

Introduction to applied bioinformatics

PETRA MATOUŠKOVÁ

2025/2026

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 trypsinogen)
 - 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!)

the cleavage of the protein. 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 [\[references\]](#) / [\[documentation\]](#) 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 97 100 104 106 107 113 116 117 121 125 127 129 133 138 145 156 158 162 165 169 177 178 179 182 185 189 191 195 205 208 212 216 221 222 223 228 229 231 233 237 238 239 252 254 258 259 266
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|
      ||      |      |      ||      |
1  MVGRRALIVLAHSERTSFNYAMKEAAAAAALKKGWEVVESDLYAMNFIISRKDITGKL  60
   -----+-----+-----+-----+-----+-----+-----+

```

```

      Tryps      Tryps      Tryps      Tryps
      Tryps|    Tryps|    Tryps|    Tryps|
      |      |      ||      |
61 KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV  120
   -----+-----+-----+-----+-----+-----+

```


Prediction of proteases cutting

 **ExPASy**
Bioinformatics Resource Portal

PeptideCutter

Home | [Contact](#)

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'):

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:

☒ 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 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)	MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWVESDLYAMNFNPIISRKDITGKLKDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERVFIGEFAYTYAAMYDKGPFRRSKKAVLSITTGGSGSMYSLQGIHGDMNVILWPIQSGILHF	179	19997.201
248	Enterokinase	CGFQVLEPQLTYSIGHTPADARIQILEGWKKRLNIWDETPLYFAPSSLFDLNFQAGFLMKKEVQDEEK	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.

1 MVGRRALIVLAHSERTSFNYAMKEAAAAALKKKGWVESDLYAMNFNPIISRKDITGKL 60

61 KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVIFQFPLQWFGVPAILKGWFERV 120

Prediction of proteases cutting

 **ExPASy**
Bioinformatics Resource Portal

PeptideCutter

Home | [Contact](#)

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

☐ 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 mass](#) 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	AVLSITTGSGSGMSYSLQGIHGDMNVILWPIQSGILHFCGFGVLEPQLTYS	59	6287.190
209	Trypsin	IQILEGWK	8	986.179
210	Trypsin	K	1	146.189
211	Trypsin	R	1	174.203
240	Trypsin	LENIWDETPLYFAPSSLFDLNFQAGFILMK	29	3407.885
241	Trypsin	K	1	146.189
248	Trypsin	EVQDEEK	7	875.888
250	Trypsin	NY	2	260.202

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 format)

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

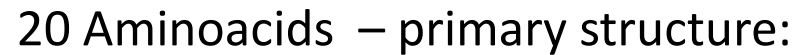
➔ **Finding 3-D structure and the domain organization of proteins**

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

...



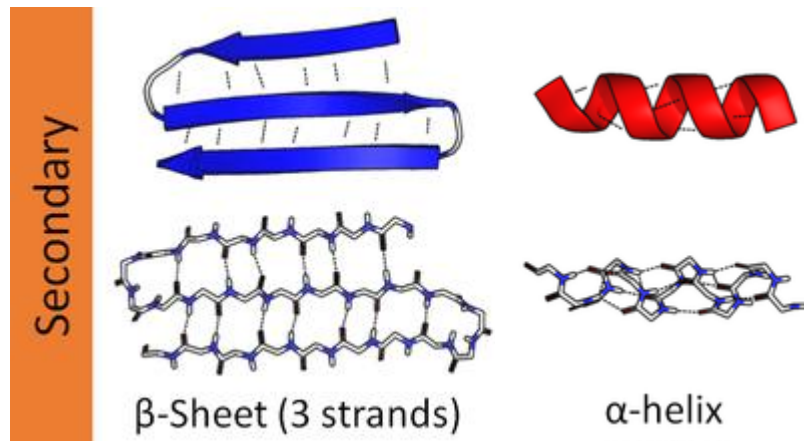
Secondary structure
Tertiary structure
Quaternary structure



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

Protein domain

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

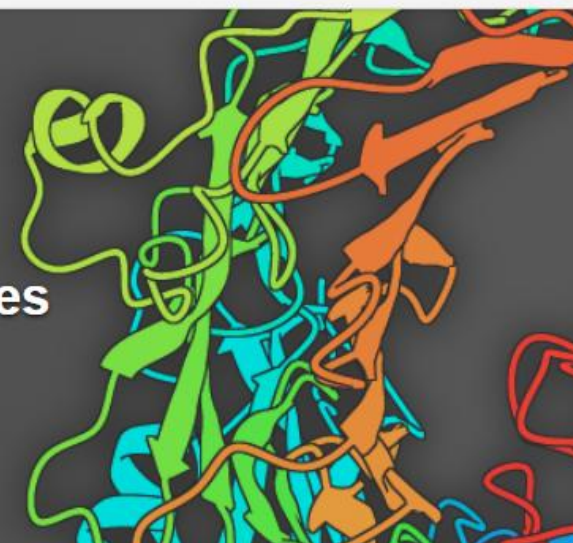


CATH / Gene3D v4.4

151 million protein domains classified into 6,573 superfamilies

Search by keywords, PDB code, GO term, etc

Search



Putative CATH annotations for predicted structural domains in AlphaFold DB are available in [The Encyclopedia of Domains \(TED\)](#). Annotations for the 21 model organisms predicted by AlphaFold (v2) are [available to download](#) ([doi:10.1038/s42003-023-04488-9](https://doi.org/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

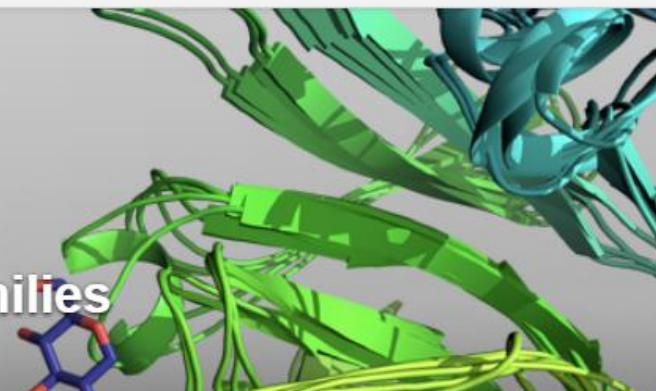
Investigate the function of your protein

[Find out more](#)

[Go](#)

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](#)
- [Download Gene3D Data](#)
- [Compare Genomes](#)
- [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	01.07.2010	about 15 hours ago

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

	Gene3D v2.1
Protein Sequences	82,665,384
CATH Domain Predictions	151,013,797

CATH – try it

- look for a superfamily of your protein

NQO1:

CATH Superfamily 1.20.1440.230

NADH-ubiquinone oxidoreductase 51kDa subunit, iron-sulphur binding domain

A Up-down Bundle
1.20

CATH ID	1.20
Topologies	104
Superfamilies	788
Domains	29676
Example Domain	2ynzC01 [PDB]



