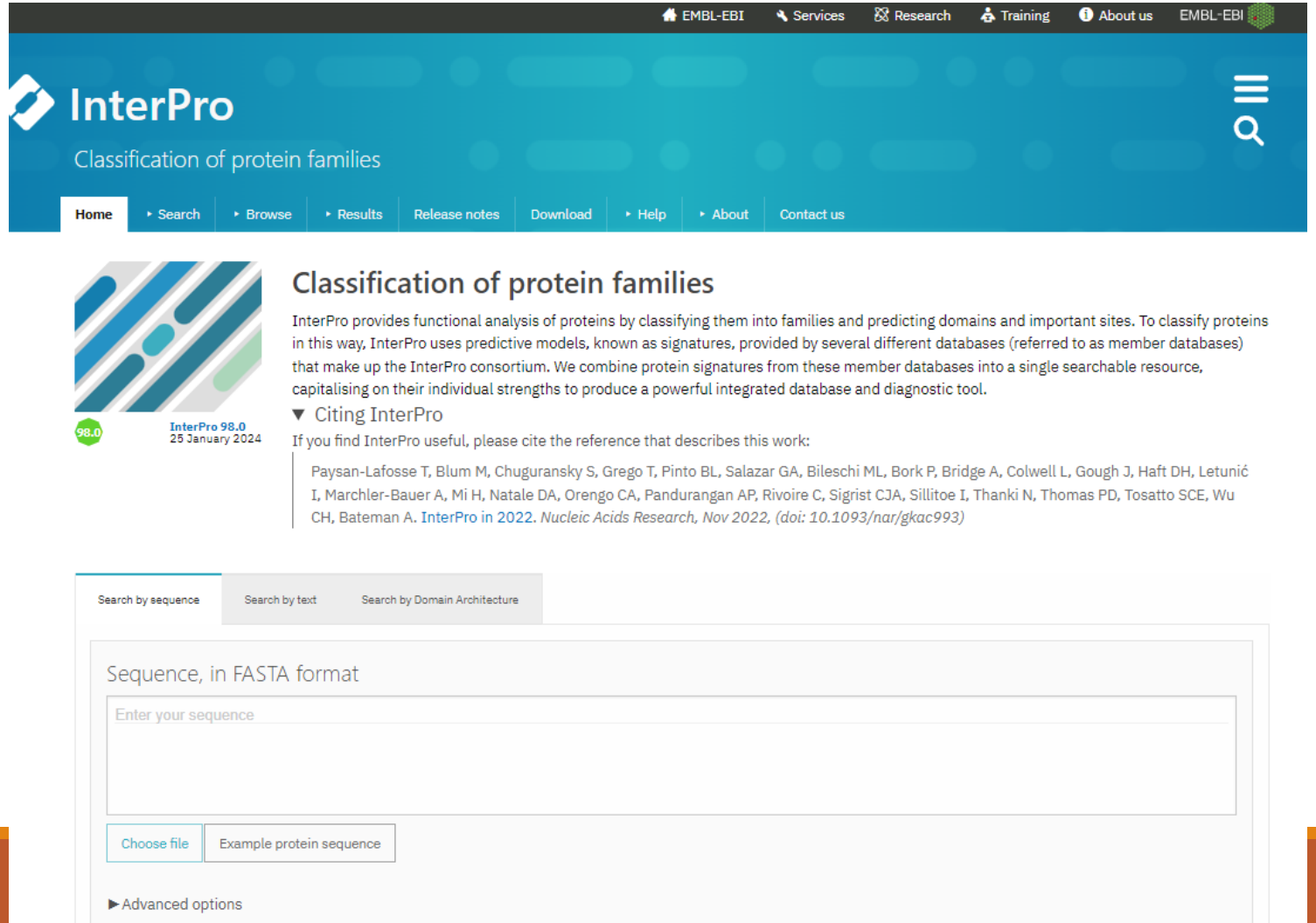
Hledání známých motivů-SMART



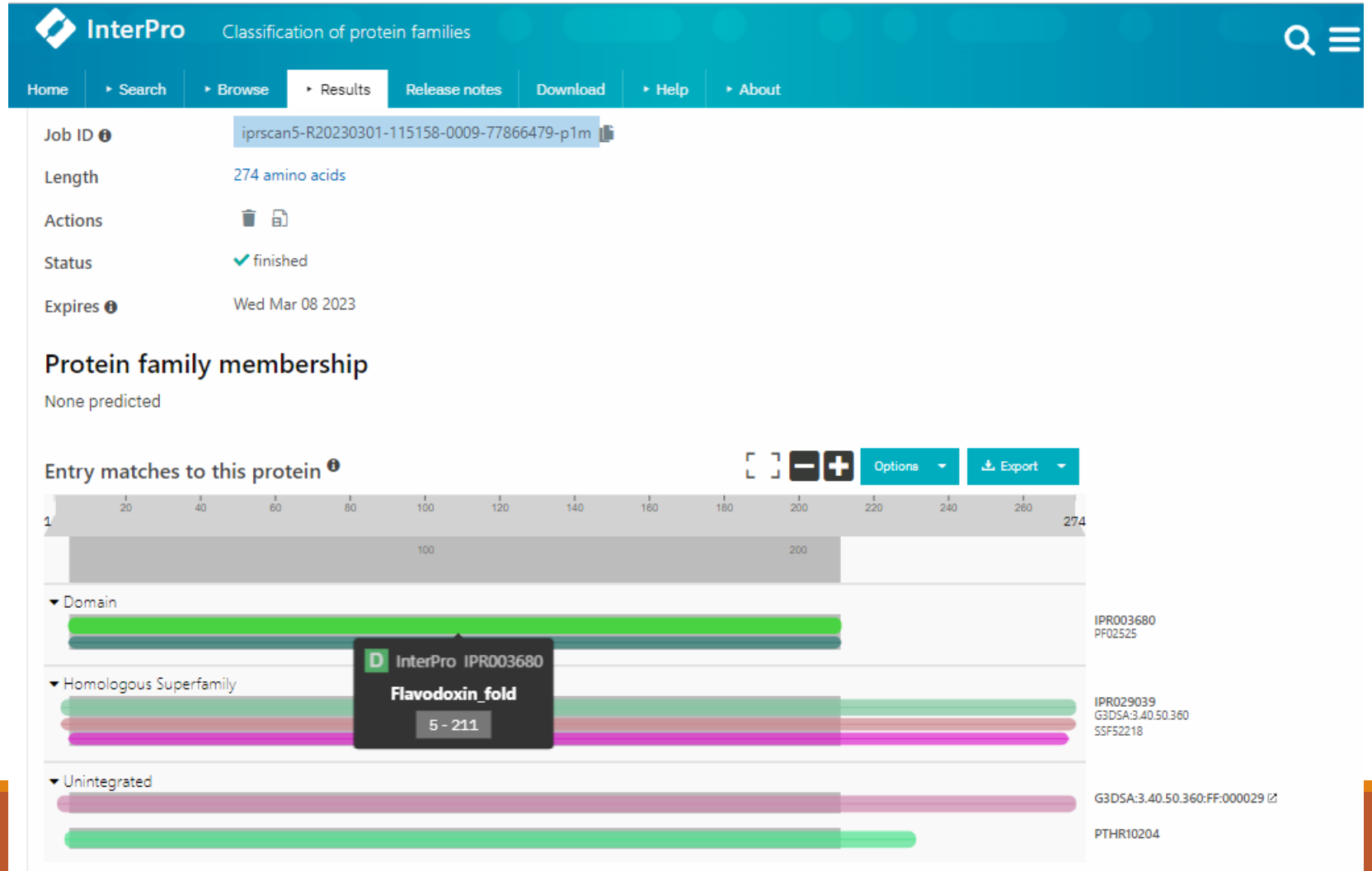
Hledání známých motivů-SMART



Hledání známých motivů-InterPro (2023)



Hledání známých motivů-InterPro (2023)

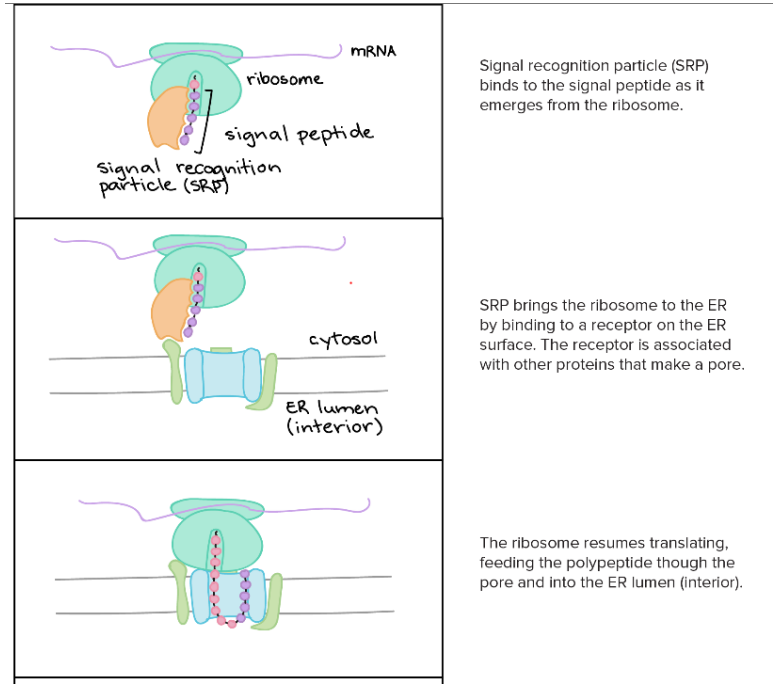


Praktická ukázka

- otevřete si soubor/stránku s vaší sekvencí
- otevřete si odkazy (každý v novém okně) a hledejte sekvenční motivy
- porovnejte výstupy z jednotlivých programů

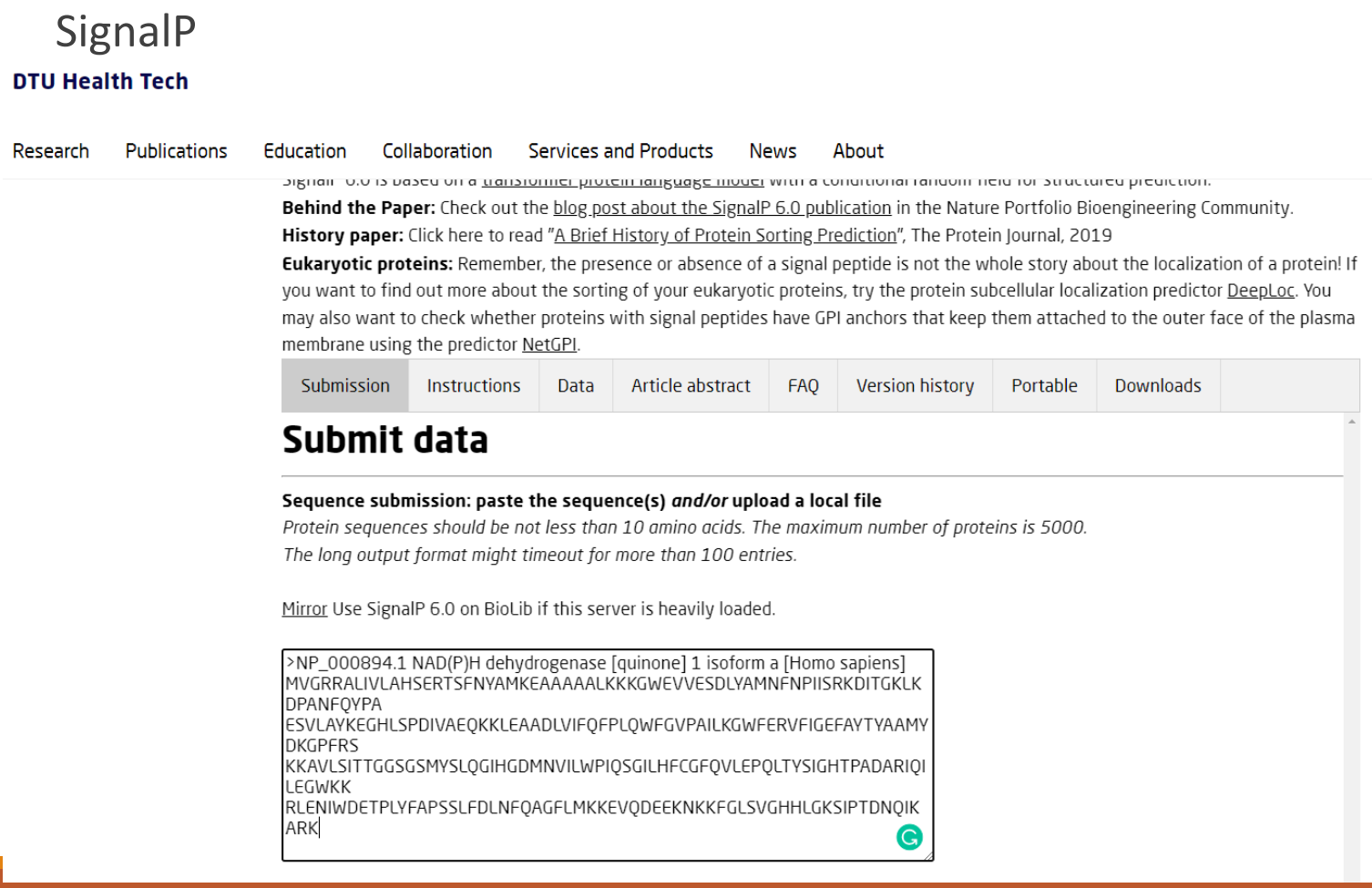
Hledání signálních peptidů

Signální peptid pro ER: 15-60 aminokyselin na N-konci proteinů



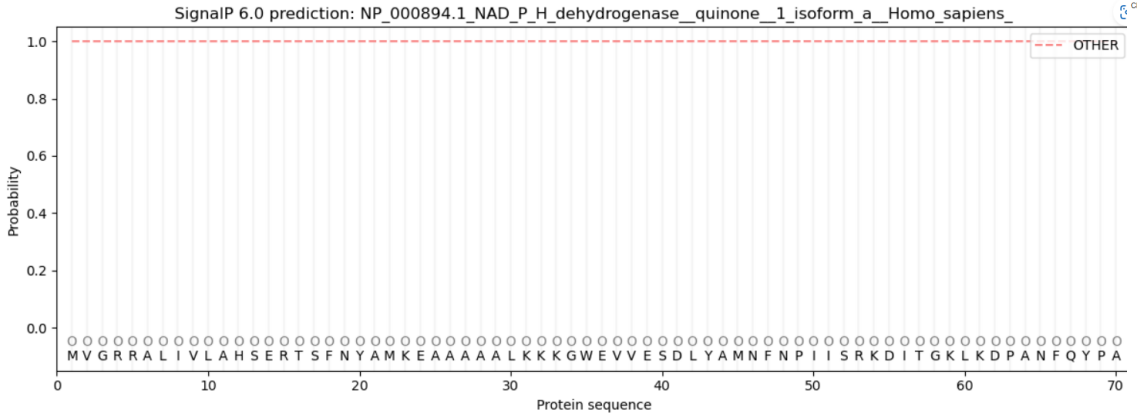
https://www.khanacademy.org/science/biology/gene-expression-central-dogma#central-dogma-transcription

