

Public Database의 이용(1) - SignalP (version 4.1)

2015. 8.

KIST 이철주

Secretion pathway prediction

ProteinCenter (Proxeon Bioinformatics, Odense, Denmark; <http://www.cbs.dtu.dk/services>)

SignalP (version 4.1)

(<http://www.cbs.dtu.dk/services/SignalP>)

Classical secretion pathway 예상

hidden Markov model algorithms 이용ⁱ

SignalP uses amino acid sequences to predict the existence and location of signal peptide cleavage sites. The hidden Markov model algorithms calculate the probability of whether the submitted sequence contains a signal peptide or not. A protein is considered classically secreted if it receives a default cutoff values Y (0.5 > D-cutoff for TM networks, 0.45 > D-cutoff for no-TM model networks).ⁱⁱ

SignalP 4.1 Server

SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide-signal peptide prediction based on a combination of several artificial neural networks.

View the [version history](#) of this server. All the previous versions are available on line, for comparison and reference.

New: SignalP has been updated to version 4.1 with two new features:

- an option to choose a D-score cutoff that reproduces the sensitivity of SignalP 3.0 (this will make the false positive rate slightly higher, but still better than that of SignalP 3.0)
- a customizable minimum length of the predicted signal peptide (default 10)

Additionally, the documentation has been rewritten. The [Instructions](#) page is expanded, the [Output format](#) page has been clarified, and there are new [Performance](#) and [FAQ](#) pages.

[FAQ](#) | [Article abstracts](#) | [Instructions](#) | [Output format](#) | [Performance](#) | [Data](#)

SUBMISSION

Paste a single amino acid sequence or several sequences in **FASTA** format into the field below:

Submit a file in **FASTA** format directly from your local disk:

파일 선택 | 번역된 파일 업로드

Organism group [\(explain\)](#)

- Eukaryotes
- Gram-negative bacteria
- Gram-positive bacteria

D-cutoff values [\(explain\)](#)

- Default (optimized for correlation)
- Sensitive (reproduces SignalP 3.0's sensitivity)
- Use optimal
- 0.45 D-cutoff for SignalPro-TM networks
- 0.50 D-cutoff for SignalP-TM networks

Method [\(explain\)](#)

- Input sequences may include TM regions
- Input sequences do not include TM regions

Graphics output [\(explain\)](#)

- No graphics
- PNG (final)
- PNG (final) and GP (as links)

Positional limits [\(explain\)](#)

- Minimal predicted signal peptide length. Default: 10
- N-terminal truncation of input sequence (0 means no truncation). Default: Truncate sequence to a length of 70 aa

Submit | **Clear fields**

Restrictions:
At most 2,000 sequences and 200,000 amino acids per submission; each sequence not more than 6,000 amino acids.

Confidentiality:
The sequences are kept confidential and will be deleted after processing.

Fig 1. SignalP homepage

■ Parameter

1. Fasta 형식 또는 파일 준비
 - 한 번의 submission에 2,000 entries, 총 200,000 amino acids가능
 - 하나의 entry는 6,000 amino acids를 넘지 않아야 함 (넘을 경우 N-terminal에서부터 6000번째까지만 사용)
2. Organism group 선택
3. Output format: data 정리 및 분석 작업을 위해서는 "short" 선택

# name	Dmax	pos	Ymax	pos	Smax	pos	Smean	D	?	Dmaxcut	Networks-used
sp_P24821_TENA_HUMAN	0.557	20	0.721	20	0.974	5	0.936	0.837	Y	0.450	SignalP-noTM
sp_095967_FBLN4_HUMAN	0.760	28	0.864	28	0.988	18	0.955	0.913	Y	0.450	SignalP-noTM
sp_Q81VNB8_SBSPO_HUMAN	0.805	21	0.863	21	0.964	3	0.925	0.897	Y	0.450	SignalP-noTM
sp_P08582_TFFM_HUMAN	0.749	20	0.840	20	0.967	13	0.947	0.898	Y	0.450	SignalP-noTM