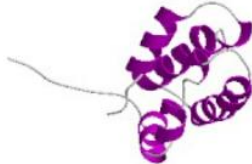
T de novo design (two linked rop proteins)
1.20.1440

CATH ID	1.20.1440
Superfamilies	33
Domains	253
Example Domain	2qsbA00 [PDB]



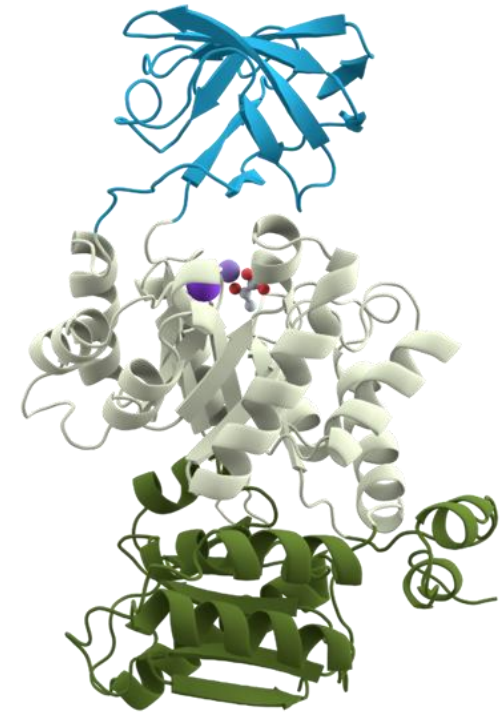
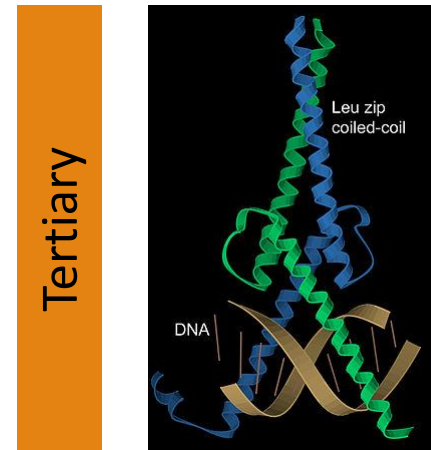
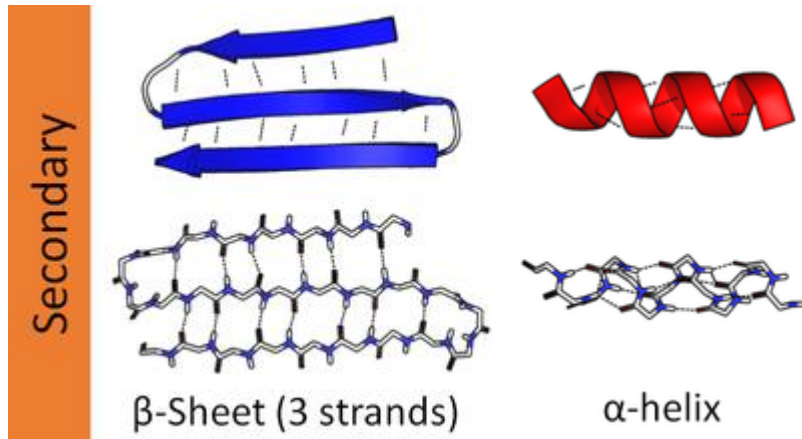
H NADH-ubiquinone oxidoreductase 51kDa subunit, iron-sulphur binding domain
[Go to Superfamily >](#)

CATH ID	1.20.1440.230
Domains	19
Example Domain	3i9v104 [PDB]



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

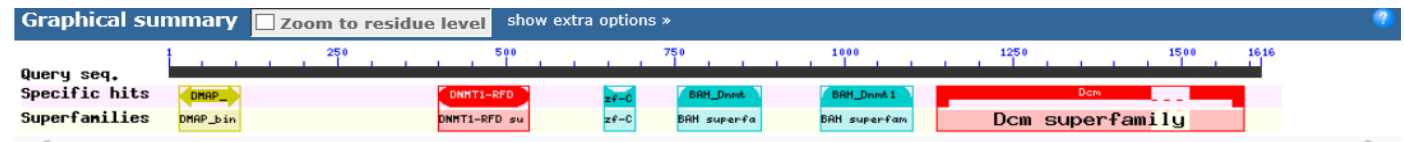


Conserved domain search

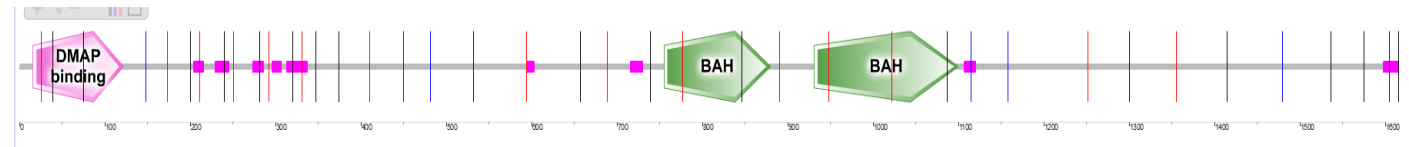
SEQUENCE $\leftrightarrow$ STRUCTURE $\leftrightarrow$ FUNCTION

- Conserved domain databases:

NCBI/CDD

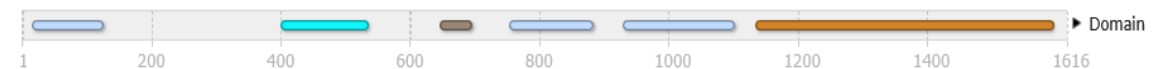


SMART



EMBL/InterPro

Domains and repeats



Pfam

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



Conserved domain search - CD (NCBI)

 [Resources](#)  [How To](#) 

[jostovap](#) [My NCBI](#) [Sign Out](#)

Conserved Domains

Conserved Domains  p15559 

Search

Advanced

Help

Conserved Domains and Protein Classification

[OVERVIEW](#) [SEARCH](#) [HOW TO](#) [HELP](#) [NEWS](#) [FTP](#) [PUBLICATIONS](#) [DISCOVER](#)

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)

NIH National Library of Medicine
National Center for Biotechnology Information

on near the bottom-right of your screen to let us know about
previous user interface will be available [here](#) for a short pe

CD-Search

Enter Query

Enter accession, GI or FASTA sequence [®] :

MVGRRALIVLAHSEKTSFNAMKEAAAAAALKKGWEVV
ESDLYAMNFPNPISRKDITGKL
KDPANFQYPAESVLAYKEGHLSPDIVAEQKKLEAADLVF
QFPLQWFGVPAILKGWFERV
FIGEFAYTYAAMYDKGPFRRSKKAVLSITGGSGSMYSLQG
IHGDMNVILWPIQSGILHFC
GFQVLEPQLTSIGHTPADARIQILEGWKKRLNIWDETPL
YFAPSSLFDLNFQAGFLMK
KEVQDEEKNKKFGLSVGHHLGKSIPTDNQIKARK

Submit

Reset

CD-Search > Navigate results

Structure Home 3D Macromolecular Structures CDART Help

Conserved domains on
[\[1cl|1|1cl|SEQMD5_24.9bc4fcdaec666c91fefe429d3f89a1\]](#)
NAD(P)H dehydrogenase [quinone] 1 isoform a
[Find similar domain architectures](#)

Protein classification: *NAD(P)H-dependent oxidoreductase* (ArchID 10006206)
NAD(P)H-dependent oxidoreductase which catalyzes the reduction or oxidation of a substrate coupled to the oxidation or reduction, respectively, of a nicotinamide adenine dinucleotide cofactor NAD(P)H or NAD(P)+
Attributes:

PubMed: 25372605|7568029

EC: 1.6.99.-|1.-.-.-

Gene Ontology: GO:0009055|GO:0050136|GO:0008753|GO:0010181

CATH: 3.40.50.360

SCOP: 3001217

Viewing Concise

H-zoom

Hide features

Download the graphic summary

Query
Specific hits
Superfamily arch.

MdaB
MdaB

1 20 40 60 80 100 120 140 160 180 200 220 240 260 274

Name	Accession	Description	Interval	E-value
MdaB	COG2249	Putative NADPH-quinone reductase (modulator of drug activity B) [General function prediction only]	5-216	4.97×10^{-61}

Conserved domain search - SMART

The screenshot shows the SMART website interface. At the top, there is a logo for SMART and a navigation bar with links: HOME, SETUP, FAQ, ABOUT, GLOSSARY, WHAT'S NEW, and FEEDBACK. A search bar is located in the top right corner with the text "keywords..." and a "Search SMART" button. Below the navigation bar, there is a section titled "Select your default SMART mode". This section contains two columns: "Normal mode" and "Genomic mode". Each column has a "SMART MODE:" label and a list of options: "NORMAL", "GENOMIC", "Simple", "Modular", "Architecture", "Research", and "Tool". The "NORMAL" option is highlighted in the "Normal mode" column, and the "GENOMIC" option is highlighted in the "Genomic mode" column. A blue box is drawn around the "Normal mode" column, and a red circle is drawn around the "NORMAL" option in the "Genomic mode" column. Below the selection section, there is a message: "Click on the images above to select your default mode." and a note: "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](#)." At the bottom, there is a footer with the SMART logo and a list of references: "Schultz et al. (1998) Proc. Natl. Acad. Sci. USA 95, 5857-5864" and "Letunic et al. (2004) Nucleic Acids Res 32, D142-D144". The "SETUP" link in the footer is circled in red.