Hledání signálních peptidů

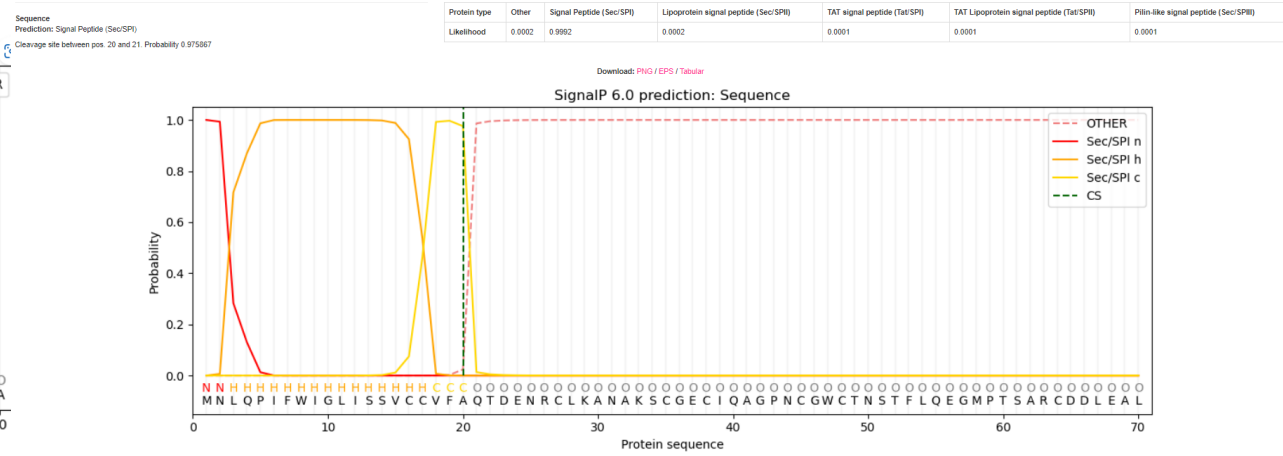


Hledání signálních peptidů

SignalP

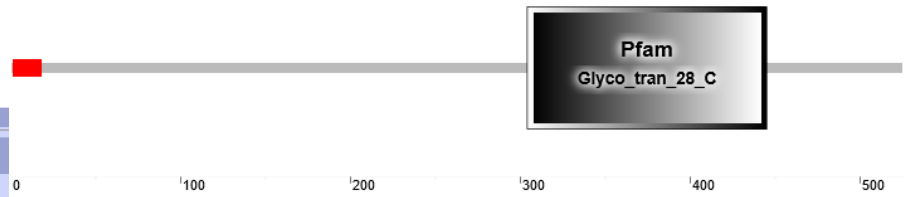


protein nemá signální peptid



protein MÁ signální peptid
(s určitou pravděpodobností)

Hledání signálních peptidů



Architecture

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 >**
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

Praktická ukázka

-podívejte se, jestli váš protein obsahuje signální peptid

- pokud nemáte, můžete se podívat jak to vypadá se sekvencí **O60656**

„Proteinová bioinformatika II“

Vyhledávání AMK sekvencí

Analýza vlastností sekvencí (aminokyselinové složení, molekulová hmotnost, isoelektrický bod…)

Štěpení proteasami

Analýza hydrofobních segmentů, transmembránových úseků

Predikce funkce, hledání známých motivů a signálních peptidů

3D-struktura, vizualizace

Vyhledání a porovnání podobných sekvencí

Evoluční příbuznost sekvencí

…

Predikce transmembránových úseků

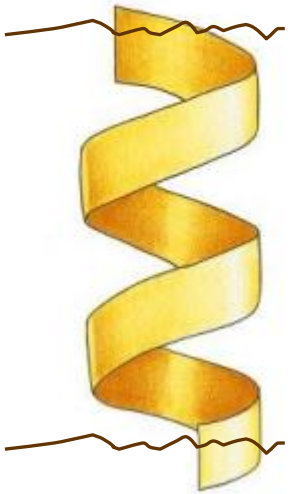
Hydrofobicita aminokyselin

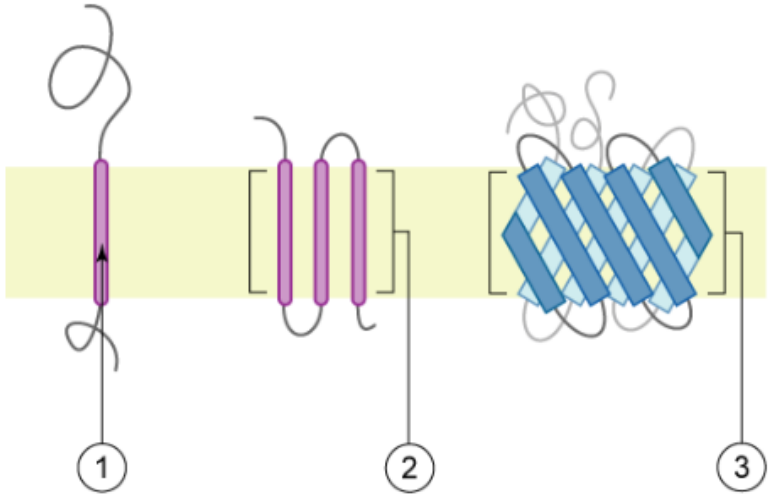
- různé programy – různé algoritmy výpočtů - různé výstupy

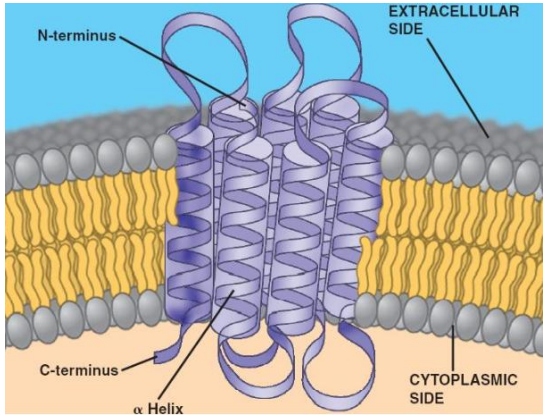
➢Předpověď topologického uspořádání (odhad, které části budou v cytoplazmě a které venku)

α helix

sbalený „naskrz membránou“





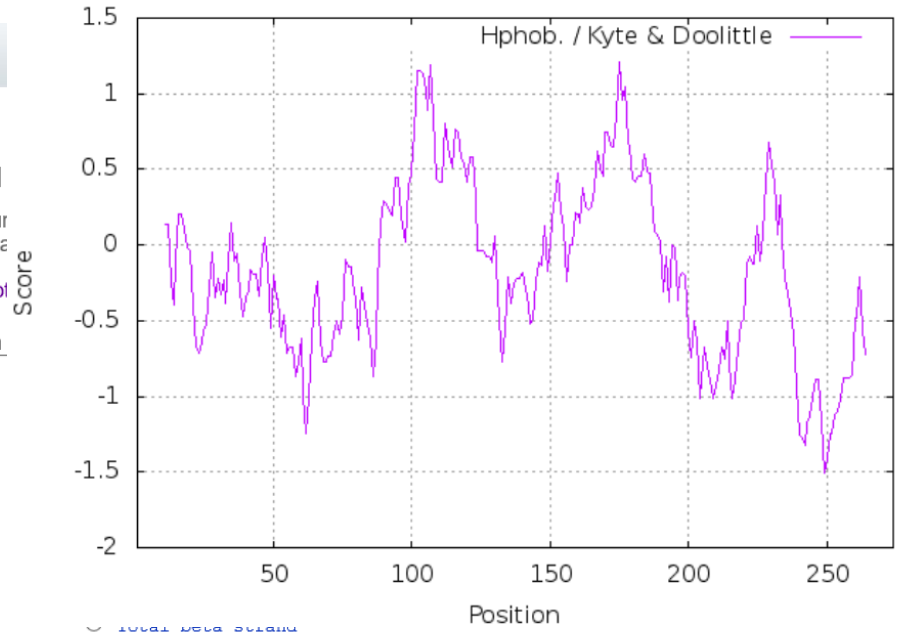


Predikce transmembránových úseků

Profil - Hydrof

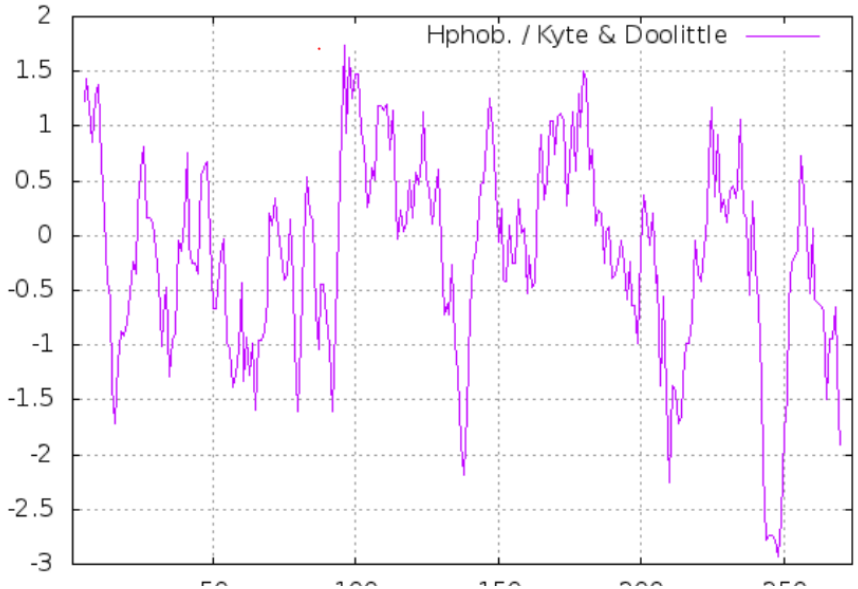
Window size: 21

cale output for user_sequence



Window size: 9

tput for user_sequence



ExPASy Bioinformatics Resource Portal

ProtScale

ProtScale [Reference / Documentation]

An **amino acid scale** is defined by a nur
but many other scales exist which are ba

Enter a UniProtKB/Swiss-Prot or UniProt

Or you can paste your own sequence in

- Hphob. / Chothia
- Ratio hetero end/side
- Average flexibility
- beta-sheet / Chou & Fasman
- alpha-helix / Deleage & Roux
- beta-turn / Deleage & Roux
- alpha-helix / Levitt
- beta-turn / Levitt
- Antiparallel beta-strand
- A.A. composition
- Relative mutability

- Parallel beta-strand
- A.A. comp. in Swiss-Prot

Window size: 9

Window size: 19 (21)

Relative weight of the window edges compared to the window center (in %): 100

Weight variation model (if the relative weight at the edges is < 100%): linear exponential

Do you want to normalize the scale from 0 to 1? yes no

If you need more information about how to set these parameters, please click here.

Submit Reset

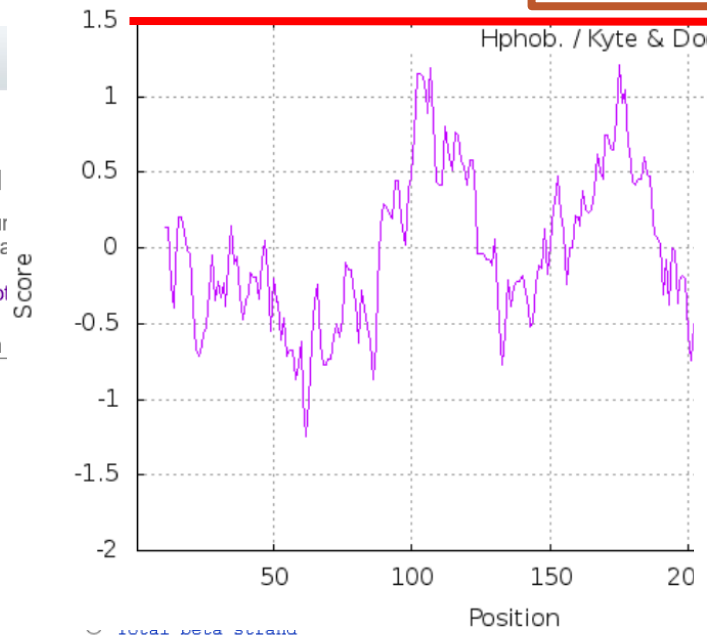
Predikce transmembránových úseků

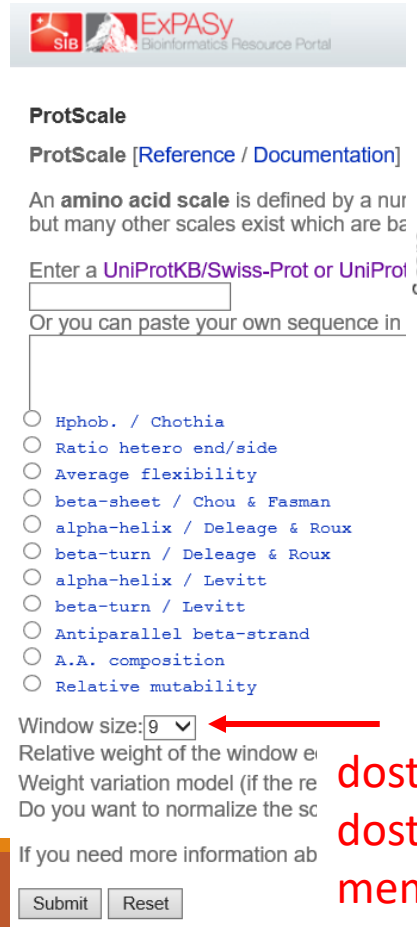
Profil - Hydrof

Window size: 21

cale output fo

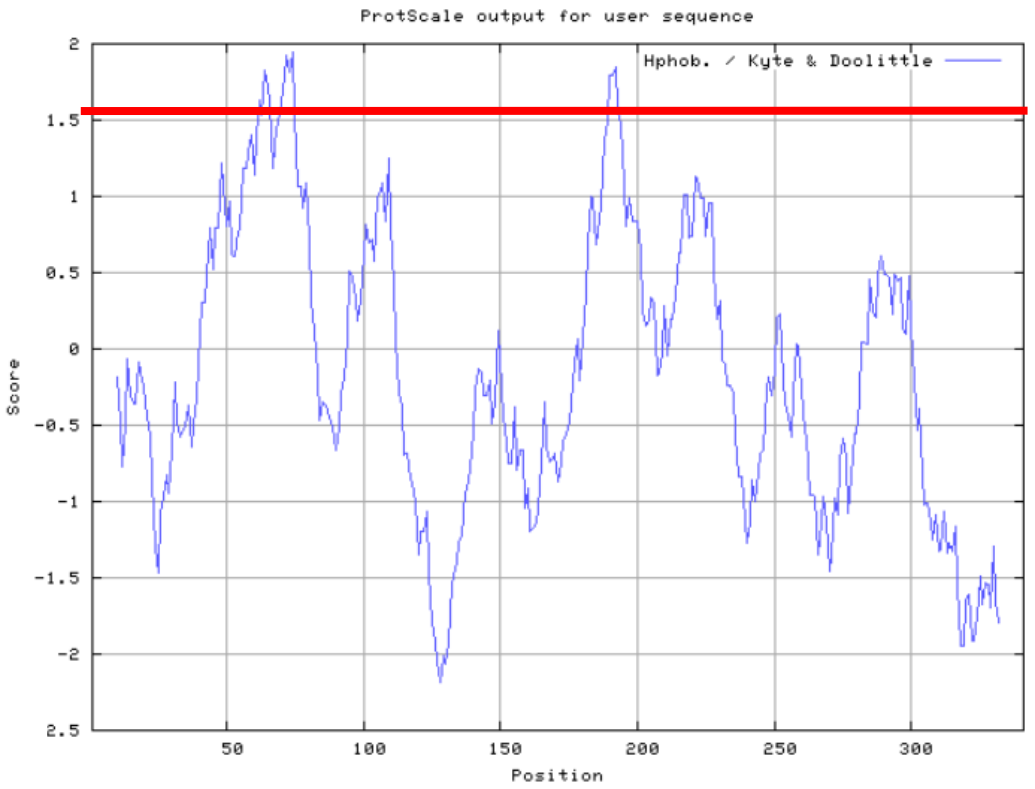
Cut off: 1.5 (1.6)