Fig 2. Output format "short"

```

# SignalP-4.1 euk predictions
>sp_P24821_TENA_HUMAN Tenascin OS: Homo sapiens GLT1NC PE.1 SV.3
# Measure Position Value Cutoff signal peptide?
max. C 20 0.557
max. Y 20 0.721
max. S 5 0.974
mean S 1-19 0.936
D 1-19 0.837 0.450 YES
Name=sp_P24821_TENA_HUMAN SP= YES Cleavage site between pos. 19 and 20: ALA-TE D=0.837 D-cutoff=0.450 Networks=SignalP-noTM
>sp_095967_FBLN4_HUMAN FGF-containing fibulin-like extracellular matrix protein 2 OS: Homo sapiens GLEFEMF2 PE.1 SV.3
# Measure Position Value Cutoff signal peptide?
max. C 28 0.760
max. Y 28 0.864
max. S 18 0.988
mean S 1-27 0.955
D 1-27 0.913 0.450 YES
Name=sp_095967_FBLN4_HUMAN SP= YES Cleavage site between pos. 27 and 28: ASP-DD D=0.913 D-cutoff=0.450 Networks=SignalP-noTM
>sp_Q81VNB8_SBSPO_HUMAN Smatomedin_B and thrombospondin type-1 domain-containing protein OS: Homo sapiens GLSBSPO PE.1 SV.2
# Measure Position Value Cutoff signal peptide?
max. C 21 0.805
max. Y 21 0.863
max. S 3 0.964
mean S 1-20 0.925
D 1-20 0.897 0.450 YES
Name=sp_Q81VNB8_SBSPO_HUMAN SP= YES Cleavage site between pos. 20 and 21: AQA-DC D=0.897 D-cutoff=0.450 Networks=SignalP-noTM
>sp_P08582_TFFM_HUMAN Melanotransferrin OS: Homo sapiens GULF12 PE.1 SV.2
# Measure Position Value Cutoff signal peptide?
max. C 20 0.749
max. Y 20 0.840
max. S 13 0.967
mean S 1-19 0.947
D 1-19 0.898 0.450 YES
Name=sp_P08582_TFFM_HUMAN SP= YES Cleavage site between pos. 19 and 20: VLG-GL D=0.898 D-cutoff=0.450 Networks=SignalP-noTM
  
```

Fig3. Output format "standard"

4. D-cutoff value: default 권장 (0.45이상 SignalP-noTM networks, 0.5이상 SignalP-TM networks)
 - *TM: transmembrane state (sequence의 hydrophobicity에 따라 integral transmembrane sequence 예상, 4개 이상 존재하면 TM으로 인지)
5. Method: default 권장 (input sequences may include TM regions)
6. Graphics output: 다수의 proteome 분석 시 default 권장 (no graphic)