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

HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK

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 are 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 will 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	SMART MODE: NORMAL GENOMIC
Simple Modular Architecture Research Tool	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):

SMART

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

HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK

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

SMART MODE:
NORMAL
GENOMIC

S
M
A
R
T
Simple
Modular
Architecture
Research
Tool

keywords...
Search SMART

HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK

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?

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](#)

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.

Architecture query

Resetovat

You can try an [Advanced Query](#) if you're familiar with SQL.

Conserved domain search - SMART

SMART MODE: **Simple**

SMART SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK

keywords... 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.

+ = - Introns SAVE Alternative representations: 1 / 2 << >>



0 100 200

Information Architecture Interactions Pathways PTMs Orthology

Length 274 aa

Source database [UniProt](#)

Identifiers NQO1_HUMAN, 9606.ENSPO00000319788, P15559, ENSPO00000319788.5, ENSPO00000319788, 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

 [SETUP](#) [FAQ](#) [ABOUT](#) [GLOSSARY](#) [WHAT'S NEW](#) [FEEDBACK](#)

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.

Alternative representations: 1 / 2



0 100 200

[Information](#) [Architecture](#) [Interactions](#) [Pathways](#) [PTMs](#) [Orthology](#)

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
 Ubiquitination	14
 Acetylation	3
 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

 **InterPro** Classification of protein families

[Home](#) [Search](#) [Browse](#) [Results](#) [Release notes](#) [Download](#) [Help](#) [About](#)

Job ID ⓘ
Length
Actions
Status
Expires ⓘ

iprscan5-R20230301-115158-0009-77866479-p1m
274 amino acids
 
✓ finished
Wed Mar 08 2023

Protein family membership

None predicted

Entry matches to this protein ⓘ

   Options ▾ Export ▾



IPR003680
PF02525

IPR029039
G3DSA:3.40.50.360
SSF52218

G3DSA:3.40.50.360:FF:000029 ⓘ

PTHR10204

 **InterPro** IPR003680
Flavodoxin_fold
5 - 211

Practical part

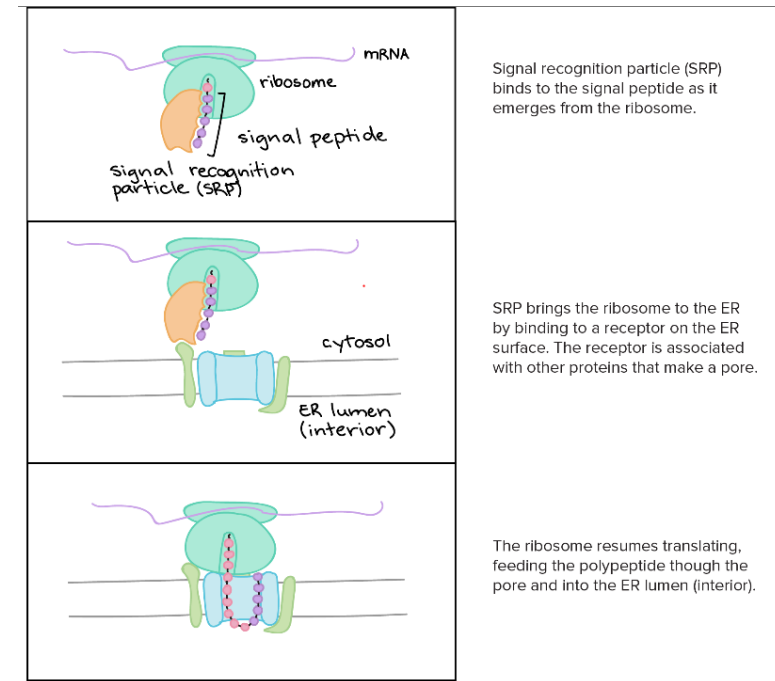
Try
CD / SMART/ InterPro
search

Find domains in your sequence

A solid orange horizontal bar at the bottom of the slide.

ER signal peptide prediction

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



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

Signal peptides

SignalP

DTU Health Tech

[Research](#) [Publications](#) [Education](#) [Collaboration](#) [Services and Products](#) [News](#) [About](#)

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](#).

Submission	Instructions	Data	Article abstract	FAQ	Version history	Portable	Downloads	
------------	--------------	------	------------------	-----	-----------------	----------	-----------	--

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.

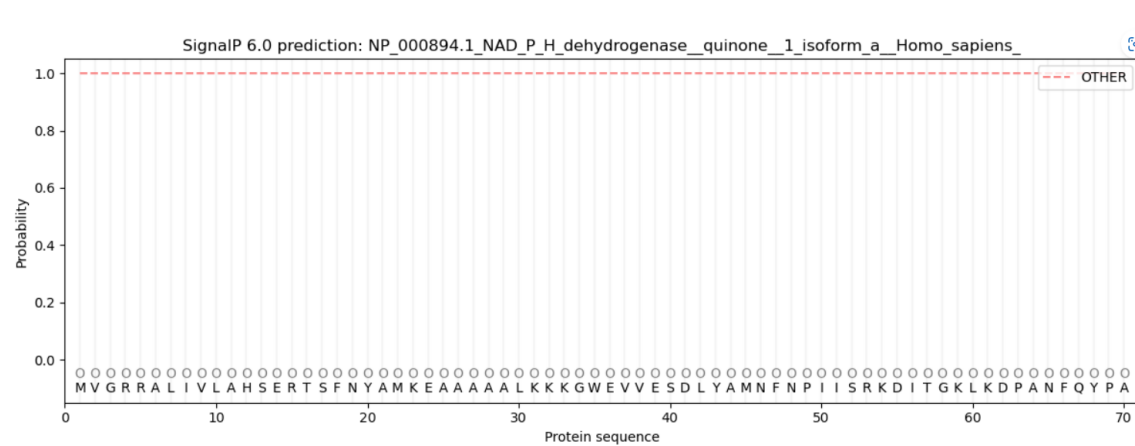
[Mirror](#) Use SignalP 6.0 on BioLib if this server is heavily loaded.