Window size: 19 (21)

dostatečně velké „okno“, které pojme dostatek aminokyselin pro průchod membránou v podobě helixu



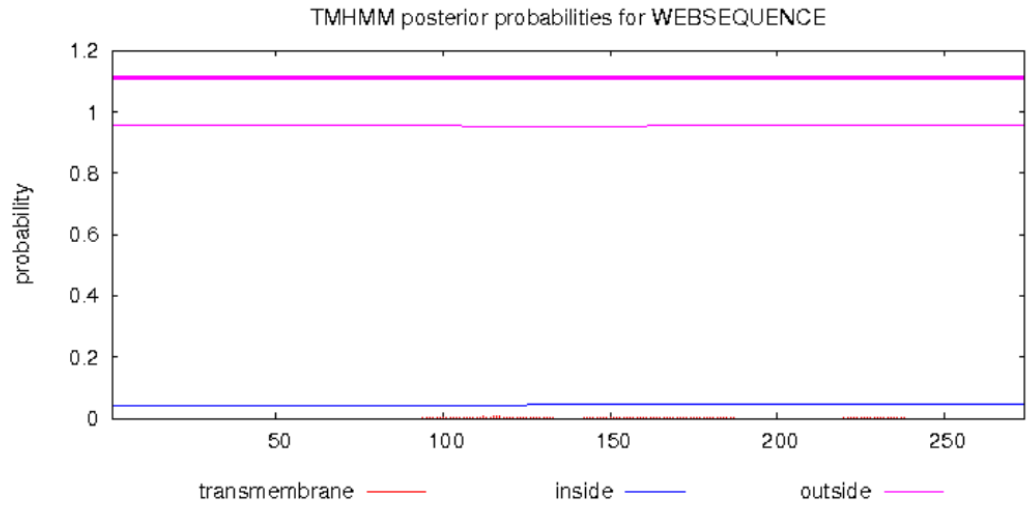
TMHMM

Desaturasa AM158251 (lišaj tabákový)

TMHMM result

HELP with output formats

```
# 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
```



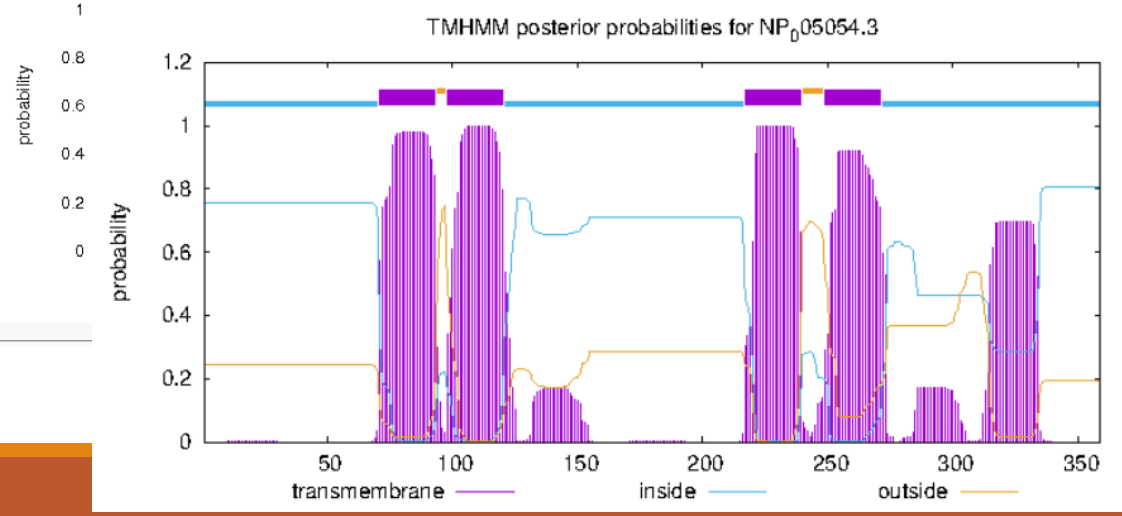
Desaturasa lidská SCD

TMHMM result

HELP with output formats

TMHMM result

```
# WEBSEQUE
# WEBSEQUE
# WEBSEQUE
# WEBSEQUE
# WEBSEQUE # NP_005054.3 Length: 359
# WEBSEQUE # NP_005054.3 Number of predicted TMHs:  4
WEBSEQUENC # NP_005054.3 Exp number of AAs in TMHs: 104.54198
WEBSEQUENC # NP_005054.3 Exp number, first 60 AAs:  0.03484
WEBSEQUENC # NP_005054.3 Total prob of N-in:        0.75464
WEBSEQUENC NP_005054.3     TMHMM2.0        inside       1    70
WEBSEQUENC NP_005054.3     TMHMM2.0        TMhelix     71    93
WEBSEQUENC NP_005054.3     TMHMM2.0        outside     94    97
WEBSEQUENC NP_005054.3     TMHMM2.0        TMhelix     98   120
WEBSEQUENC NP_005054.3     TMHMM2.0        inside     121   216
WEBSEQUENC NP_005054.3     TMHMM2.0        TMhelix    217   239
           NP_005054.3     TMHMM2.0        outside    240   248
           NP_005054.3     TMHMM2.0        TMhelix    249   271
           NP_005054.3     TMHMM2.0        inside     272   359
```



Phobius

A combined transme

Normal prediction | Constr predi

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

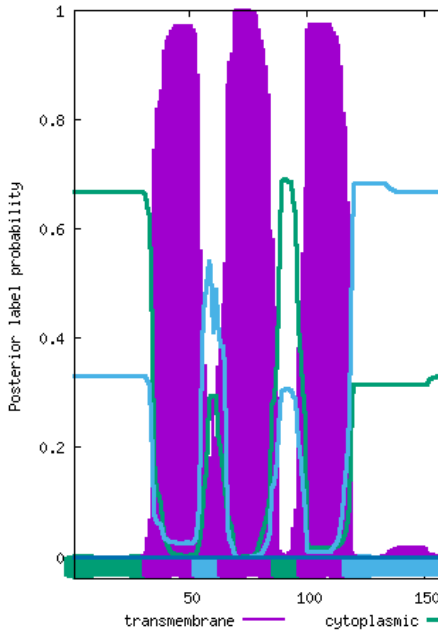
Odeslat Resetovat

Phobius prediction

Prediction of UNNAMED

```
ID   UNNAMED
FT   TOPO_DOM      1     33       CYTOPLASM
FT   TRANSMEM     34     54
FT   TOPO_DOM     55     65       NON CYTOP
FT   TRANSMEM     66     88
FT   TOPO_DOM     89     99       CYTOPLASM
FT   TRANSMEM    100    118
FT   TOPO_DOM    119    180       NON CYTOP
FT   TRANSMEM    181    202
FT   TOPO_DOM    203    213       CYTOPLASM
FT   TRANSMEM    214    234
FT   TOPO_DOM    235    344       NON CYTOP
//
```

Phobius posterior



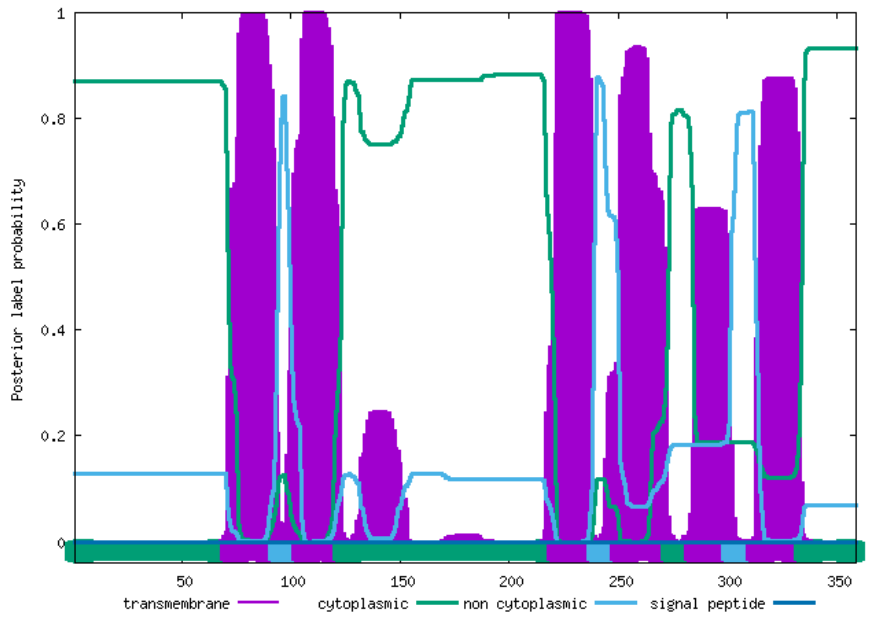
The probability data used in the plot is found here, and the

Phobius prediction

Prediction of NP_005054.3

```
ID   NP_005054.3
FT   TOPO_DOM      1     71       CYTOPLASMIC.
FT   TRANSMEM     72     93
FT   TOPO_DOM     94    104       NON CYTOPLASMIC.
FT   TRANSMEM    105    123
FT   TOPO_DOM    124    221       CYTOPLASMIC.
FT   TRANSMEM    222    239
FT   TOPO_DOM    240    250       NON CYTOPLASMIC.
FT   TRANSMEM    251    273
FT   TOPO_DOM    274    284       CYTOPLASMIC.
FT   TRANSMEM    285    301
FT   TOPO_DOM    302    312       NON CYTOPLASMIC.
FT   TRANSMEM    313    334
FT   TOPO_DOM    335    359       CYTOPLASMIC.
//
```

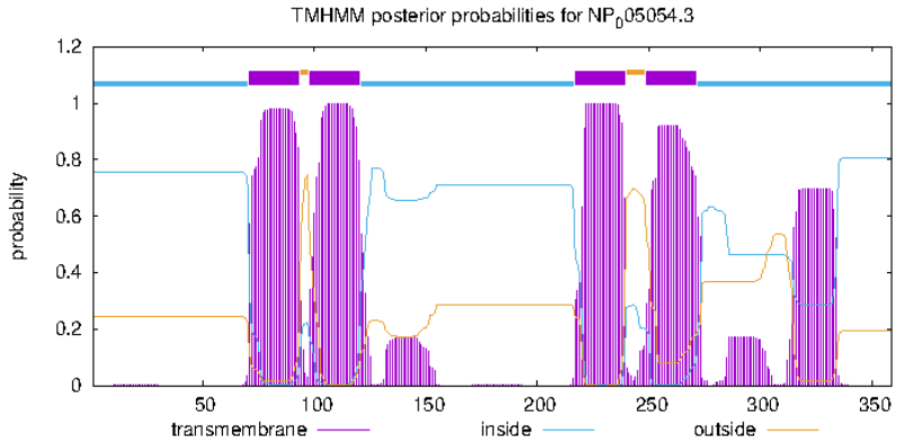
Phobius posterior probabilities for NP_005054.3



Predikce transmembránových úseků (SCD)

TMHMM result

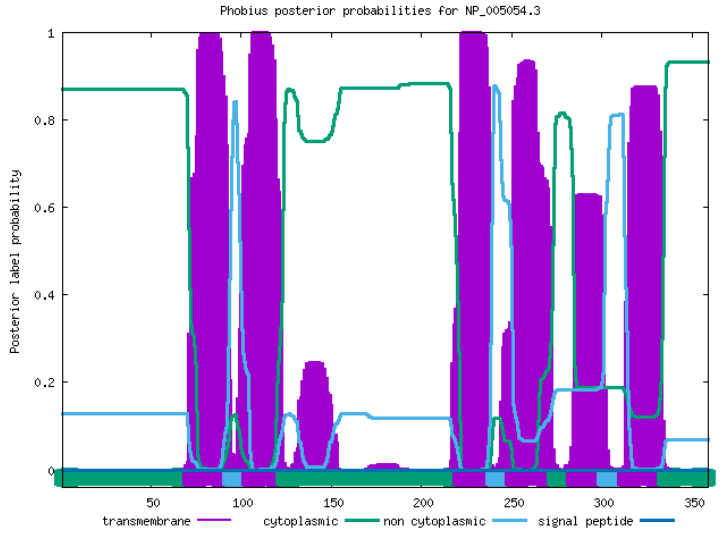
```
# NP_005054.3 Length: 359
# NP_005054.3 Number of predicted TMHs:  4
# NP_005054.3 Exp number of AAs in TMHs: 104.54198
# NP_005054.3 Exp number, first 60 AAs:  0.03484
# NP_005054.3 Total prob of N-in:        0.75464
NP_005054.3     TMHMM2.0        inside      1    70
NP_005054.3     TMHMM2.0        TMhelix     71   93
NP_005054.3     TMHMM2.0        outside     94   97
NP_005054.3     TMHMM2.0        TMhelix     98   120
NP_005054.3     TMHMM2.0        inside      121  216
NP_005054.3     TMHMM2.0        TMhelix     217  239
NP_005054.3     TMHMM2.0        outside     240  248
NP_005054.3     TMHMM2.0        TMhelix     249  271
NP_005054.3     TMHMM2.0        inside      272  359
```