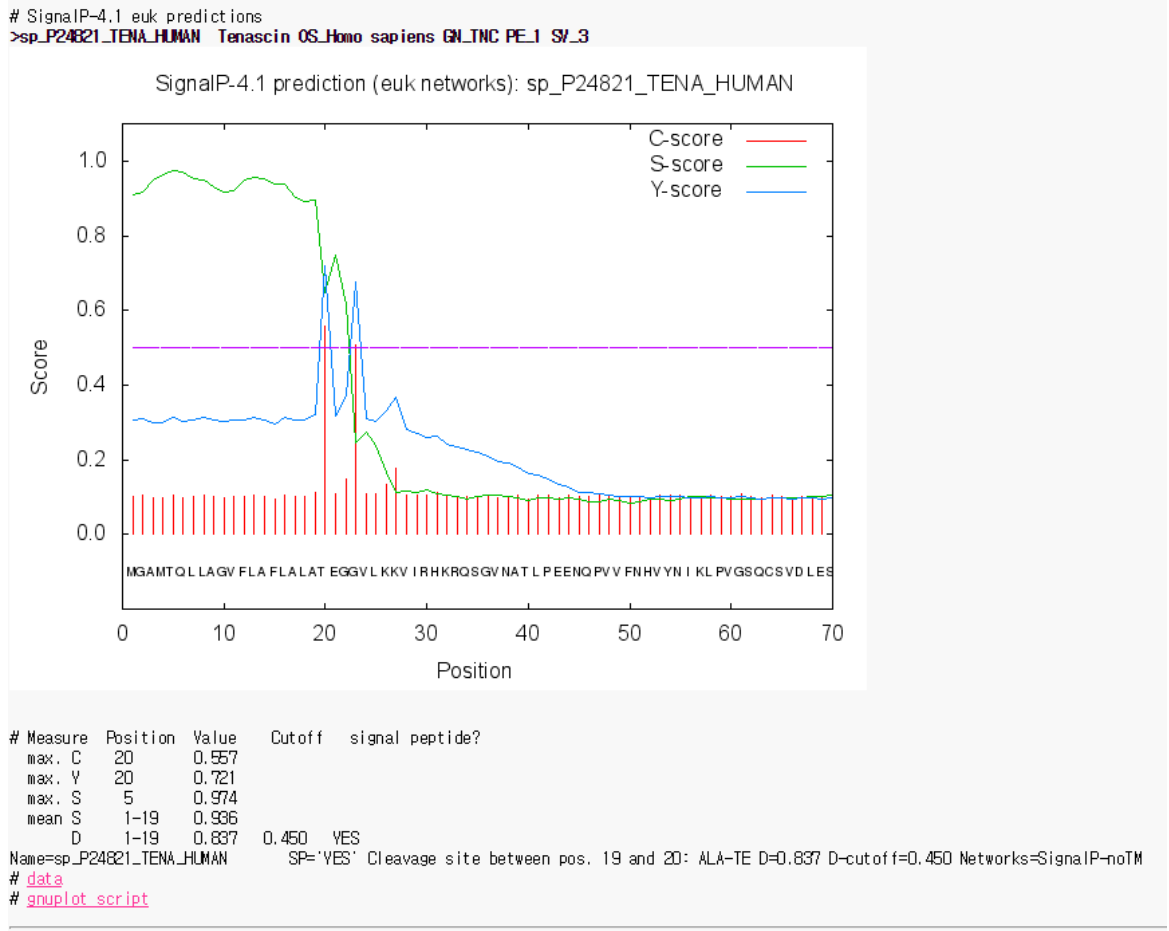


Fig 4. Graphics output "PNG (inline)" 선택 시, output

■ 결과 해석

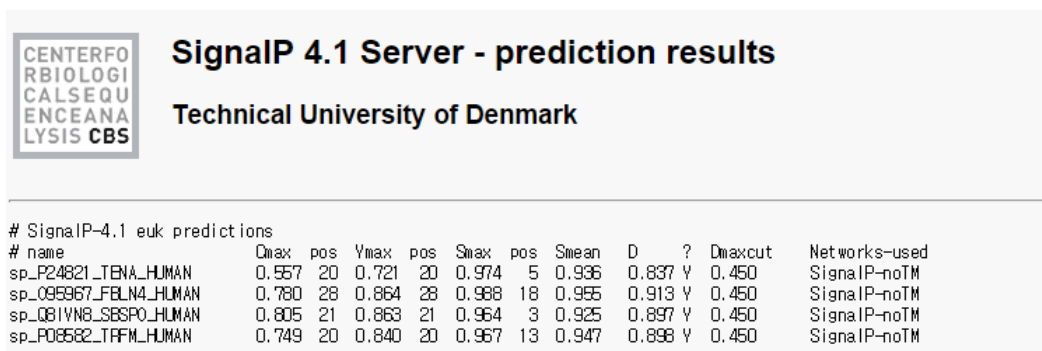


Fig 5. SignalP의 prediction results

D (D-cutoff)

? (Y: D-cutoff $0.45 \geq$ SignalP-noTM, $0.5 \geq$ SignalP-TM, N: $0.45 <$ SignalP-noTM, $0.5 <$ SignalP-TM)

Public Database의 이용(2) - SecretomeP (version 2.0)

2015. 8.

KIST 이철주

SecretomeP (version 2.0)

- (<http://www.cbs.dtu.dk/services/SecretomeP>)
- nonclassical secretion pathway 예상
- neural network 이용

SecretomeP uses a neural network that combines six protein characteristics to determine if a protein is nonclassically secreted. These characteristics include the number of atoms, number of positively charged residues, presence of transmembrane helices, presence of low-complexity regions, presence of pro-peptides and subcellular localization. A protein is considered nonclassically secreted if it receives an NN-score ≥ 0.5 .ⁱⁱⁱ

CENTER FOR BIOLOGICAL ENGINEERING ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

CBS >> CBS Prediction Servers >> SecretomeP

SecretomeP 2.0 Server

Prediction of non-classical protein secretion

The SecretomeP 2.0 server produces *ab initio* predictions of non-classical i.e. not signal peptide triggered protein secretion. The method queries a large number of other feature prediction servers to obtain information on various post-translational and localizational aspects of the protein, which are integrated into the final secretion prediction.