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

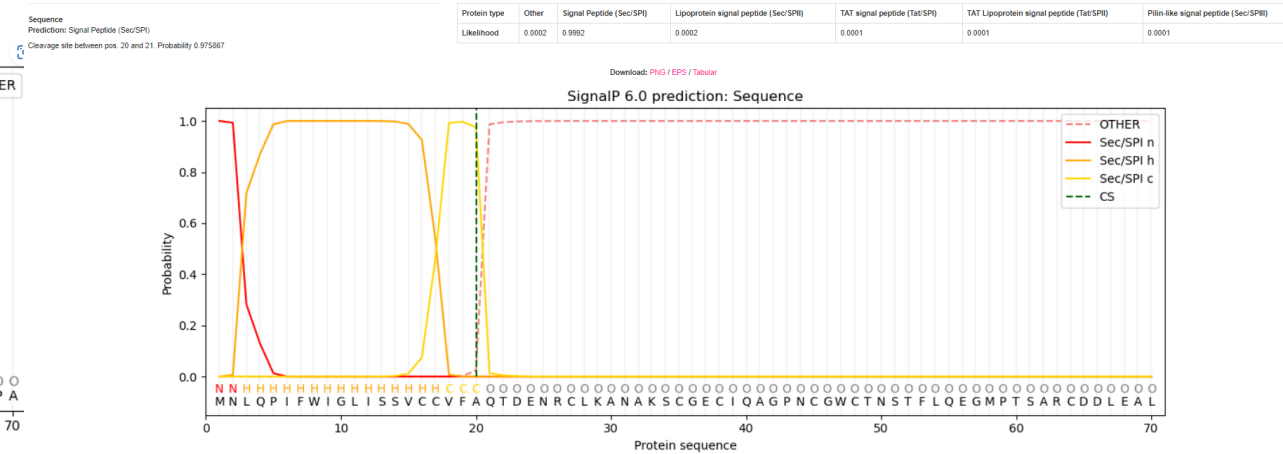


Signal peptides

SignalP



protein does not have signal peptide



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

Signal peptides



SMART MODE:
NORMAL
GENOMIC

Simple
Modular
Architecture
Research
Tool

Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* 95, 5857-5864
Letunic et al. (2014) *Nucleic Acids Res* doi: 10.1093/nar/gku949
[HOME](#) [SETUP](#) [FAQ](#) [ABOUT](#) [GLOSSARY](#) [WHAT'S NEW](#) [FEEDBACK](#)

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

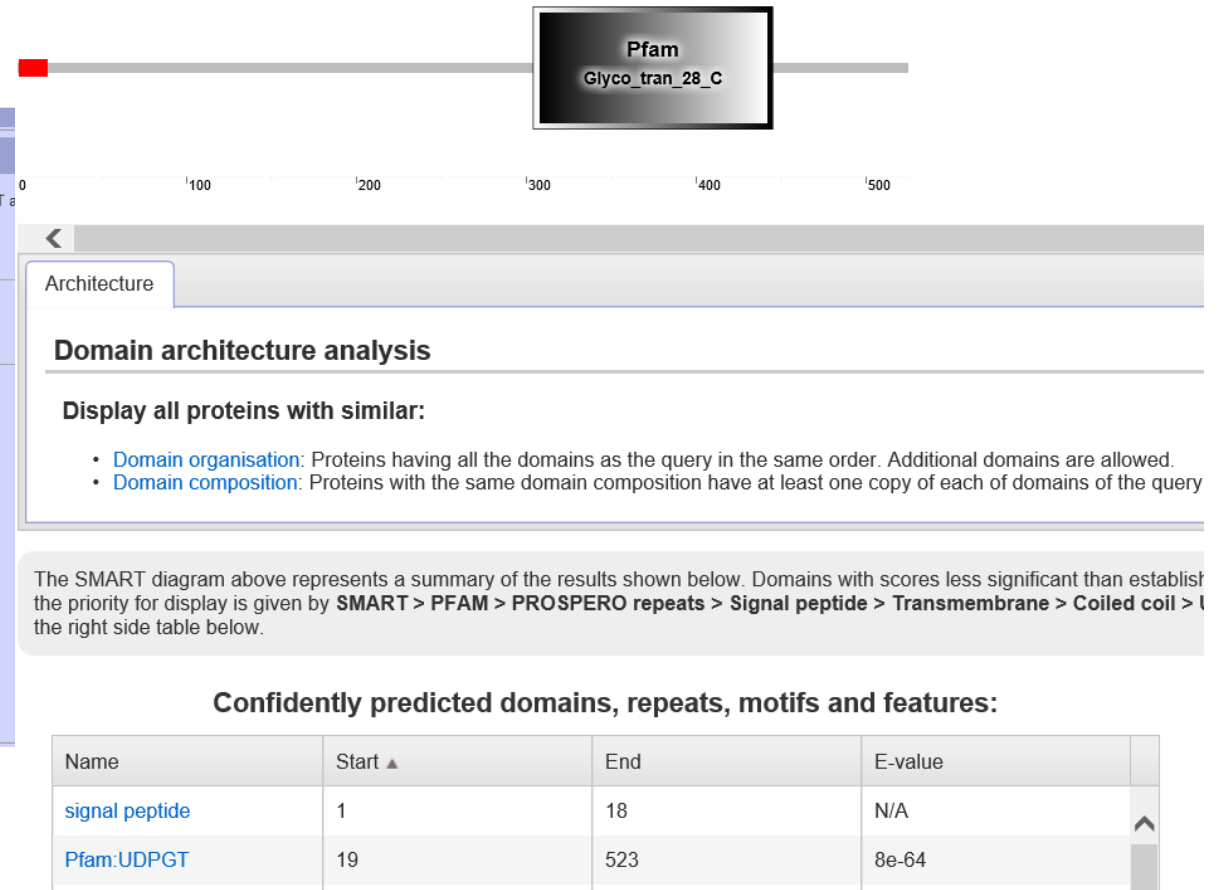
Protein sequence

MVAAATVAAAMLLLWAAACAQQEQDFYDFKAVNIRGKLVSLKYRGSVSLVVNVA
SECGFT
DQHYRALQQLQRLDGPHHFNVLAFFPCNQFGQQEPDSNKEIESFARRTYSVSFFPM
FSKIAV
TGTGAHPAFKYLQTSKGKEPTWNEWKYLVAPDGKVVGAWDFTVSVEEVRPQITA
LVRKLI
LLKREDL

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

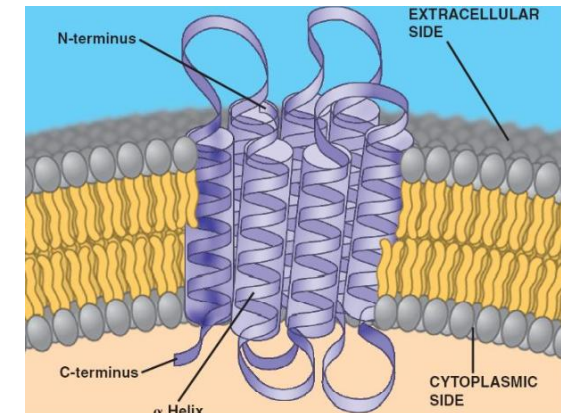
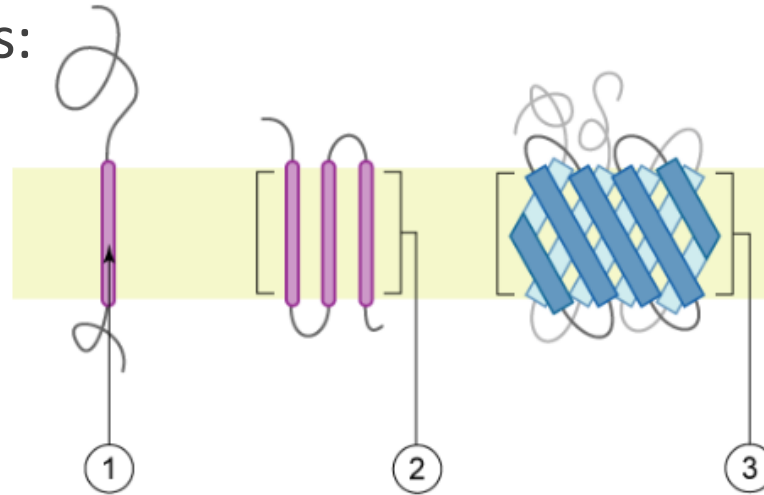


Practical part

search for signal peptide in
your sequence

Prediction of transmembrane helices

Transmembrane proteins:



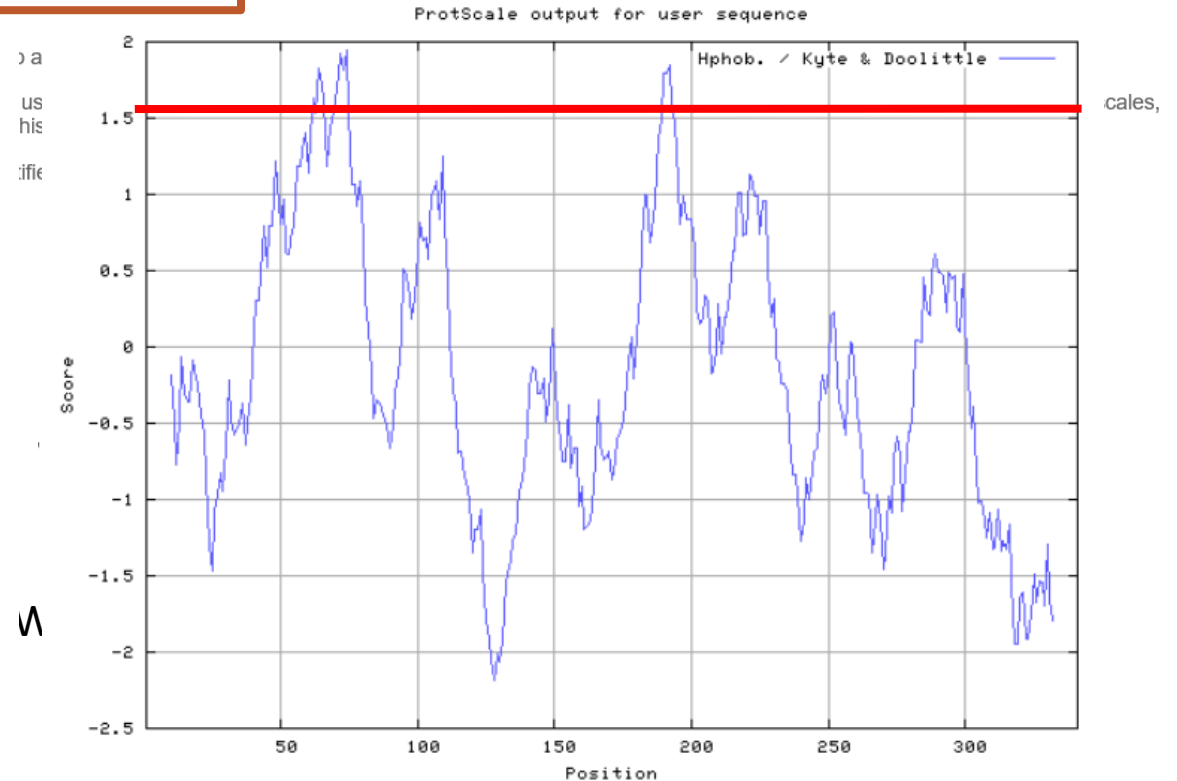
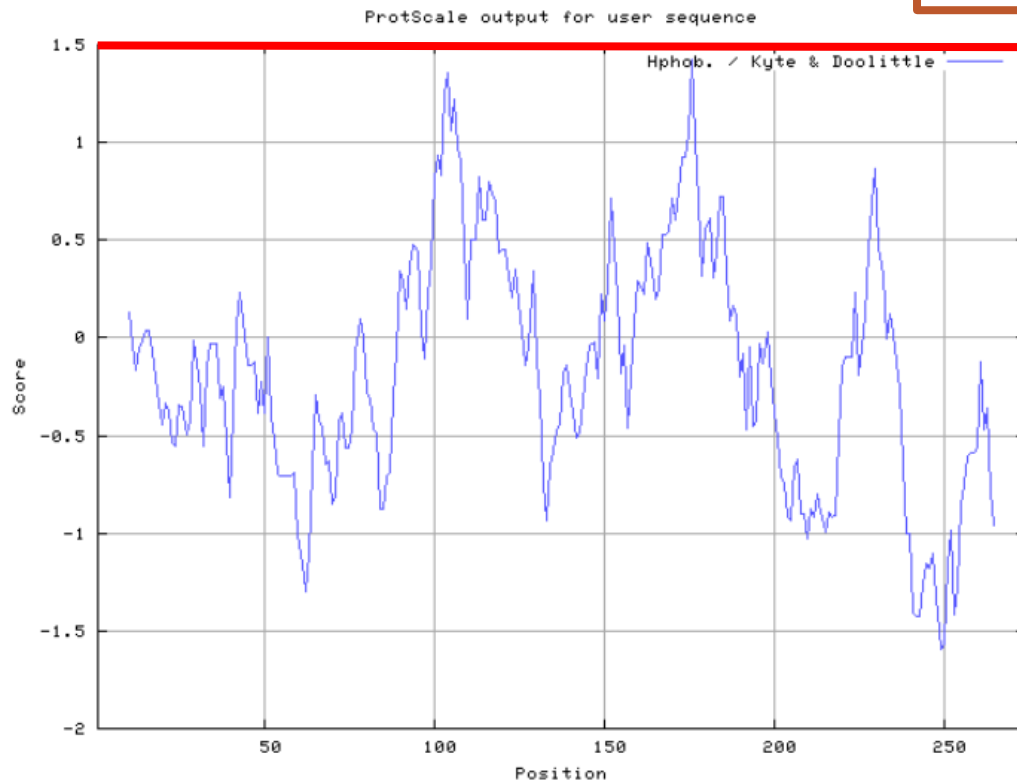
Amino acid Hydrophobicity

- various programs – different algorithms – 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

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS CBS

EVENTS

STAFF

NEWS

CONTACT