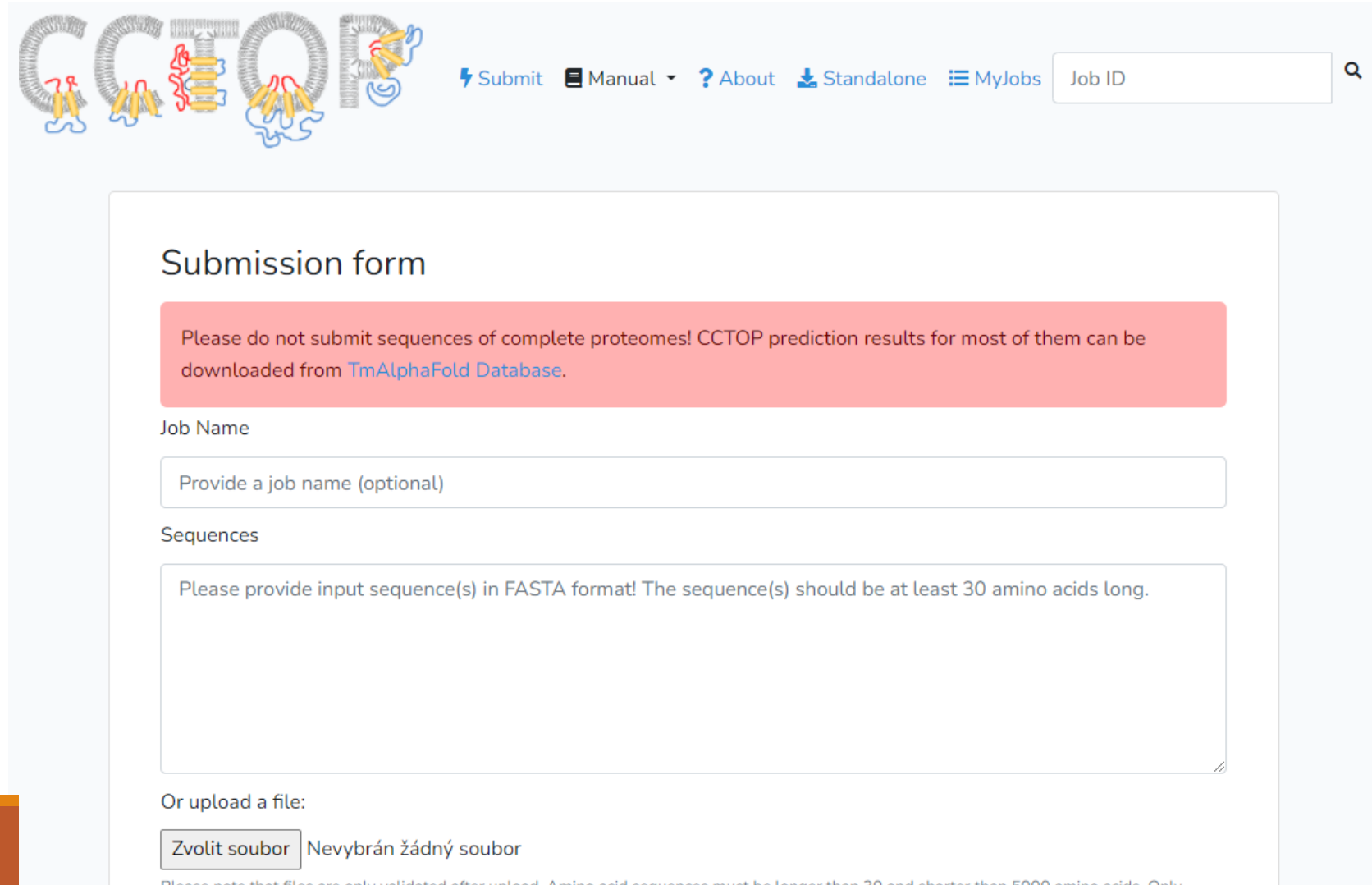
Phobius prediction

Prediction of NP_005054.3

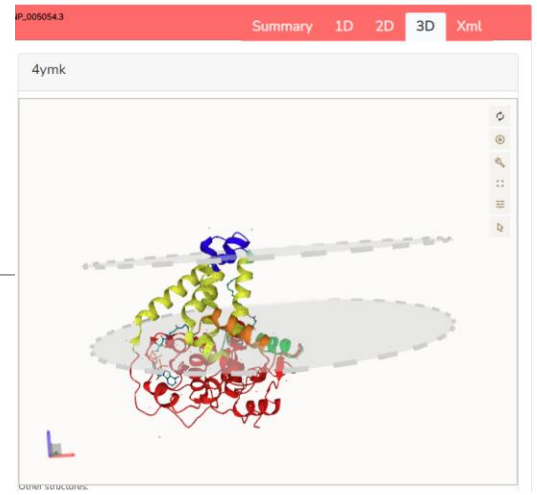
```
ID   NP_005054.3
FT   TOPO_DOM      1     71       CYTOPLASMIC.
FT   TRANSMEM     72     93
FT   TOPO_DOM     94    104       NON CYTOPLASMIC.
FT   TRANSMEM    105    123
FT   TOPO_DOM    124    221       CYTOPLASMIC.
FT   TRANSMEM    222    239
FT   TOPO_DOM    240    250       NON CYTOPLASMIC.
FT   TRANSMEM    251    273
FT   TOPO_DOM    274    284       CYTOPLASMIC.
FT   TRANSMEM    285    301
FT   TOPO_DOM    302    312       NON CYTOPLASMIC.
FT   TRANSMEM    313    334
FT   TOPO_DOM    335    359       CYTOPLASMIC.
//
```



CCTOP


Submit
Manual
About
Standalone
MyJobs
Job ID
Submission form
Please do not submit sequences of complete proteomes! CCTOP prediction results for most of them can be downloaded from TmAlphaFold Database.
Job Name
Provide a job name (optional)
Sequences
Please provide input sequence(s) in FASTA format! The sequence(s) should be at least 30 amino acids long.
Or upload a file:
Zvolit soubor
Nevybrán žádný soubor

CCTOP



Submit Manual About Standalone MyJobs Job ID

Results for job 92e87f663074d81d642456ea6e9d63a6

Download results

XML file

Proteins:

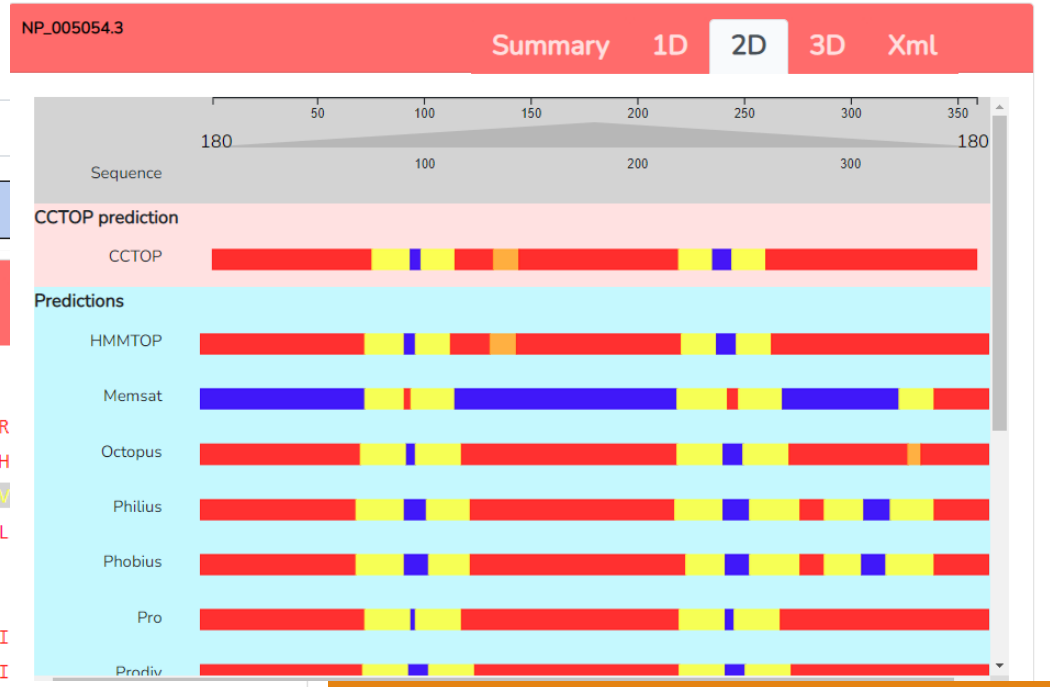
sequences: 1/1

> NP_005054.3

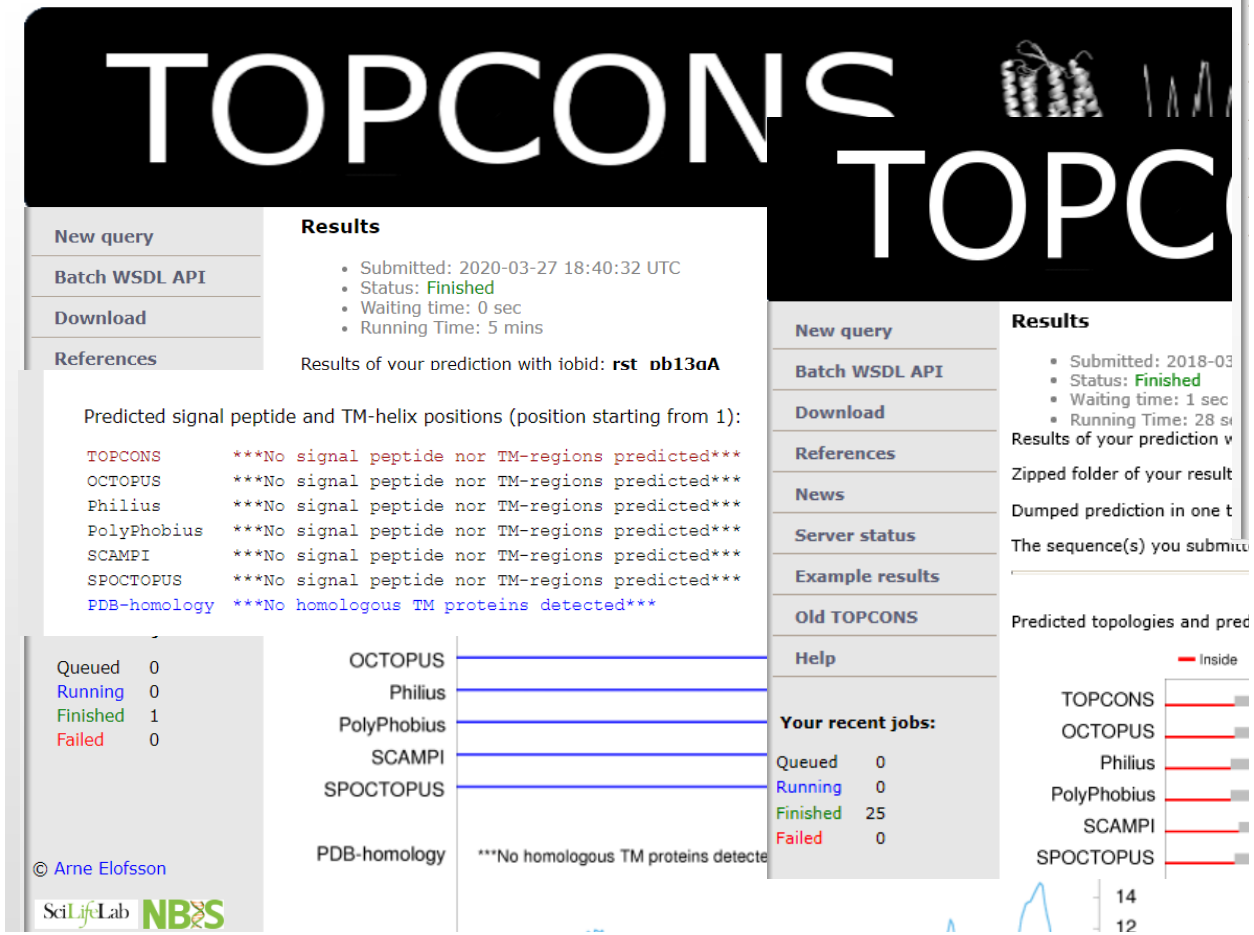
Control panel

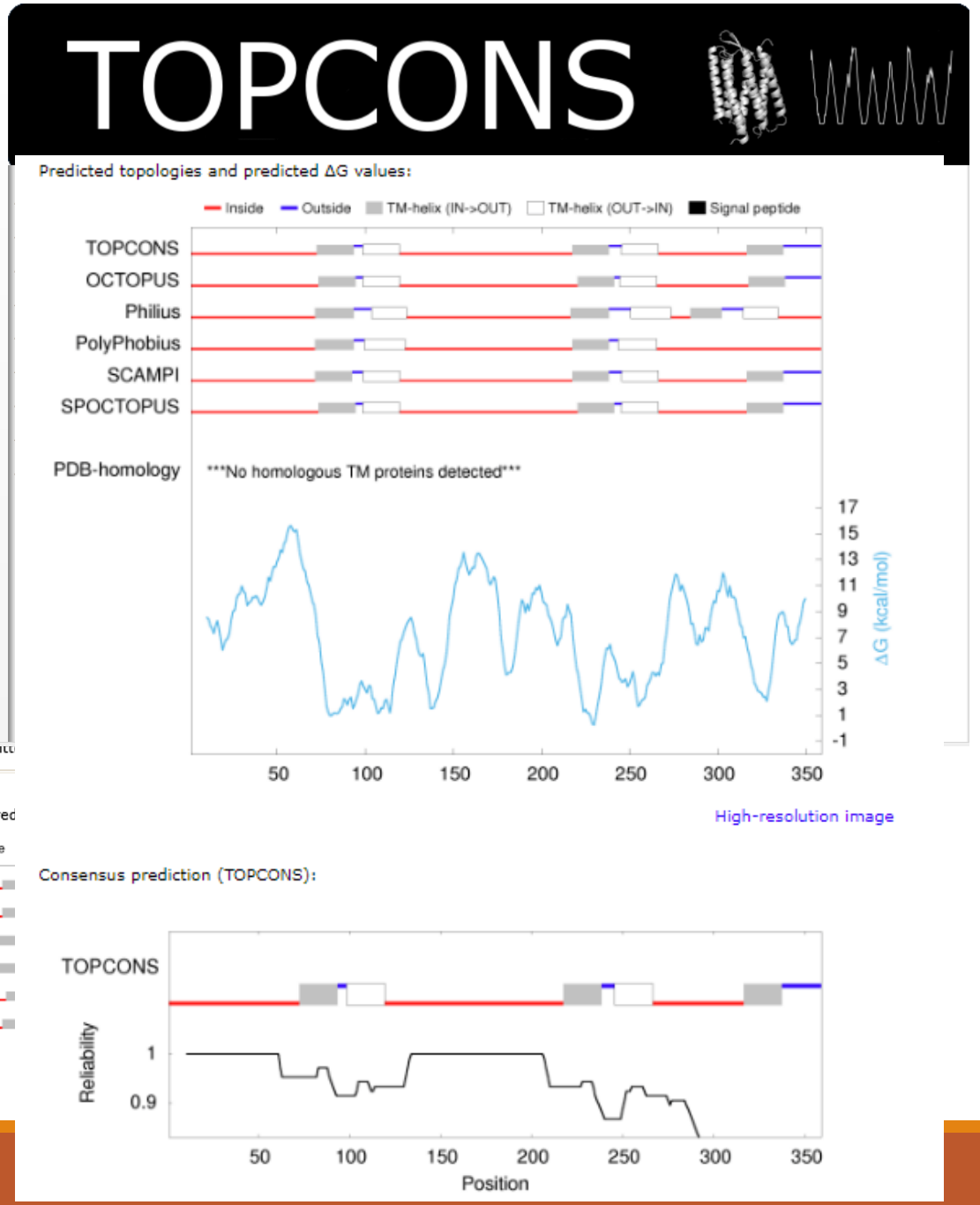
NP_005054.3 Summary 1D 2D

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



TOPCONS

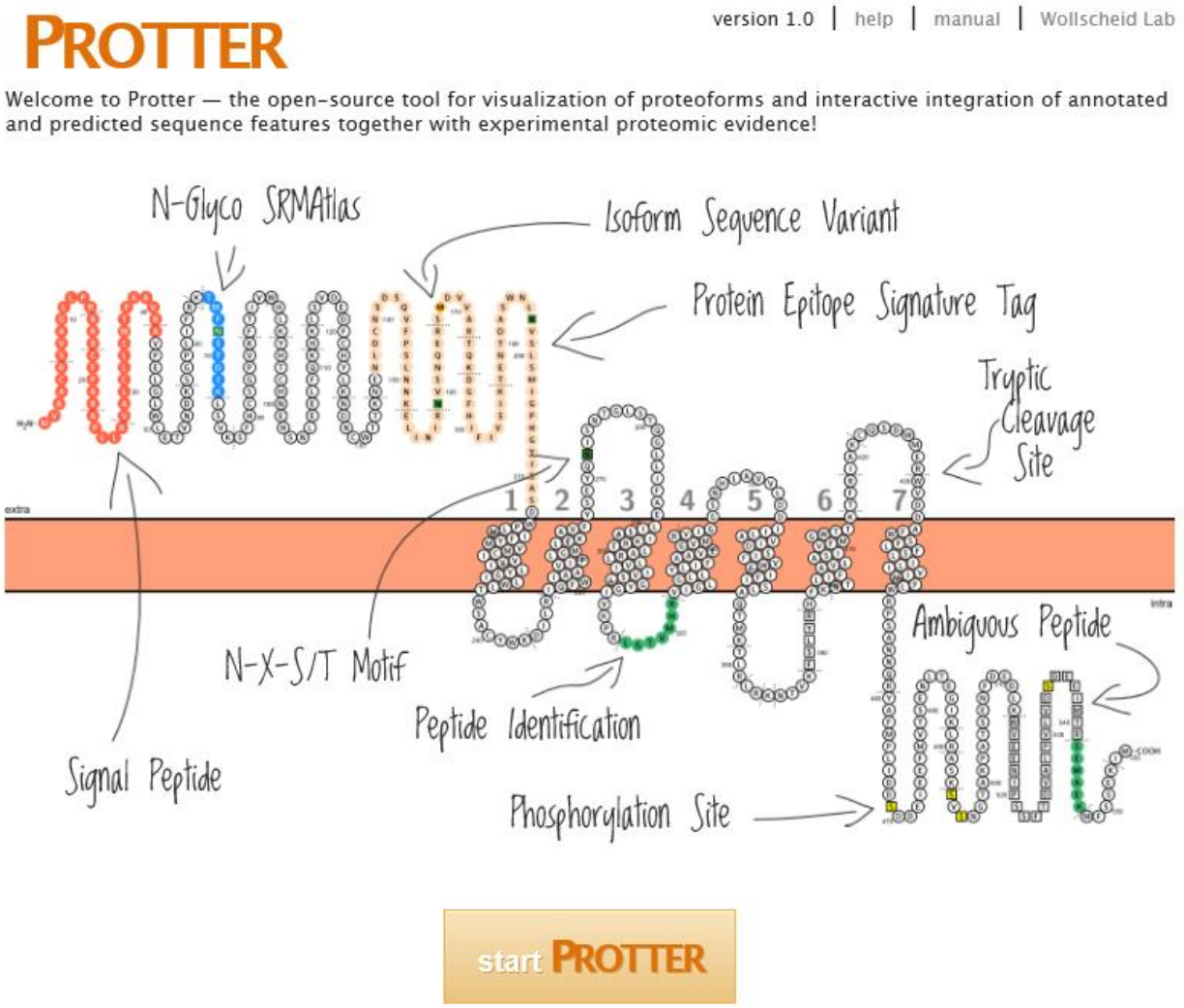

TOPCONS
New query
Batch WSDL API
Download
References
Results
Submitted: 2020-03-27 18:40:32 UTC
Status: Finished
Waiting time: 0 sec
Running Time: 5 mins
Predicted signal peptide and TM-helix positions (position starting from 1):
TOPCONS ***No signal peptide nor TM-regions predicted***
OCTOPUS ***No signal peptide nor TM-regions predicted***
Philius ***No signal peptide nor TM-regions predicted***
PolyPhobius ***No signal peptide nor TM-regions predicted***
SCAMPI ***No signal peptide nor TM-regions predicted***
SPOCTOPUS ***No signal peptide nor TM-regions predicted***
PDB-homology ***No homologous TM proteins detected***
Queued 0
Running 0
Finished 1
Failed 0
OCTOPUS
Philius
PolyPhobius
SCAMPI
SPOCTOPUS
PDB-homology
© Arne Elofsson
SciLifeLab NBIS
TOPC
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 s
Results of your prediction
Zipped folder of your result
Dumped prediction in one t
The sequence(s) you submit
Predicted topologies and pred
Inside
TOPCONS
OCTOPUS
Philius
PolyPhobius
SCAMPI
SPOCTOPUS
14
12


TOPCONS
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
17
15
13
11
9
7
5
3
1
-1
ΔG (kcal/mol)
50
100
150
200
250
300
350
High-resolution image
Consensus prediction (TOPCONS):
TOPCONS
Reliability
1
0.9
50
100
150
200
250
300
350
Position

PROTTER-obrázek!

-pracuje s daty z UniProt

(nepočítá sám)



PROTTER-obrázek!

-pracuje s daty z UniProt

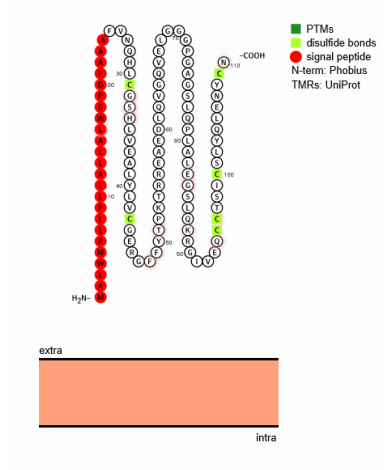
(nepočítá sám)

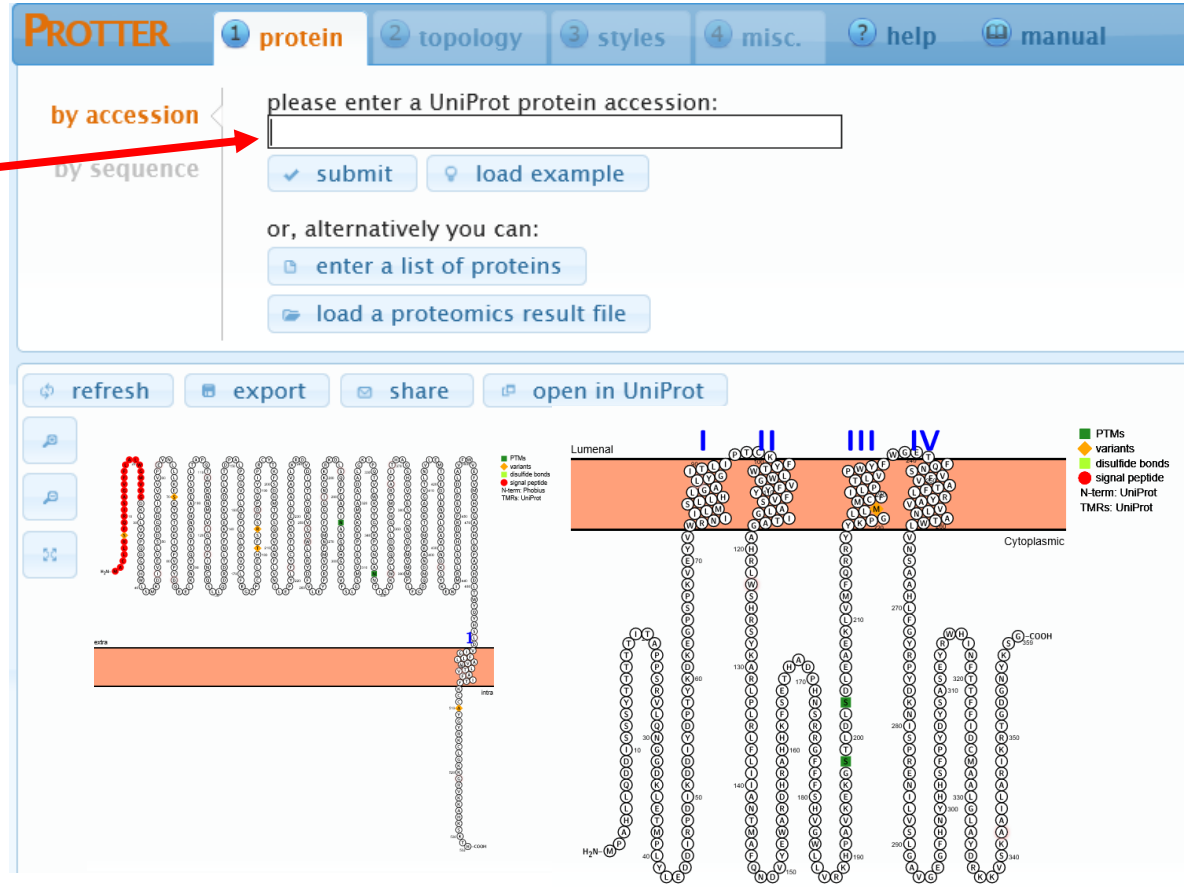
-zadává se přístupový kód:

UGT1A6 (P19224)

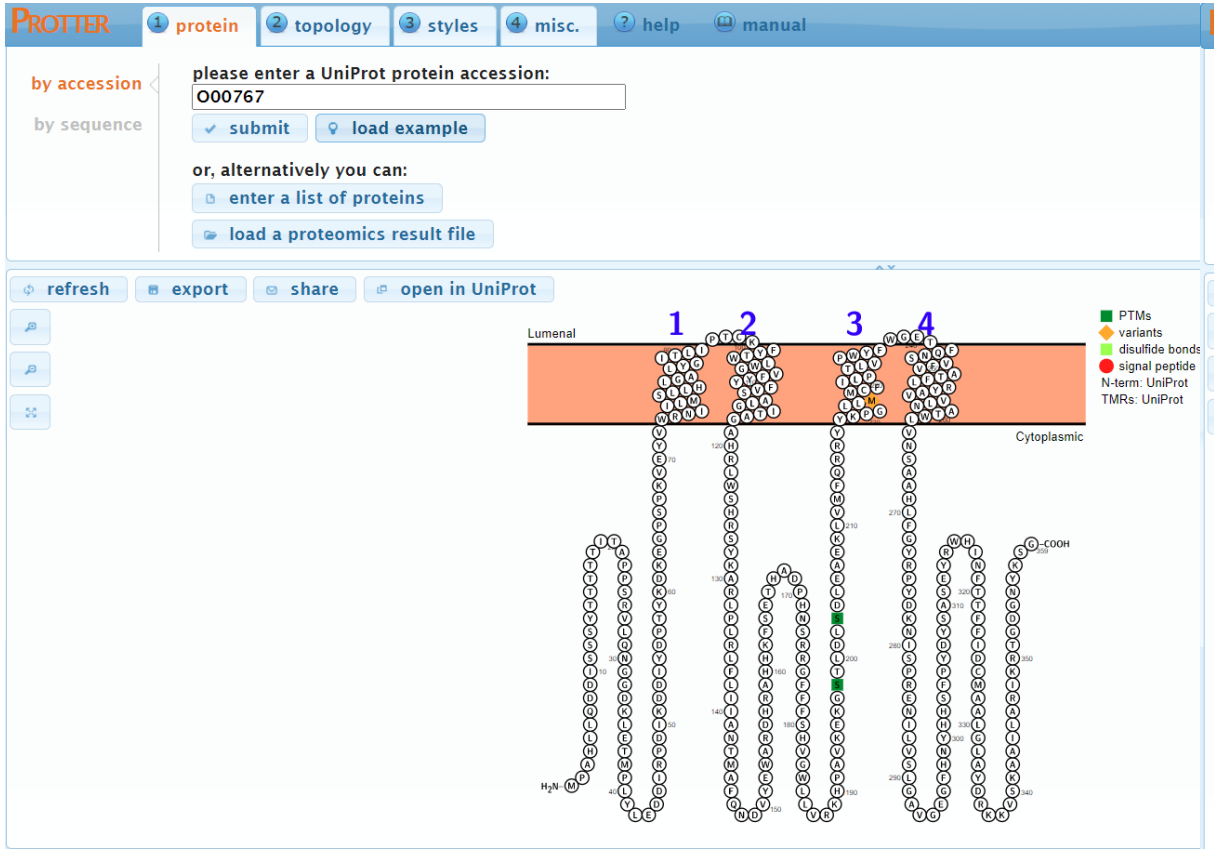
Desaturasa (O00767)