View the [version history](#) of this server.

[Instructions](#) | [Output format](#) | [Article abstracts](#) | [Data sets](#) | [Supplementary material](#)

SUBMISSION

Paste a single sequence or several sequences in **FASTA** format into the field below:

Submit a file in **FASTA** format directly from your local disk:
파일 선택 | 선택된 파일 없음

Gram-negative bacteria
 Gram-positive bacteria
 Mammalian

Restrictions:
At most 100 sequences and 200,000 amino acids per submission; each sequence not less than 40 and not more than 4,000 amino acids.

Confidentiality:
The sequences are kept confidential and will be deleted after processing.

Fig 6. SecretomeP homepage

■ **Parameter**

1. Fasta 형식 또는 파일 준비
 - 한 번의 submission에 100 entries, 총 200,000 amino acids가능
 - 하나의 entry는 40~4,000 amino acids여야 함. (넘을 경우 N-terminal에서부터 4,000번째까지만 사용)
2. Organism group 선택

■ **결과 해석**

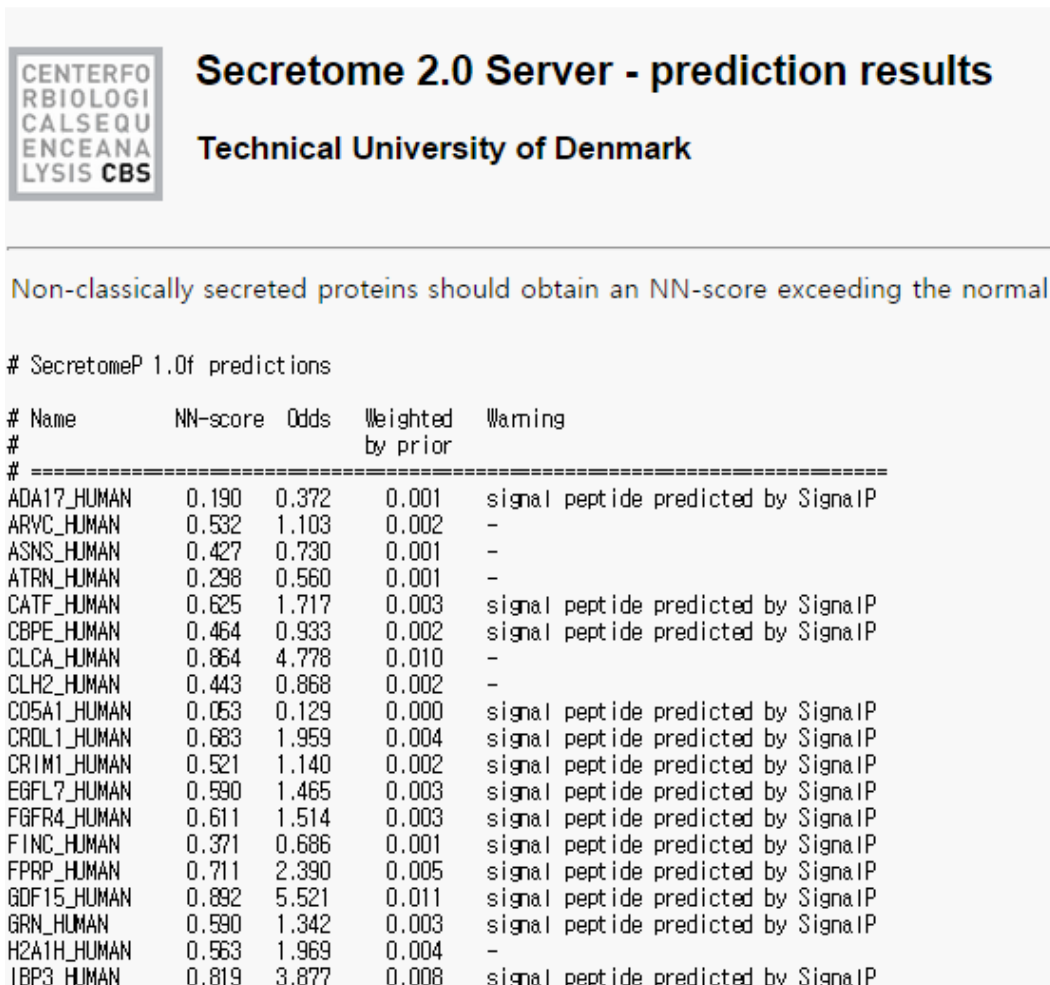


Fig 7. SecretomeP의 prediction results

SignalP 에 의해 signal peptide가 예상된 된 단백질을 제외하고 NN-score가 0.5이상인 경우, nonclassical secretion pathway를 통해 secretion 되었다고 예상