RESEARCH GROUPS

ABOUT CBS

CBS PREDICTION SERVERS

INTERNAL

CBS DATA SETS

CBS BIOINFORMATICS TOOLS

PUBLICATIONS

CBS COURSES

EDUCATION

OTHER BIOINFORMATICS LINKS

CBS >> CBS Prediction Servers >> TMHMM

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

Submission of a local file in FASTA format

OR by pasting sequence(s) in FASTA format

Output format:

Other options:

Restrictions:

Confidentiality:

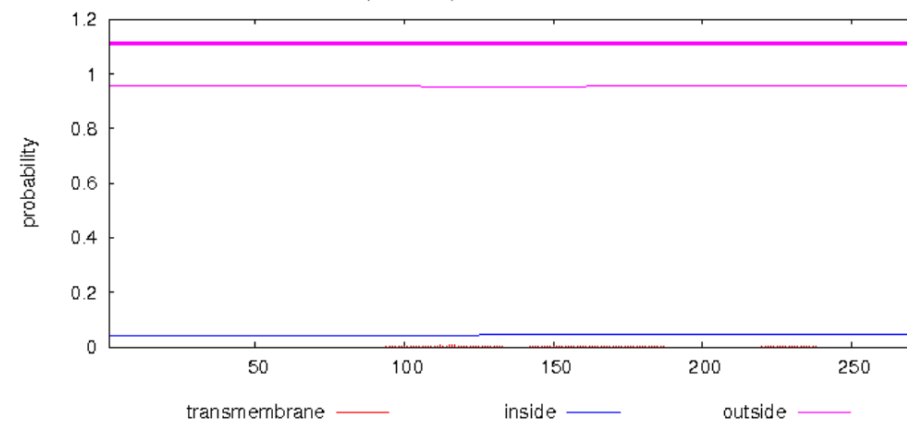
TMHMM result

[HELP](#) with output formats

no TM helix

```
# WEBSEQUENCE Length: 274
# WEBSEQUENCE Number of predicted TMHs: 0
# WEBSEQUENCE Exp number of AAs in TMHs: 0.20324
# WEBSEQUENCE Exp number, first 60 AAs: 0
# WEBSEQUENCE Total prob of N-in: 0.04315
WEBSEQUENCE TMHMM2.0 outside 1 274
```

TMHMM posterior probabilities for WEBSEQUENCE



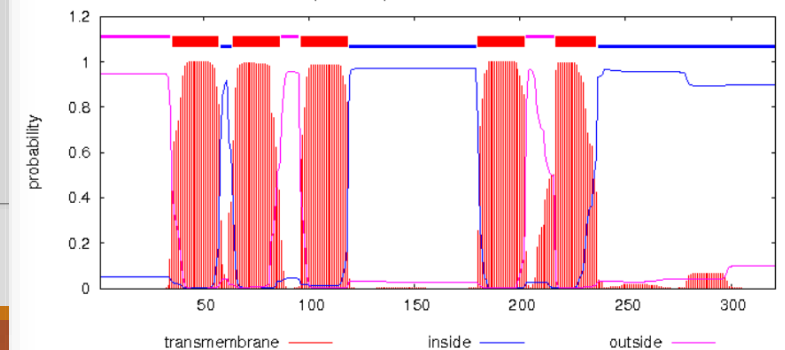
TMHMM result

[HELP](#) with output formats

five predicted TM helices

```
# WEBSEQUENCE Length: 321
# WEBSEQUENCE Number of predicted TMHs: 5
# WEBSEQUENCE Exp number of AAs in TMHs: 108.47546
# WEBSEQUENCE Exp number, first 60 AAs: 21.62676
# WEBSEQUENCE Total prob of N-in: 0.05151
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE TMHMM2.0 outside 1 34
WEBSEQUENCE TMHMM2.0 TMhelix 35 57
WEBSEQUENCE TMHMM2.0 inside 58 63
WEBSEQUENCE TMHMM2.0 TMhelix 64 86
WEBSEQUENCE TMHMM2.0 outside 87 95
WEBSEQUENCE TMHMM2.0 TMhelix 96 118