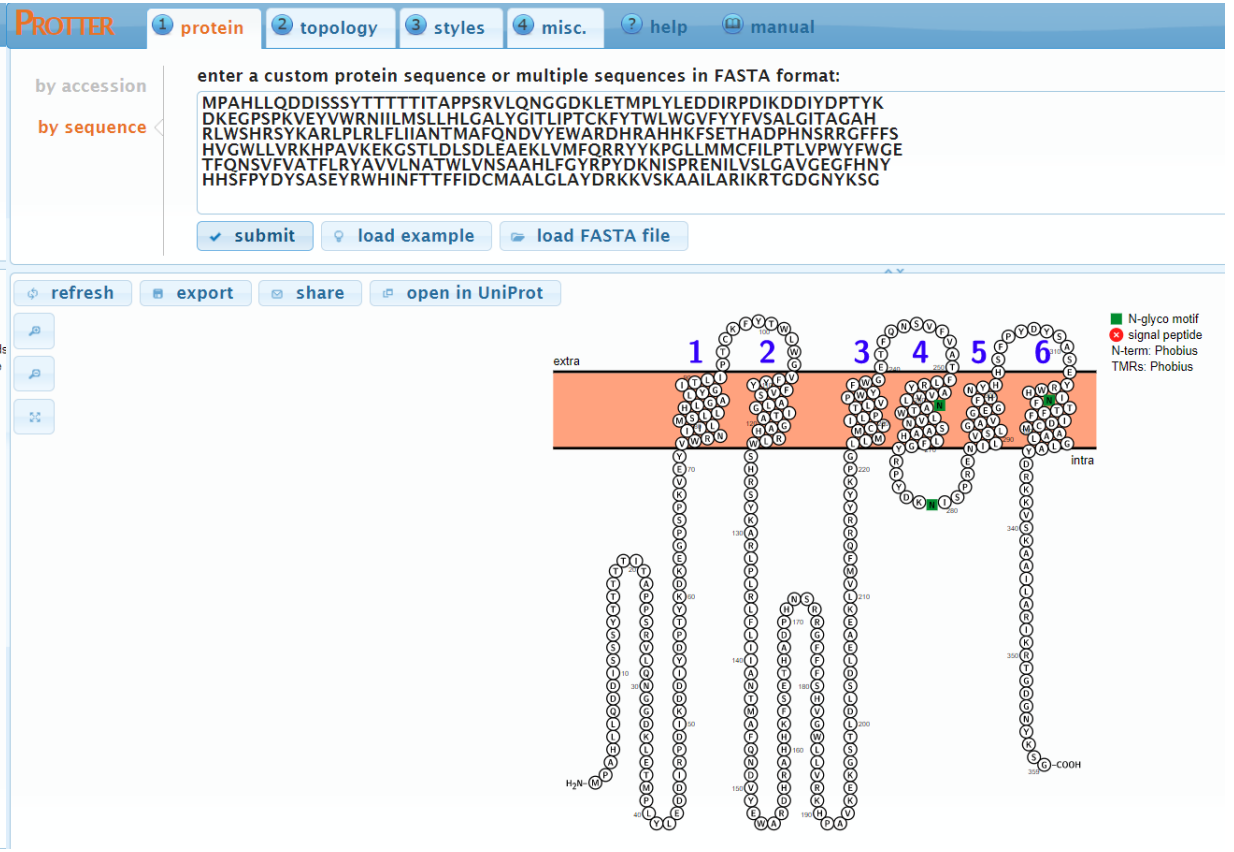
(prepro) insulin (P01308)





PROTTER-obrázek!





Predikce transmembránových úseků

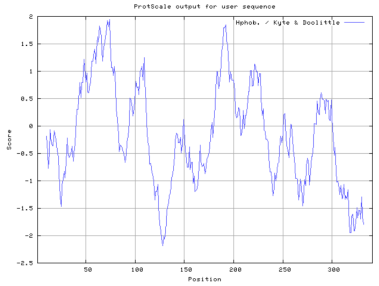
ProtScale

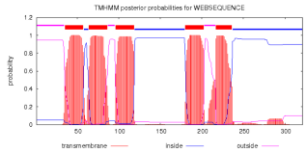
TMHMM

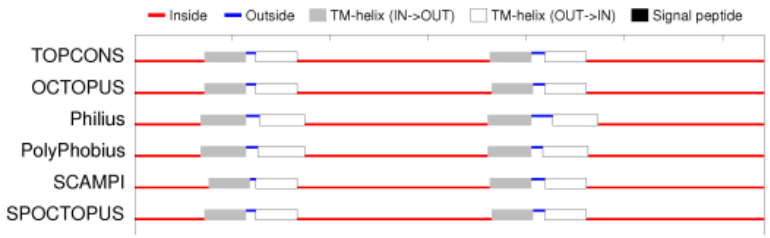
Phobius…

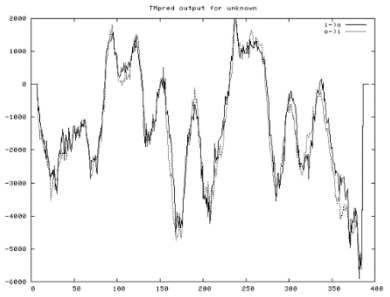
TOPCONS

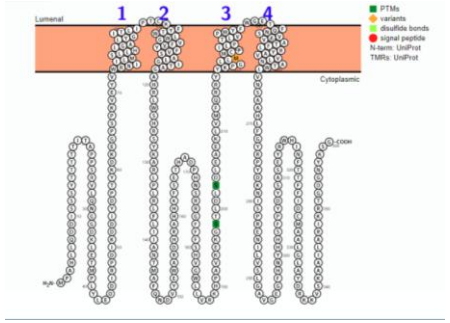
CCTOP











- **Vyzkoušet více programů!**

Praktická ukázka

- otevřete si soubor/stránku s vaší sekvencí

- otevřete si odkazy (každý v novém okně) a zjistěte zda je to rozpustný protein, případně kolik má předpovězených transmembránových úseků

- pokud nemáte TM úsek, můžete se podívat opět na sekvenci **O60656**

„Proteinová bioinformatika“

Vyhledávání AMK sekvencí

Analýza vlastností sekvencí (aminokyselinové složení, molekulová hmotnost, isoelektrický bod…)

Štěpení proteasami

Analýza hydrofobních segmentů, transmembránových úseků

Predikce sekundární struktury ⟶ Hledání známých motivů

3D-struktura, vizualizace

Vyhledání a porovnání podobných sekvencí

Evoluční příbuznost sekvencí

…

Hledání podobnosti / porovnávání (alignment)

Podobnost (similarity) x homologie

BLAST: Basic Local Alignment and Search Tool

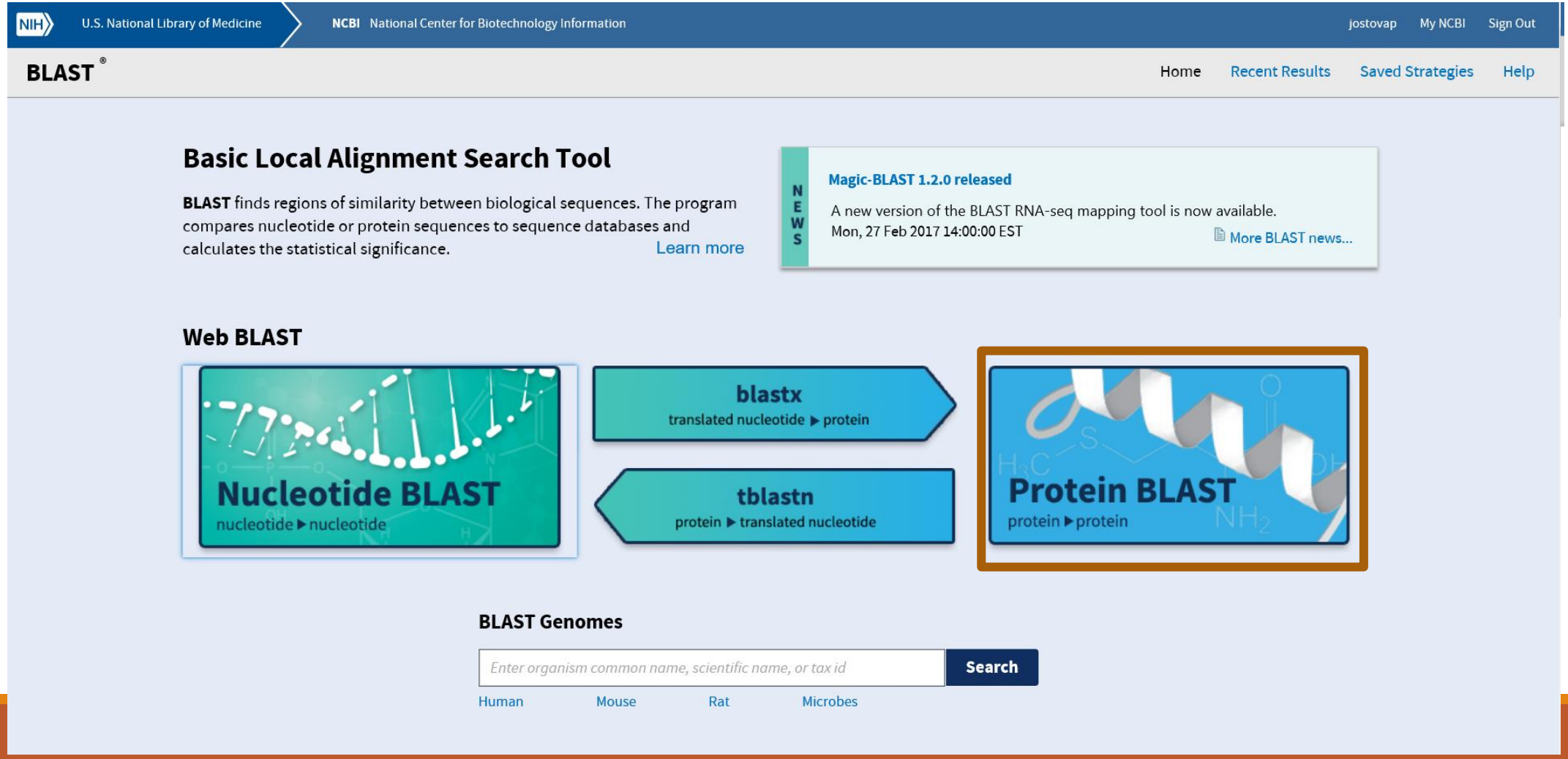
Hledá podobné sekvence z vybraných databáze pomocí krátkých „slov“ (sekvencí vytvořených ze zadané sekvence), s využitím „substituční matice“, která udává míru důležitosti rozdílných aminokyselin v zadané sekvenci a vyhledané sekvenci

„Leucin je podobnější Isoleucinu než Histidinu“

BLAST je součástí téměř všech dostupných databází (včetně genomových).