Public Database의 이용(3) - TMHMM (version 2.0)

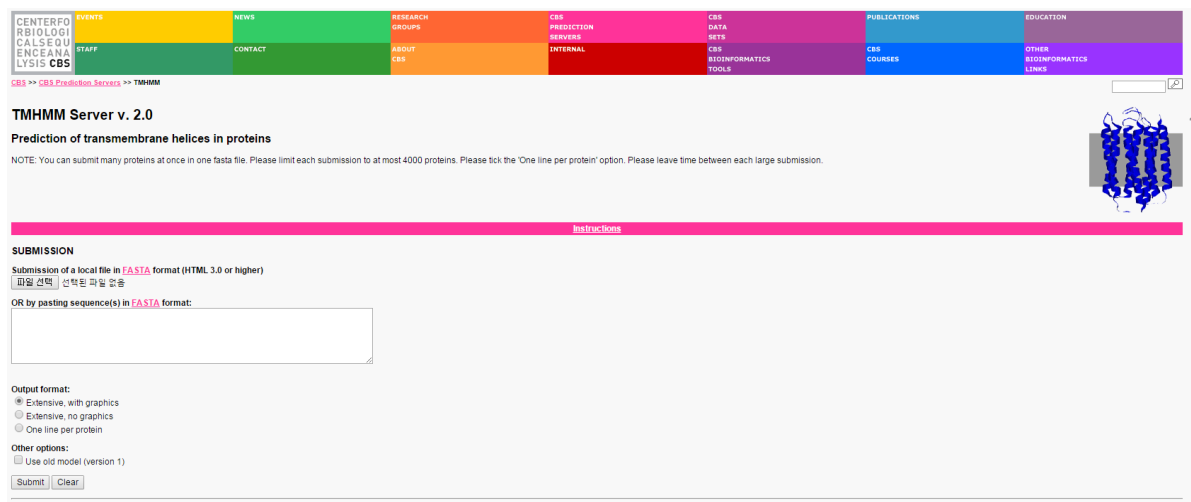
2015. 8.

KIST 이철주

TMHMM (version 2.0)

- (<http://www.cbs.dtu.dk/services/TMHMM>)
- transmembrane helices in integral membrane proteins 예상
- hidden Markov model 이용

The TMHMM program is based on a hidden Markov model approach. The hidden Markov model can incorporate hydrophobicity, charge bias, helix lengths and grammatical constraints into one model for which algorithms for parameter estimate and prediction already exist.^{iv}



CENTER FOR BIOLOGICAL SEQUENCING AND ANALYSIS CBS

EVENTS NEWS RESEARCH GROUPS CBS PREDICTION SERVICES CBS DATA PUBLICATIONS EDUCATION

STAFF CONTACT ABOUT CBS INTERNAL CBS DATA SETS CBS BIOPHARMACEUTICALS TOOLS CBS COURSES OTHER BIOPHARMACEUTICALS LINKS

CBS >> CBS Prediction Services >> TMHMM

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

NOTE: You can submit many proteins at once in one fasta file. Please limit each submission to at most 4000 proteins. Please tick the 'One line per protein' option. Please leave time between each large submission.

[Instructions](#)

SUBMISSION

Submission of a local file in **FASTA** format (HTML 3.0 or higher)
파일 선택 | 선택된 파일 없음

OR by pasting sequence(s) in **FASTA** format:

Output format:

Extensive, with graphics
 Extensive, no graphics
 One line per protein

Other options:

Use old model (version 1)