WEBSEQUENCE TMHMM2.0 inside 119 179
WEBSEQUENCE TMHMM2.0 TMhelix 180 202
WEBSEQUENCE TMHMM2.0 outside 203 216
WEBSEQUENCE TMHMM2.0 TMhelix 217 236
WEBSEQUENCE TMHMM2.0 inside 237 321
```

TMHMM posterior probabilities for WEBSEQUENCE

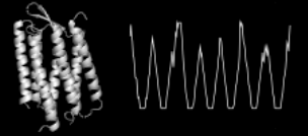


TOPCONS

TOPCONS

TOPCONS

TOPCONS



New query

Batch WSDL API

Download

References

News

Server status

Example results

Old TOPCONS

Help

Consensus prediction of membrane peptides

Please paste your amino acid sequences in [FASTA](#) format. Allowed characters: "ABCDEFGHIJKLMNOPQRSTUVWXYZ*", < > (Sequences should be no shorter than 10 amino acids)

Alternatively, upload a text file in FASTA format up to 100 MB

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

[Procházet...](#)

[Submit](#) [Clear](#) [Generate example input](#)

© Arne Elofsson

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: 2018-03-14
- Status: **Finished**
- Waiting time: 1 sec
- Running Time: 28 sec

Results of your prediction with jobid: **rst_BKOIKK**

Zipped folder of your result can be found in [rst_BKOIKK.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](#)

© Arne Elofsson

Predicted topologies and predicted ΔG values:



Results

- Submitted: 2018-03-05 15:59:14
- Status: **Finished**
- Waiting time: 0 sec
- Running Time: 0 sec

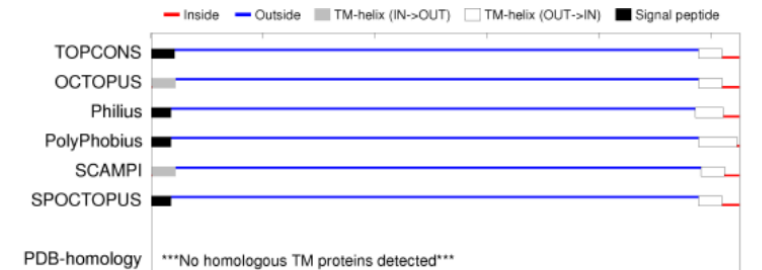
Results of your prediction with jobid: **rst_BKOIKK**