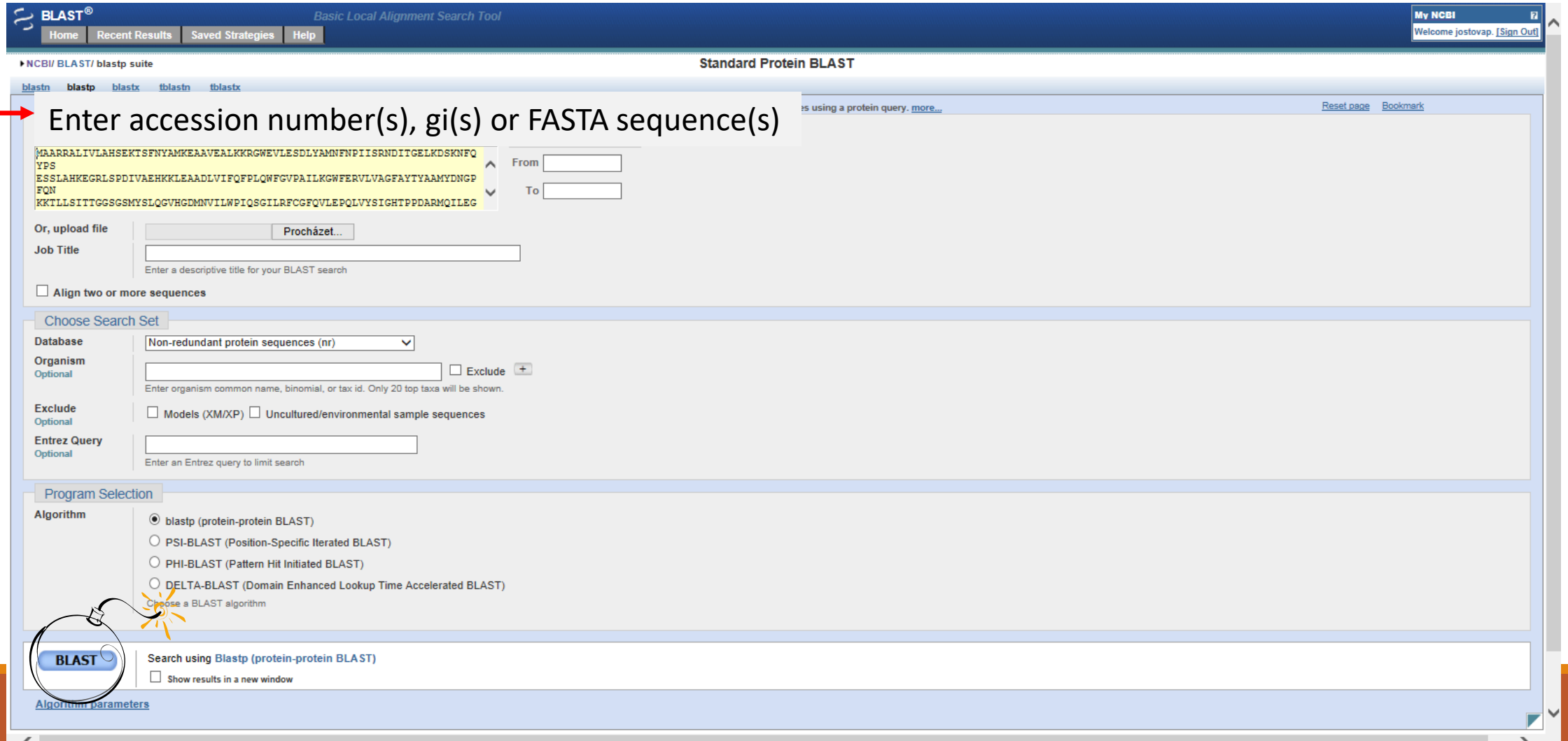
NCBI/BLASTp

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

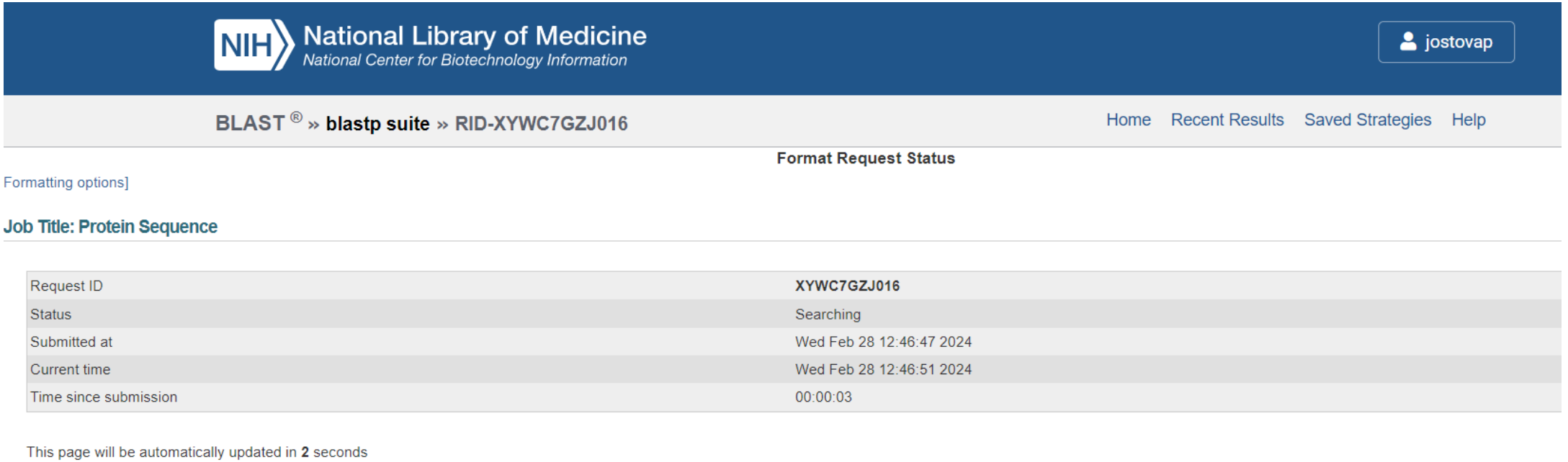


http://www.youtube.com/watchBLAST

NCBI/BLASTp

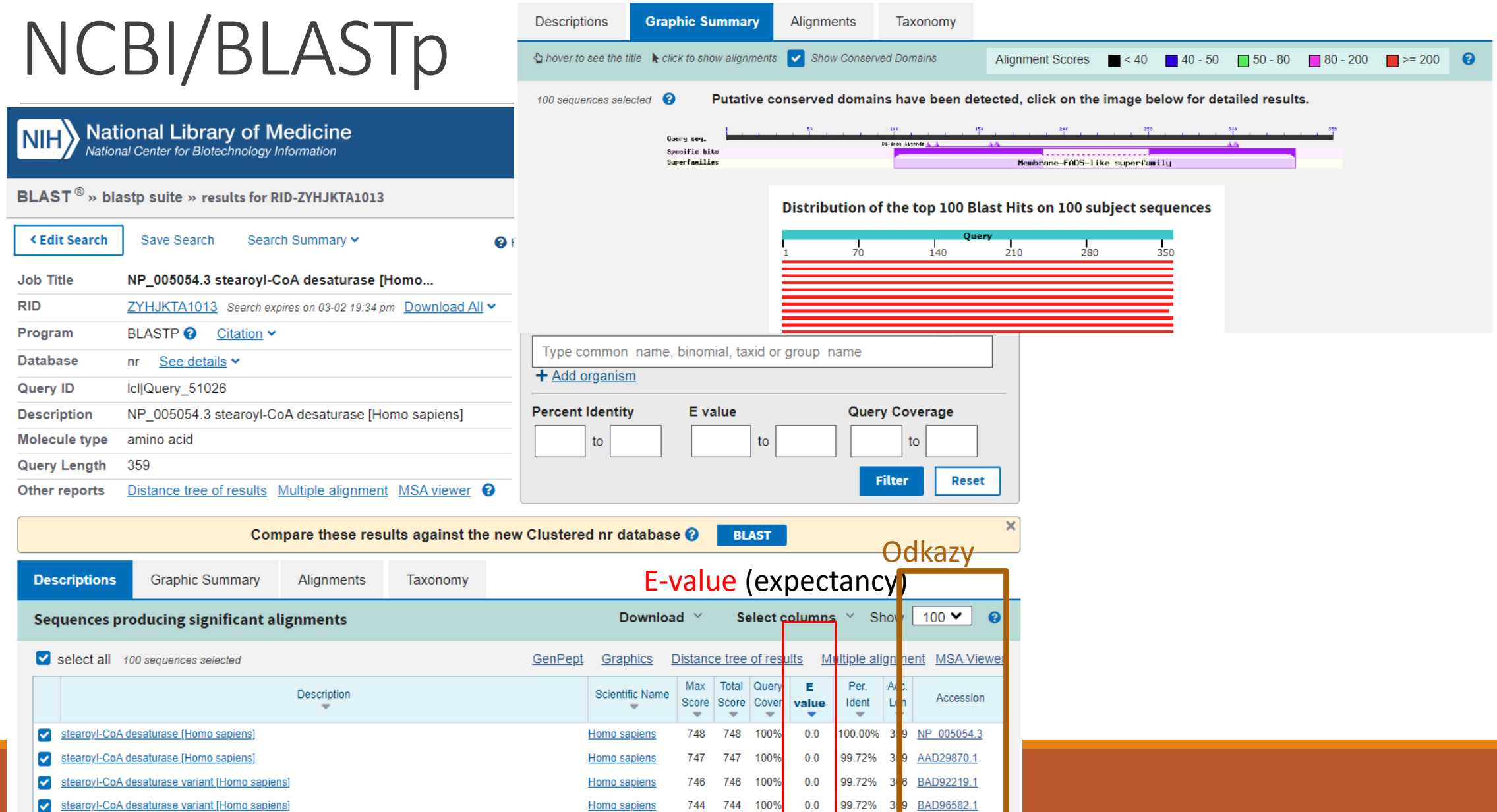


NCBI/BLASTp



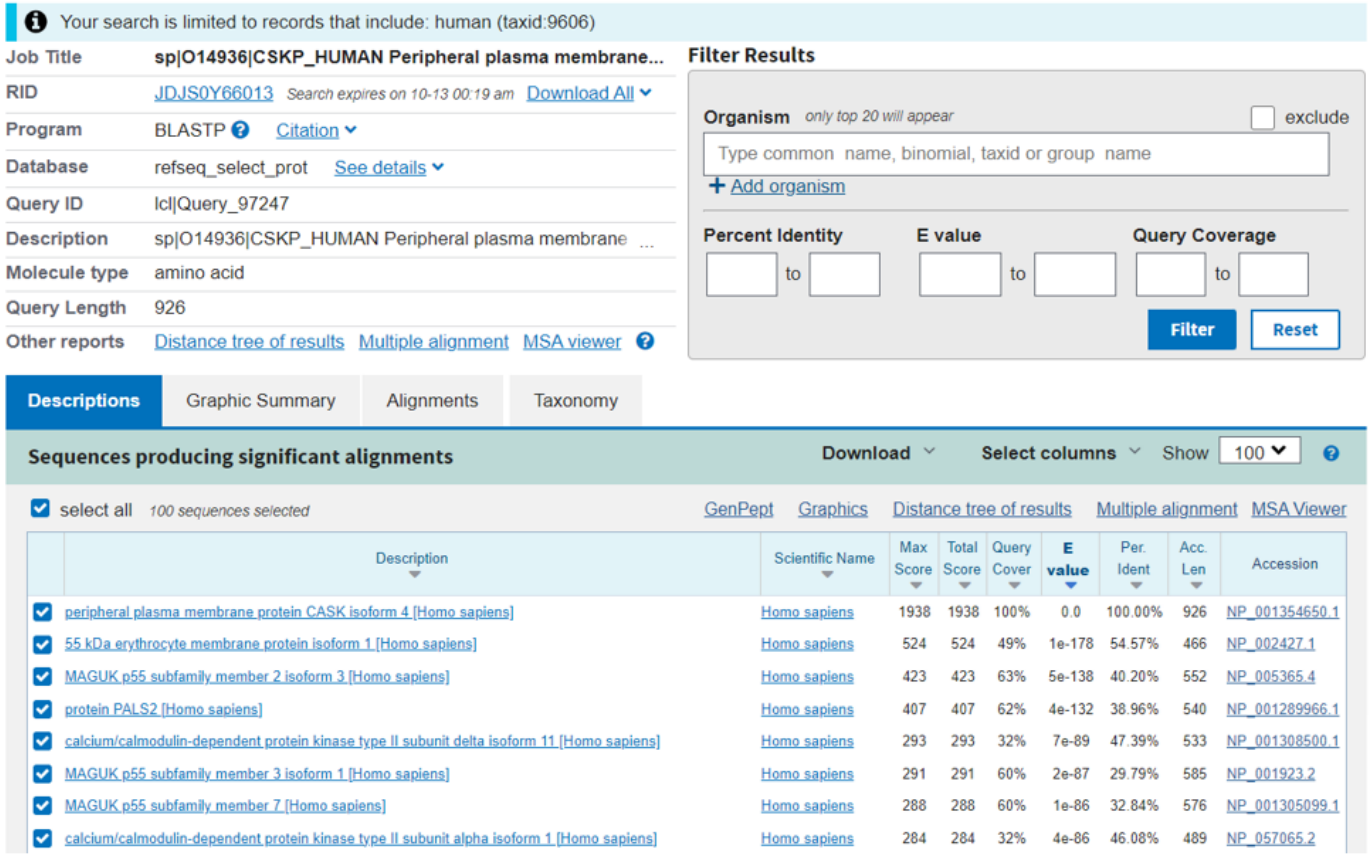
Na začátku to ukazuje toto:…..trpělivost

NCBI/BLASTp


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
NIH
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
Odkazy
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