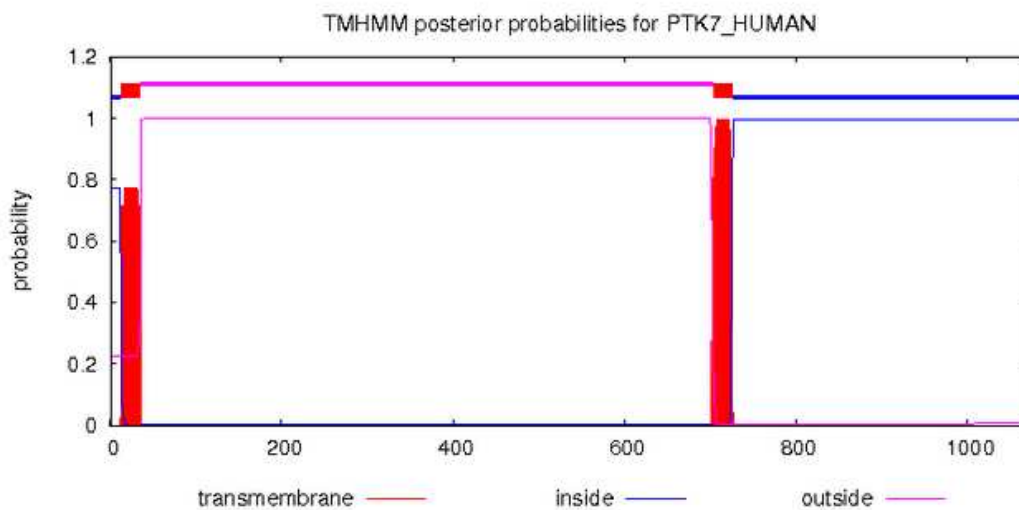
Fig 8. TMHMM homepage

■ Parameter

1. Fasta 형식 또는 파일 준비
 - 한 번의 submission에 4,000 entries 가능
2. Output format 선택 : one line per protein 추천

■ 결과해석

```
# PTK7_HUMAN Length: 1070
# PTK7_HUMAN Number of predicted TMHs: 2
# PTK7_HUMAN Exp number of AAs in TMHs: 39.834649999999999
# PTK7_HUMAN Exp number, first 60 AAs: 17.43834
# PTK7_HUMAN Total prob of N-in: 0.77345
# PTK7_HUMAN POSSIBLE N-term signal sequence
PTK7_HUMAN TMHMM2.0 inside 1 12
PTK7_HUMAN TMHMM2.0 TMhelix 13 35
PTK7_HUMAN TMHMM2.0 outside 36 703
PTK7_HUMAN TMHMM2.0 TMhelix 704 726
PTK7_HUMAN TMHMM2.0 inside 727 1070
```



```
# plot in postscript, script for making the plot in gnuplot, data for plot
```

Fig 9. TMHMM의 prediction results (Extensive, with graphics)