Zipped folder of your result can be found in [rst_BKOIKK.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:



PDB-homology ***No homologous TM proteins detected***

Paste your protein sequence here in Fasta format:

Copyright 2019 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. WCN 02-200-203

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

Select output format:

- ☐ Short
- ☐ Long without Graphics
- ☒ Long with Graphics

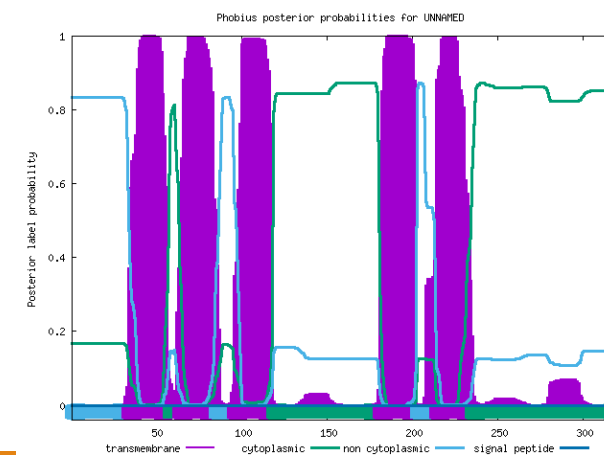
Odeslat Resetovat

A combined transmembrane topology and signal peptide predictor

Phobius prediction

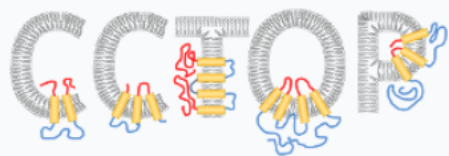
Prediction of UNNAMED

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



The probability data used in the plot is found [here](#), and the gnuplot script is [here](#).

CCTOP



 Submit

Manual ▾

[? About](#)

 Standalone

☰ MyJobs

Job ID



Results for job 92e87f663074d81d642456ea6e9d63a6

[Download results](#)

XML file

Proteins:

```
# sequences: 1/1
```

NP_005054.3

Control panel

NP_005054.3

Summary

1D

2D

```
>sequence
```

MPAHLQLQDDISSYTTTTTITAPPSRVLQNGGDKLETMPLYLEDDIRPDIKDDIYDPTYKDKEGPSKVEYVWR
LYGELIIP¹TC²K³F⁴Y⁵TWLVGVF⁶Y⁷F⁸V⁹S¹⁰ALG¹¹ITAGAHRLWSHRSYKARLPLRLFLIIANTMAFQNDVYEWARHRAH
SRRGFFFSHVGWLLVVRKHPAVKEKGSTLDLSLEAEKLVMFQRRYYK²⁰PGLLMMCFILPTLVPM²¹YFWGETFQ²²N²³V²⁴
NA²⁵TWLVNSAAHLFGYRPPYDKNISPRENILVSLGAVGEGFHNYHHSFPYDYSASEYRWHINF³⁰TFFIDCMAALGL