pro DÚ-3: „větší výstřižek“ = informativnější

5) Bylo nalezeno 100 a více sekvencí



NCBI/BLASTp

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

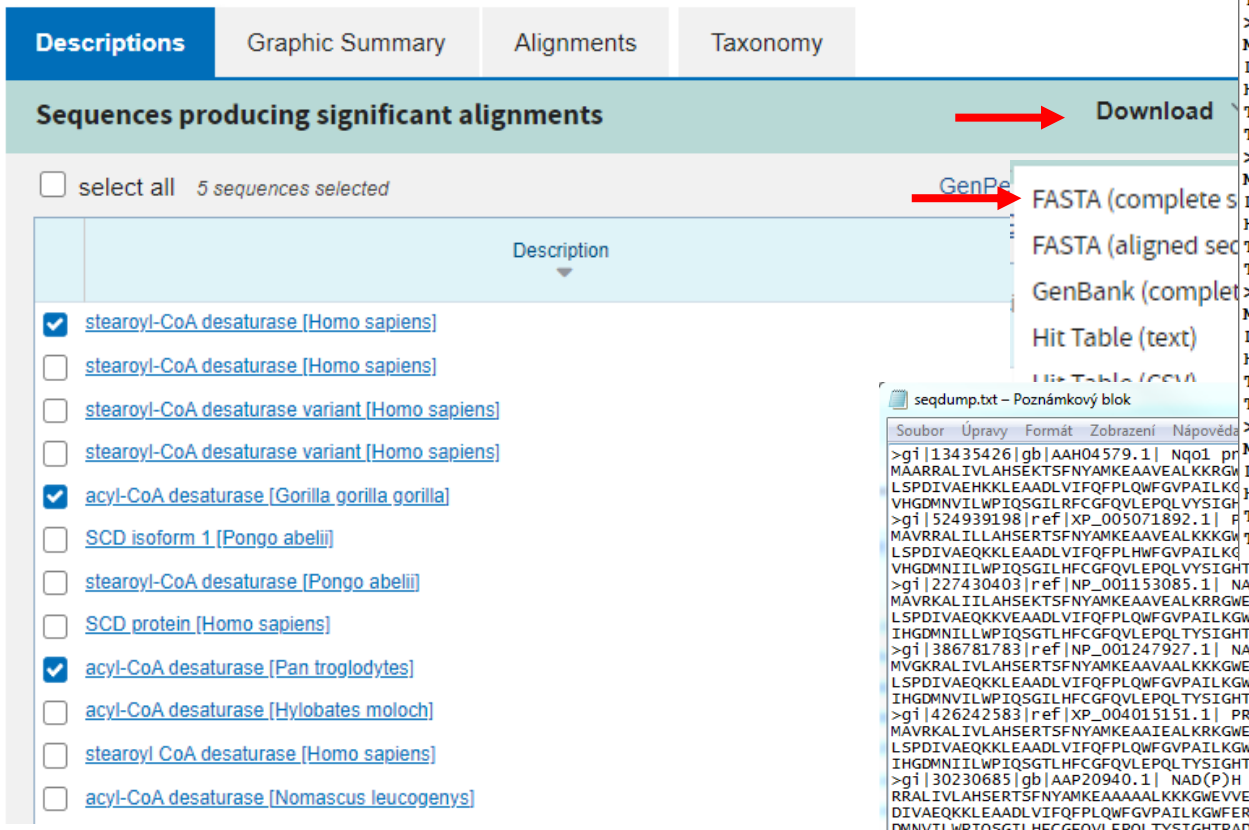
Gene - associated gene details

Genome Data Viewer - aligned genomic context

Identical Proteins - Identical proteins to NP_005054.3

→upravit uložené sekvence, tak aby v názvu byl jenom organismus

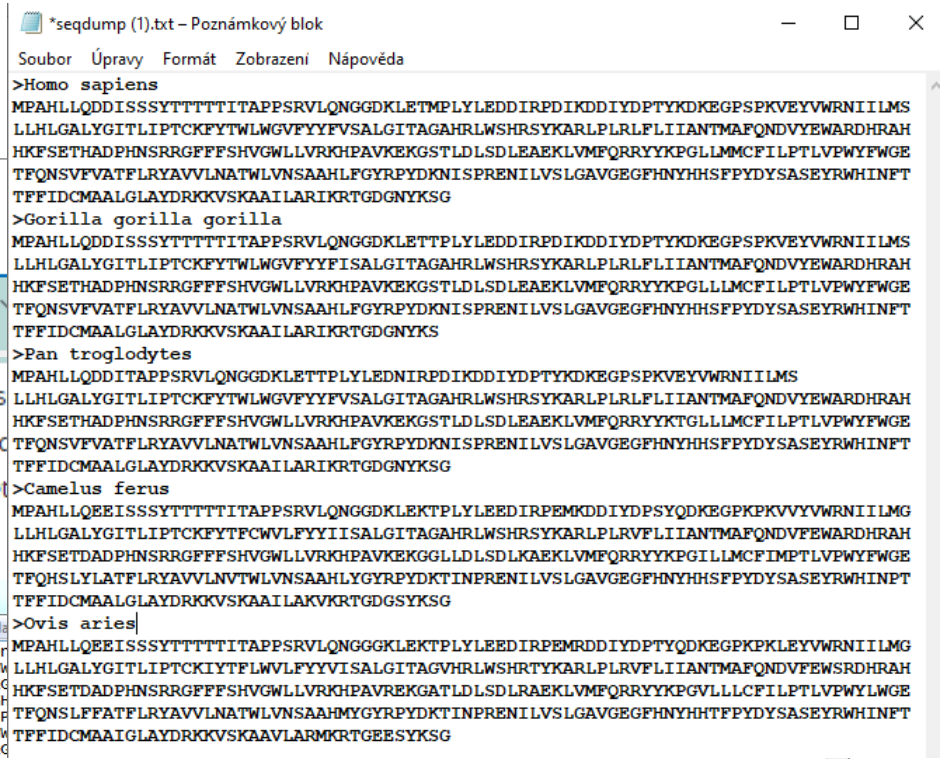
NCBI/BLASTp



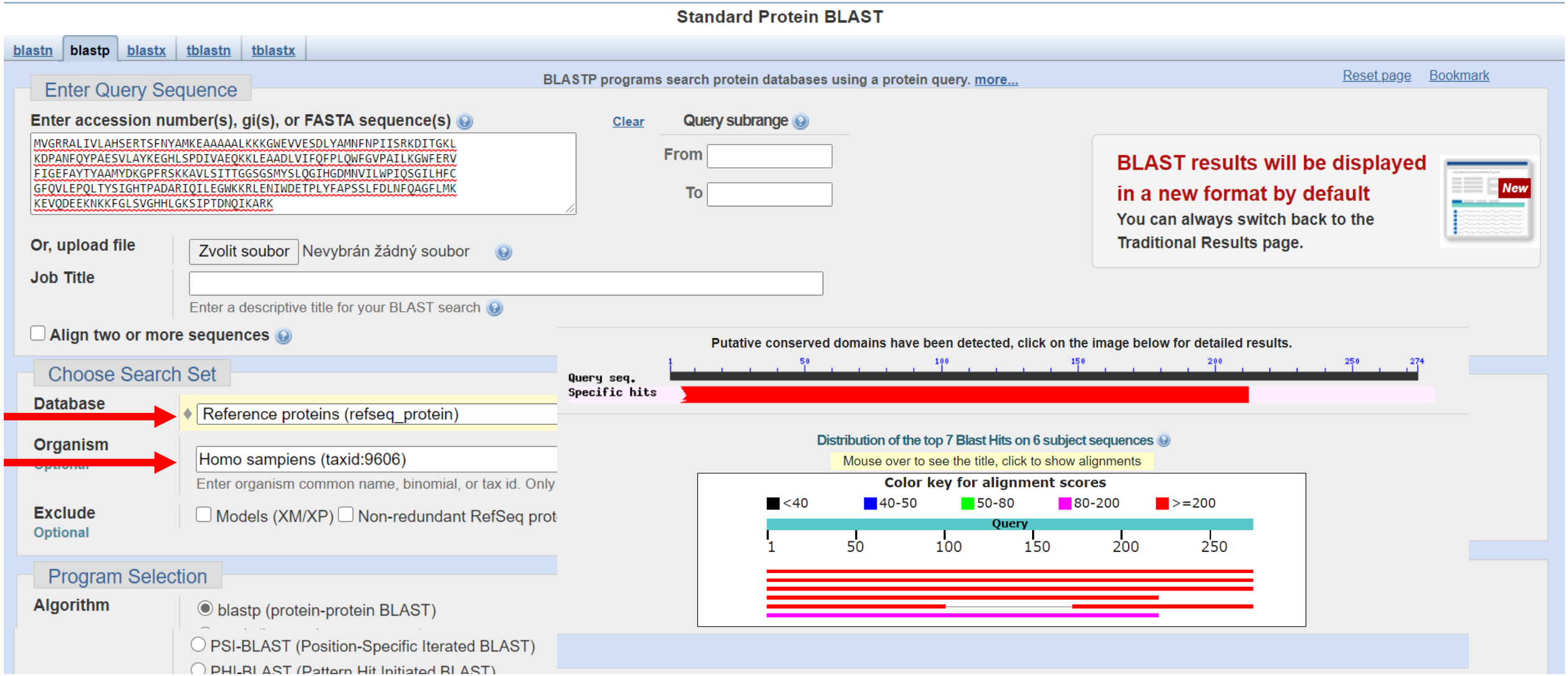
√

√

√

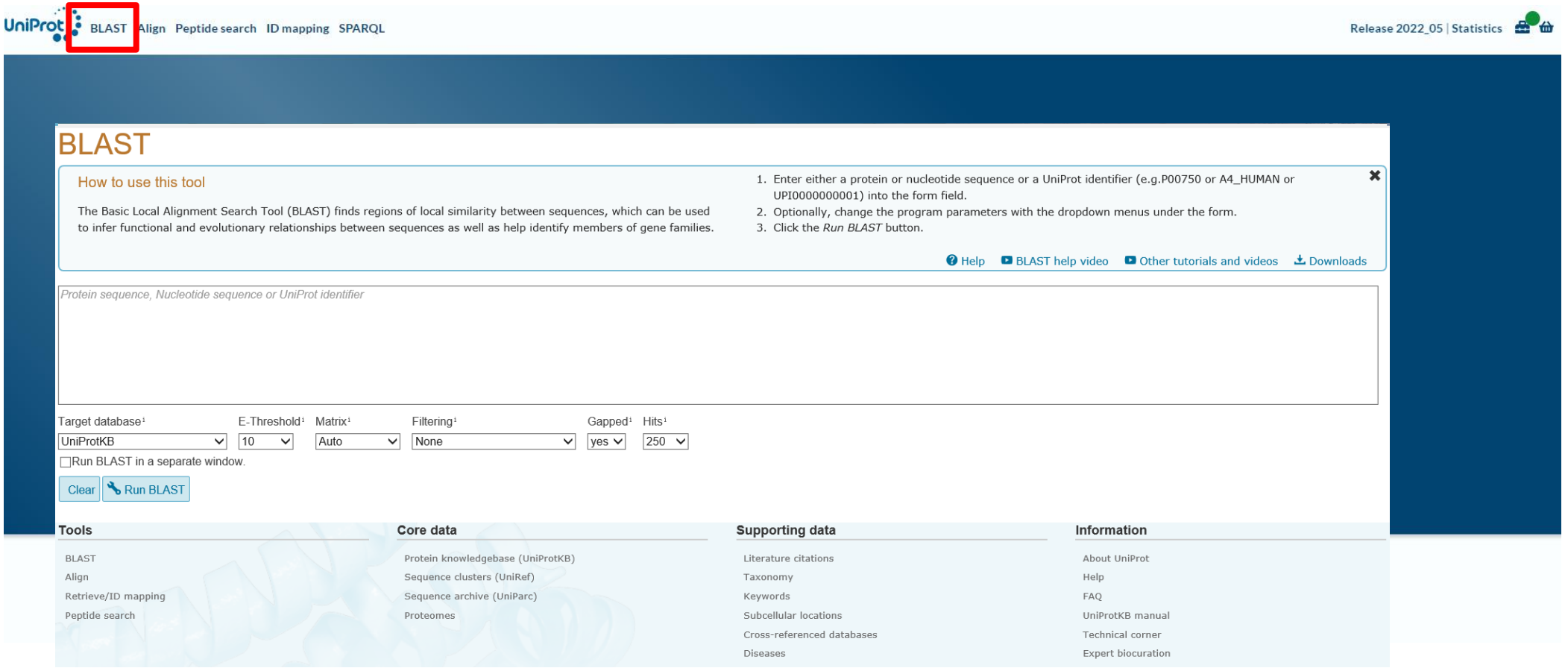


NCBI/BLAST-podobné sekvence u jednoho org.

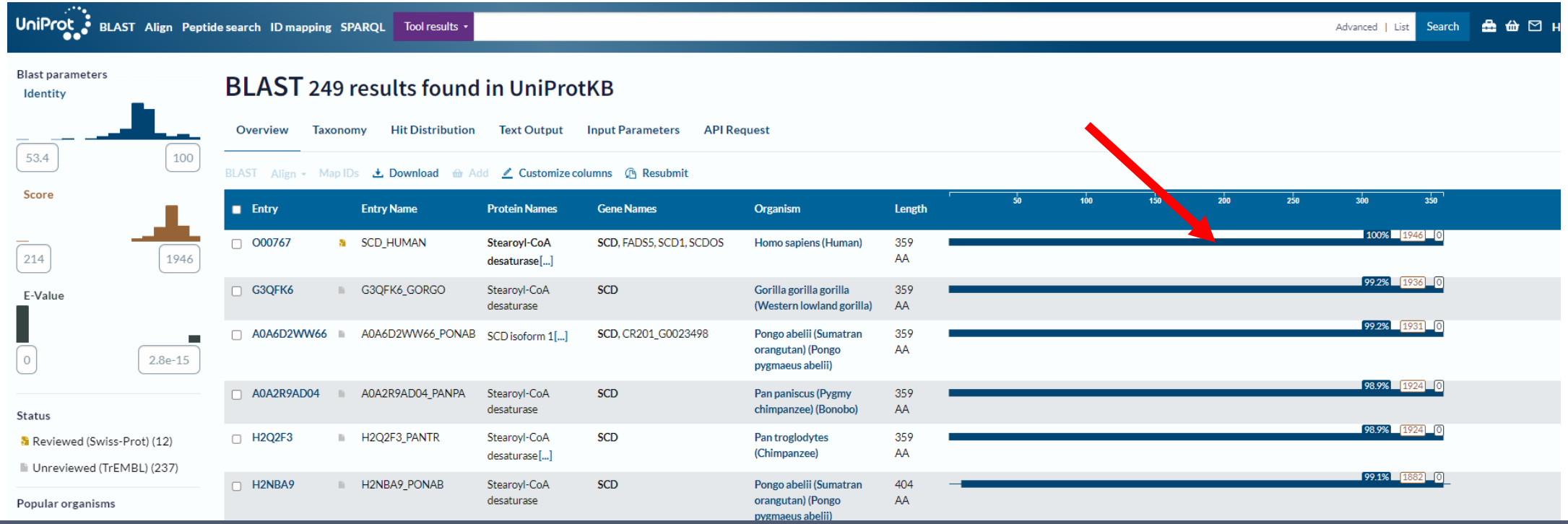


Uniprot/BLAST

https://www.youtube.com/watch?v=UPaConHNP7E



Uniprot/BLAST

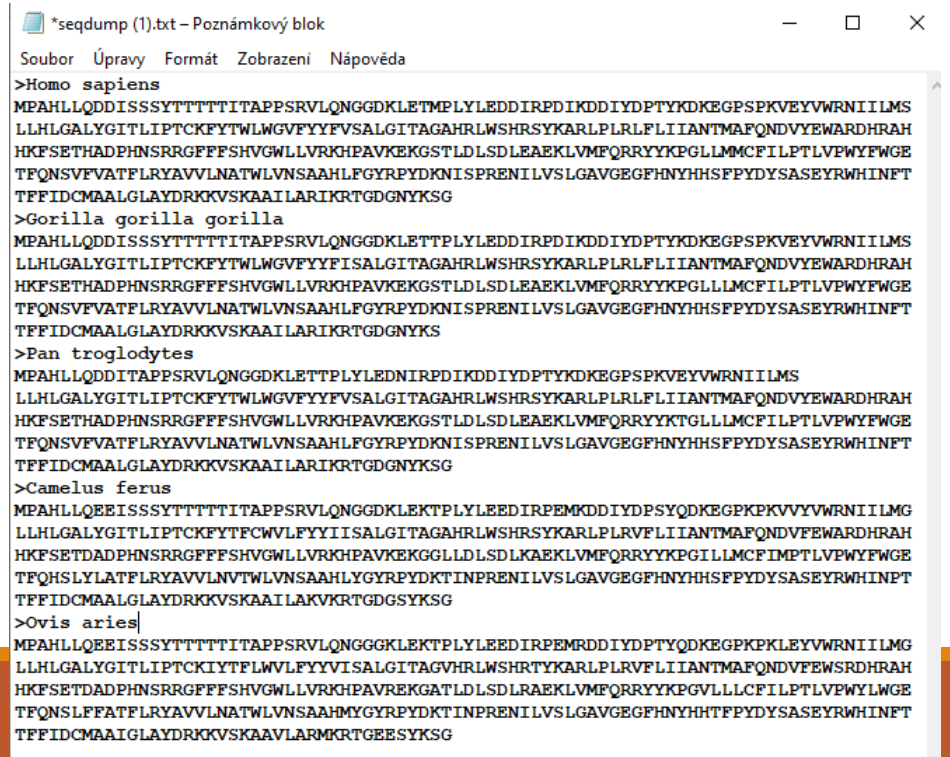


Vyzkoušejte si BLAST

Uložte si pět sekvencí z různých organismů ve FASTA formátu (do poznámkového bloku).

- identifikátor ve fasta formátu upravte jen na organismus (vhodné pro porovnávání v příští lekci)

→upravit uložené sekvence, tak aby v názvu byl jenom organismus



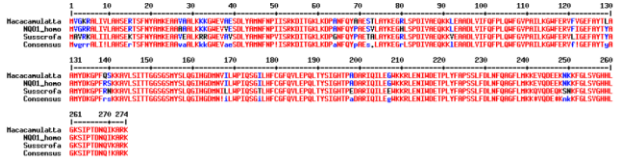
shrnutí I - Rešeršní projekt: NQO1

Vyhledejte dostupné informace o NQO1, získejte co nejvíce literárních, sekvenčních, případně i strukturních a dalších údajů o tomto genu/proteinu.

Nalezněte podobné sekvence a porovnejte je na proteinové úrovni.

- NAD(P)H:chinonoxidoreduktasa / reference
- Protein: NP_000894, P15559 / 274 AMK; 30,8 kDa; trypsin štěpí 33x (59AMK nejdelší peptid)…
- **Flavodoxinová doména**

, N-terminální signální peptid
- **Cytosolický/Extracelulární (Rozpustný) protein**
- **Podobné proteiny vyhledány, FASTA soubory uloženy, sekvence porovnány**



DÚ3: analýza proteinů

Pracujte s „vaším“ genem/proteinem/enzymem (př. NQO1)

1) Zjistěte kolikrát bude vaše **celá** sekvence štěpena trypsinem
2) Obsahuje váš protein nějaké typické motivy? (vytřihněte výstup z jednoho programu)
3) Je to transmembránový protein? (-II-)
4) Má signální peptid? (-II-)
5) Vyberte pět podobných sekvencí (vyhledaných BLASTem) z jiných organismů – uložte si sekvence

➢ vše zpracujte do wordu/pdf

Příště:

Téma 4/10

Proteinová bioinformatika III

Cíle:

Student bude schopen porovnat sekvence dvou či více proteinů, určit míru identity, zobrazit jednoduchý fylogenetický strom vybraných proteinů a najít 3D strukturu a interakční partnery proteinu…