TMHMM result

[HELP](#) with output formats

```

GDF15_HUMAN len=308 ExpAA=3.03 First60=3.03 PredHel=0 Topology=0
LOXL3_HUMAN len=753 ExpAA=0.48 First60=0.47 PredHel=0 Topology=0
GRDL1_HUMAN len=450 ExpAA=0.01 First60=0.01 PredHel=0 Topology=0
MMP10_HUMAN len=476 ExpAA=0.51 First60=0.46 PredHel=0 Topology=0
TGFBI_HUMAN len=390 ExpAA=7.17 First60=7.16 PredHel=0 Topology=0
LAMC2_HUMAN len=1193 ExpAA=0.26 First60=0.26 PredHel=0 Topology=0
QPCT_HUMAN len=361 ExpAA=4.36 First60=4.29 PredHel=0 Topology=0
LTBP1_HUMAN len=1721 ExpAA=2.46 First60=2.46 PredHel=0 Topology=0
len=593 ExpAA=0.03 First60=0.03 PredHel=0 Topology=0
COSA1_HUMAN len=1838 ExpAA=5.38 First60=5.37 PredHel=0 Topology=0
IBP3_HUMAN len=291 ExpAA=4.22 First60=4.22 PredHel=0 Topology=0
LAMB3_HUMAN len=1172 ExpAA=0.00 First60=0.00 PredHel=0 Topology=0
STC1_HUMAN len=247 ExpAA=0.01 First60=0.01 PredHel=0 Topology=0
EBF1L7_HUMAN len=273 ExpAA=5.36 First60=5.36 PredHel=0 Topology=0
RINT2_HUMAN len=256 ExpAA=8.19 First60=8.19 PredHel=0 Topology=0
INHBA_HUMAN len=426 ExpAA=0.01 First60=0.01 PredHel=0 Topology=0
MIS_HUMAN len=560 ExpAA=0.57 First60=0.56 PredHel=0 Topology=0
FINC_HUMAN len=2366 ExpAA=0.17 First60=0.17 PredHel=0 Topology=0
CBPE_HUMAN len=476 ExpAA=10.01 First60=10.00 PredHel=0 Topology=0
H2A1H_HUMAN len=128 ExpAA=1.63 First60=1.13 PredHel=0 Topology=0
MFGM_HUMAN len=387 ExpAA=2.94 First60=2.94 PredHel=0 Topology=0
RS12_HUMAN len=132 ExpAA=0.01 First60=0.01 PredHel=0 Topology=0
PTK7_HUMAN len=1070 ExpAA=39.83 First60=17.44 PredHel=2 Topology=i13-35o704-726i
ATRN_HUMAN len=1429 ExpAA=26.26 First60=0.05 PredHel=1 Topology=o1279-1301i
MET_HUMAN len=1390 ExpAA=25.70 First60=0.61 PredHel=1 Topology=o933-955i
RPESP_HUMAN len=264 ExpAA=0.00 First60=0.00 PredHel=0 Topology=0
MSLN_HUMAN len=630 ExpAA=19.57 First60=13.00 PredHel=1 Topology=i9-31o
FPRP_HUMAN len=879 ExpAA=22.24 First60=0.06 PredHel=1 Topology=o831-853i
ASNS_HUMAN len=561 ExpAA=0.02 First60=0.01 PredHel=0 Topology=0
NOTUN_HUMAN len=496 ExpAA=0.04 First60=0.02 PredHel=0 Topology=0
CR1M1_HUMAN len=1036 ExpAA=32.05 First60=10.15 PredHel=1 Topology=o939-961i
TRFM_HUMAN len=738 ExpAA=0.16 First60=0.06 PredHel=0 Topology=0
UBZD4_HUMAN len=147 ExpAA=0.01 First60=0.01 PredHel=0 Topology=0
ADA17_HUMAN len=624 ExpAA=21.54 First60=0.01 PredHel=1 Topology=o672-694i
SF3B4_HUMAN len=424 ExpAA=0.00 First60=0.00 PredHel=0 Topology=0
ARVC_HUMAN len=962 ExpAA=0.00 First60=0.00 PredHel=0 Topology=0
CATF_HUMAN len=484 ExpAA=0.45 First60=0.44 PredHel=0 Topology=0
LTBP4_HUMAN len=1624 ExpAA=7.83 First60=7.83 PredHel=0 Topology=0
FGFR4_HUMAN len=802 ExpAA=6.96 First60=0.09 PredHel=0 Topology=0
CLCA_HUMAN len=248 ExpAA=0.00 First60=0.00 PredHel=0 Topology=0
CLH2_HUMAN len=1640 ExpAA=1.26 First60=0.00 PredHel=0 Topology=0
NDRG3_HUMAN len=375 ExpAA=1.16 First60=0.00 PredHel=0 Topology=0

```

Fig 10. TMHMM prediction results (one line per protein)

- Length ("len="): the length of the protein sequence.
- Number of predicted TMHs: The number of predicted transmembrane helices.
- Exp number of AAs in TMHs ("ExpAA="): The expected number of amino acids intramembrane helices. If this number is larger than 18 it is very likely to be a transmembrane protein (OR have a signal peptide).
- Exp number, first 60 AAs ("First60="): The expected number of amino acids in transmembrane helices in the first 60 amino acids of the protein. If this number more than a few, you should be warned that a predicted transmembrane helix in the N-term could be a signal peptide.
- Total prob of N-in: The total probability that the N-term is on the cytoplasmic side of the membrane.
- POSSIBLE N-term signal sequence: a warning that is produced when "Exp number, first 60 AAs" is larger than 10.
- "PredHel=": The number of predicted transmembrane helices by N-best.
- "Topology=": The topology predicted by N-best. ('i' if the loop is on the inside or 'o' if it is on the outside)

ⁱ Nielsen H, Krogh A. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol.* 1998;6:122-30.

ⁱⁱ Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods.* 2011;8:785-6.

ⁱⁱⁱ Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel.* 2004;17:349-56.

^{iv} Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 2001;305:567-80.