LARIKRTGDGNGYKSG

```
>topology
```

NP_005054.3

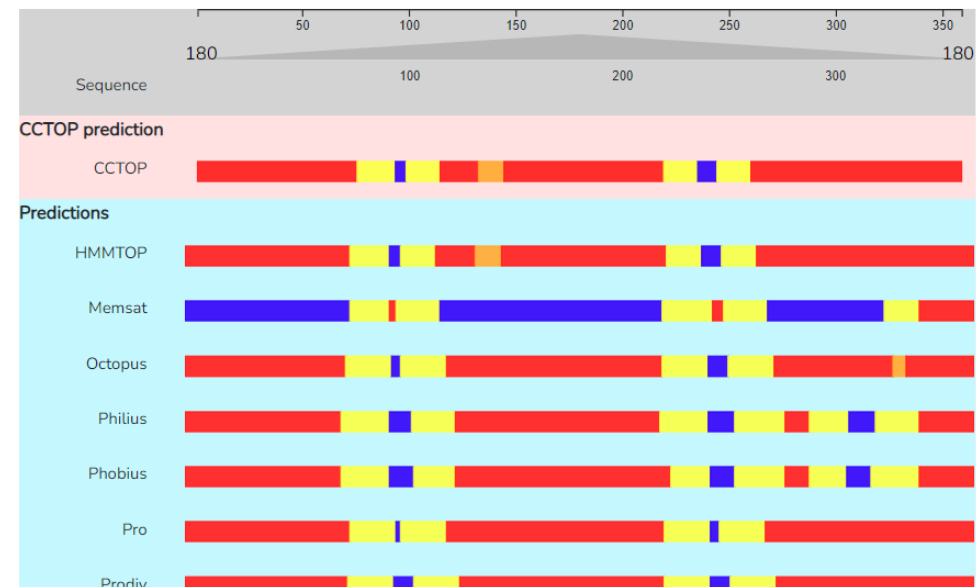
Summary

1D

2D

3D

Xml



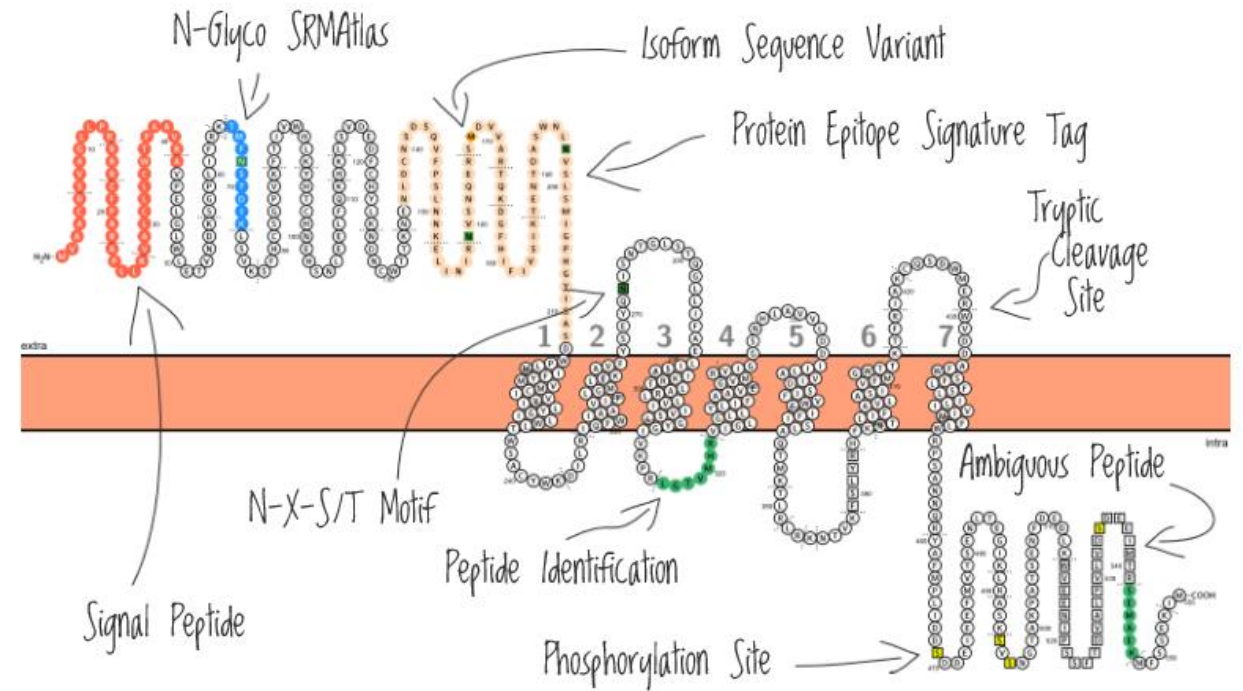
PROTTER-figure!

-creates figure from the UniProt data

PROTTER

version 1.0 | [help](#) | [manual](#) | [Wollscheid Lab](#)

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!



start **PROTTER**

PROTTER-figure!

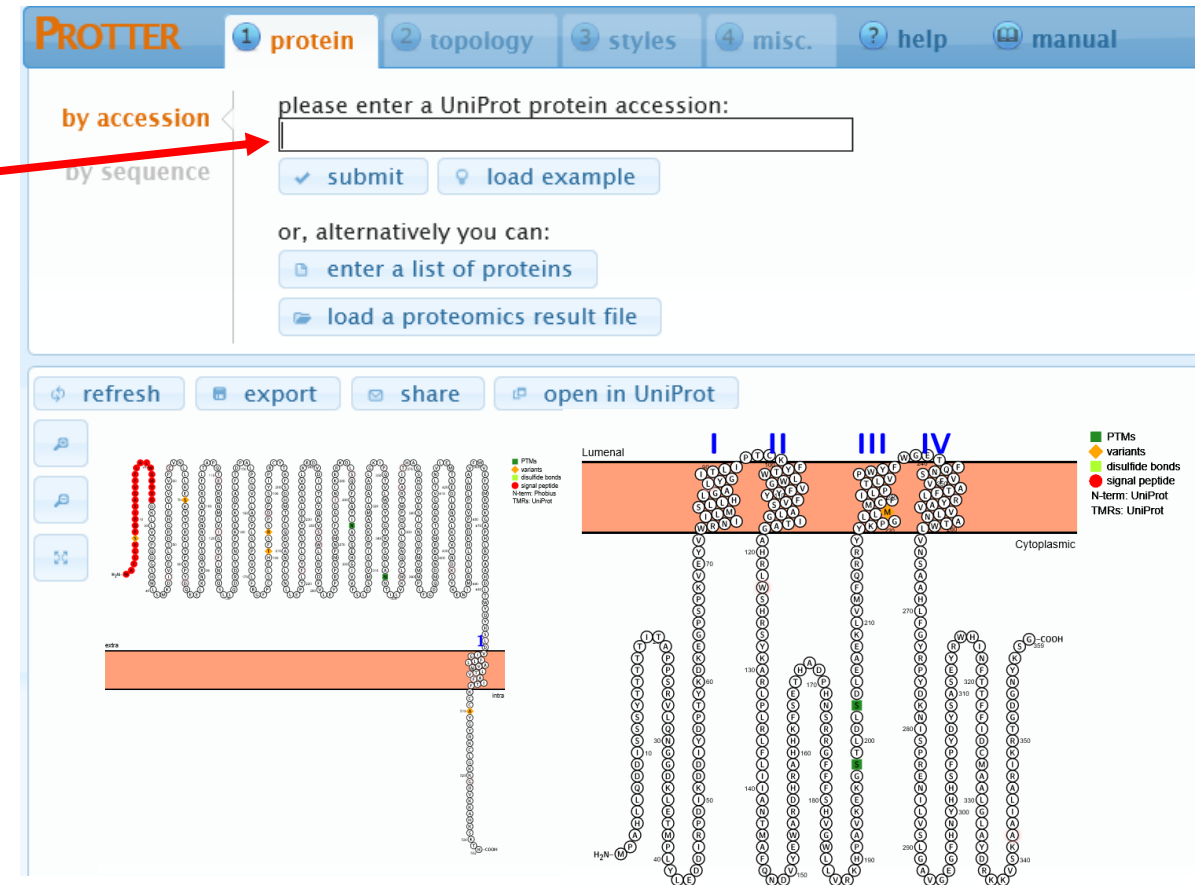
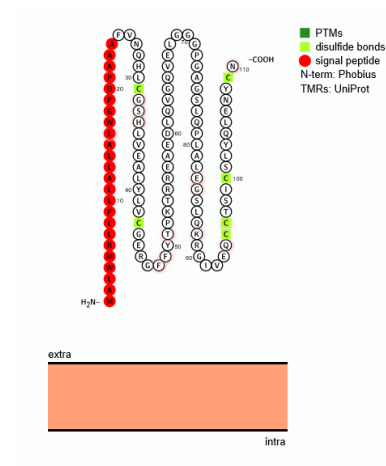
-creates figure from the UniProt data

-uses UniProt ID:

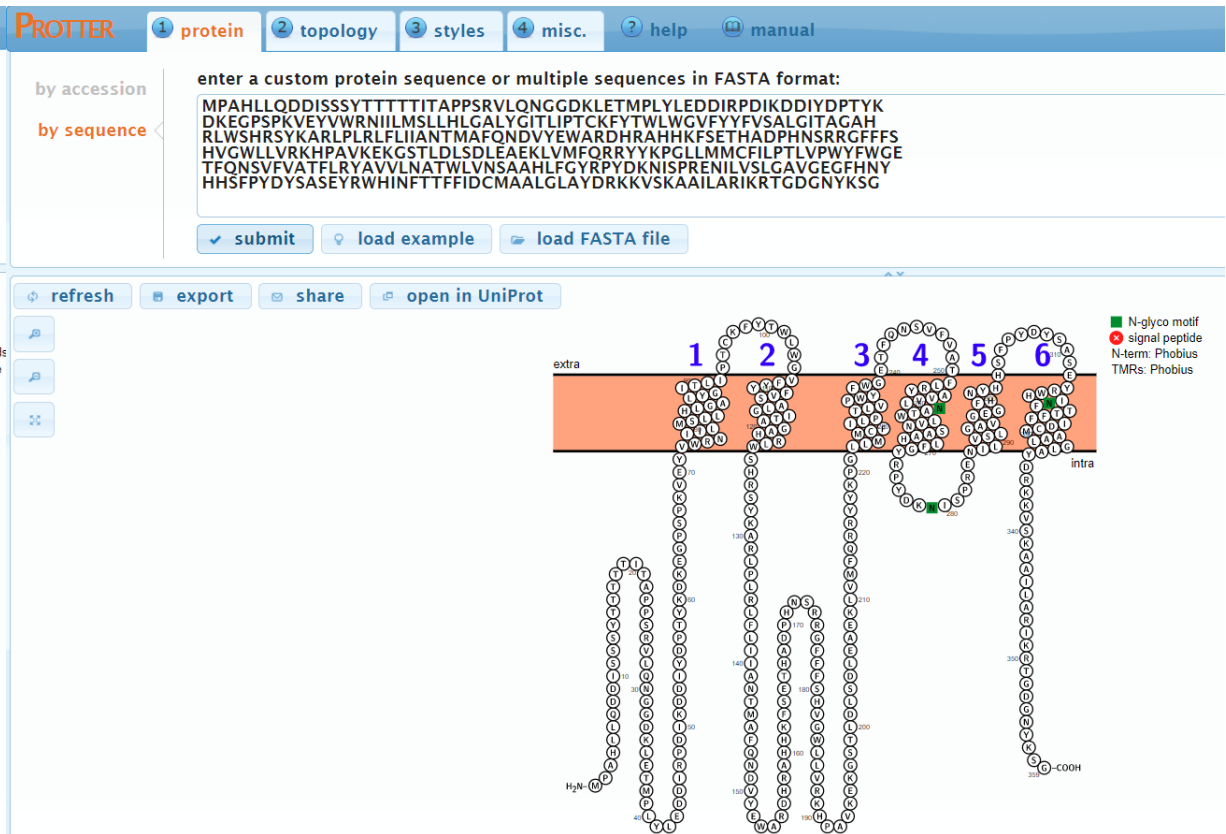
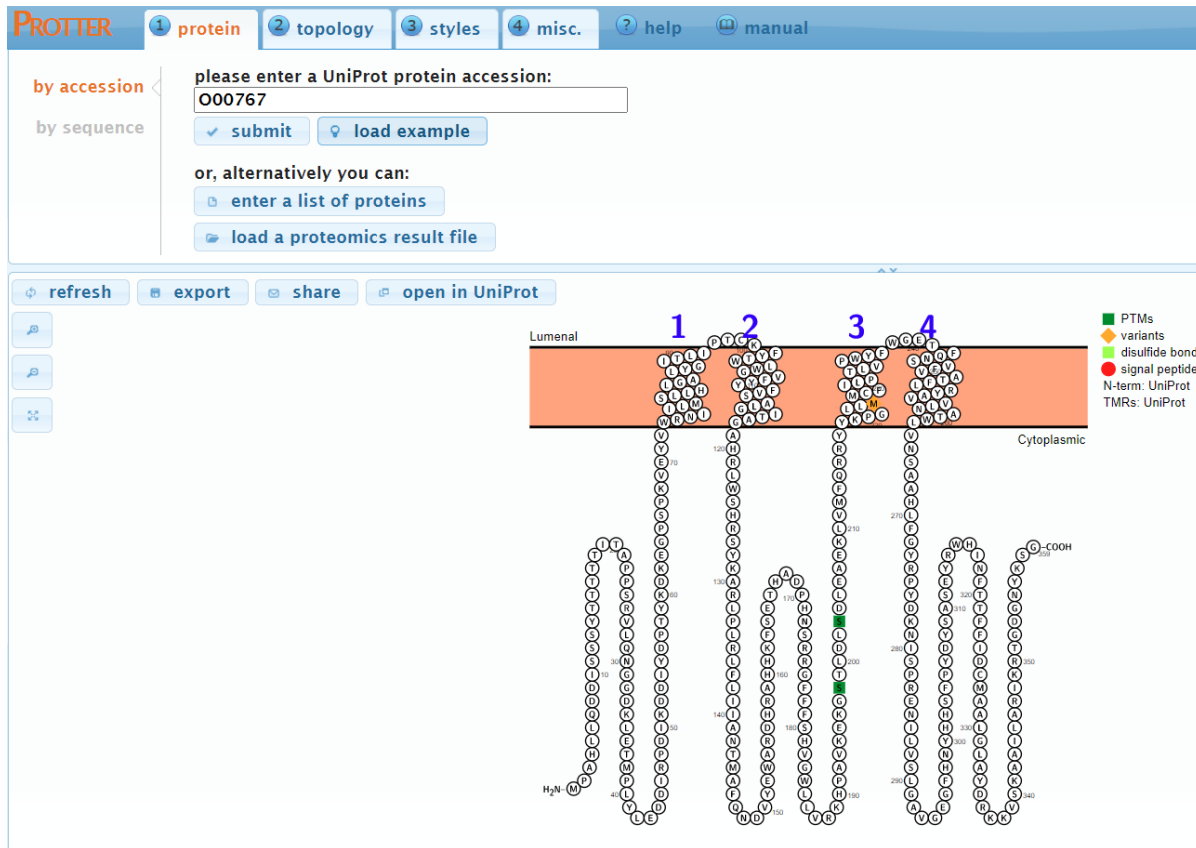
UGT1A6 (P19224)

Desaturase (O00767)

(prepro) insulin (P01308)

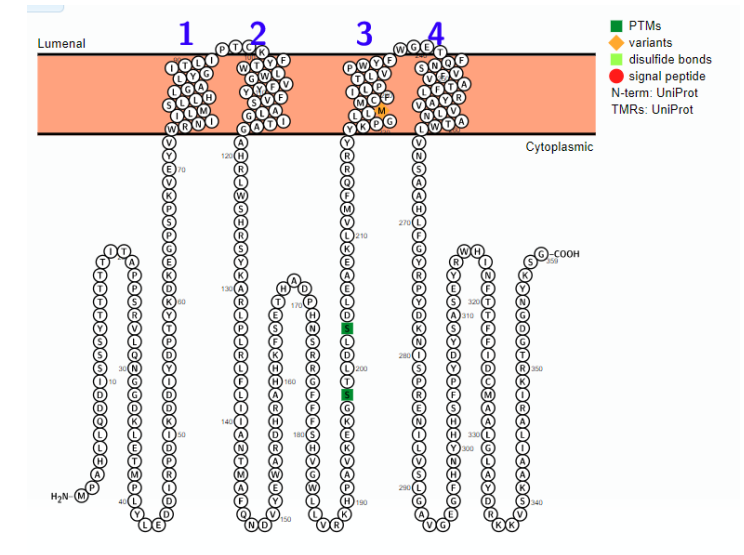
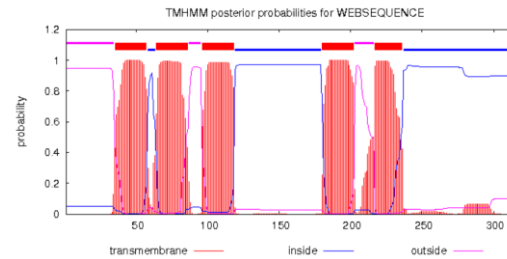
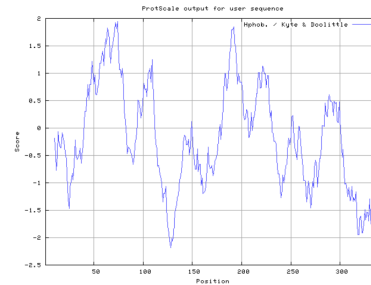


PROTTER-figure!



Prediction of transmembrane helices

ProtScale
TMHMM
TOPCONS



➤ Always try more programs!

Practical part

Try more programs.

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

„Protein bioinformatics I“ next lecture

Retrieving protein sequences from databases (Uniprot: FASTA format)

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

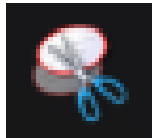
Computing the optimal alignment between two or more protein sequences

...

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

C

2)

1)



3) 2)

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

4)

