TMHMM2.0 User's guide

This server is for prediction of transmembrane helices in proteins.
The method (version 1) is described in

Erik L.L. Sonnhammer, Gunnar von Heijne, and Anders Krogh:
A hidden Markov model for predicting transmembrane helices in protein sequences.
In Proc. of Sixth Int. Conf. on Intelligent Systems for Molecular Biology, p 175-182
Ed J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff, and C. Sensen
Menlo Park, CA: AAAI Press, 1998

Download compressed postscript file

Download pdf file

Please cite.

Press here to see other material (model, training data, etc). .

Version 2 is very similar to version one, but it builds on a new model, so predictions are not identical. The web server has been improved (hopefully) a little. A paper is submitted describing TMHMM in more detail (publication details not available yet).

Input

The program takes proteins in **FASTA format.** It recognizes the 20 amino acids and B, Z, and X, which are all treated equally as unknown. Any other character is changed to X, so please make sure the sequences are sensible proteins

This is an example (one protein):

```
>5H2A_CRIGR you can have comments after the ID
MEILCEDNTSLSSIPNSLMQVDGDSGLYRNDFNSRDANSSDASNWTIDGENRTNLSFEGYLPPTCLSILHL
QEKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMSLAIADMLLGFLVMPVSMLTILYGYRWP
LPSKLCAVWIYLDVLFSTASIMHLCAISLDRYVAIQNPIHHSRFNSRTKAFLKIIAVWTISVGVSMPIPVF
GLQDDSKVFKQGSCLLADDNFVLIGSFVAFFIPLTIMVITYFLTIKSLQKEATLCVSDLSTRAKLASFSFL
PQSSLSSEKLFQRSIHREPGSYTGRRTMQSISNEQKACKVLGIVFFLFVVMWCPFFITNIMAVICKESCNE
HVIGALLNVFVWIGYLSSAVNPLVYTLFNKTYRSAFSRYIQCQYKENRKPLQLILVNTIPALAYKSSQLQA
GQNKDSKEDAEPTDNDCSMVTLGKQQSEETCTDNINTVNEKVSCV
```

How to run it

Either give the name of the local file in which you have the proteins in the top half of the window, or paste the sequence(s) into the lower part of the window. Then press `Submit'. (It should be possible to

both give it a local file and paste sequences if you really want.)

Output

There are two output formats: Long and short.

Long output format

For the long format (default), the server gives some statistics and a list of the location of the predicted transmembrane helices and the predicted location of the intervening loop regions.

Here is an example:

```
# COX2_BACSU Length: 278
# COX2_BACSU Number of predicted TMHs:  3
# COX2_BACSU Exp number of AAs in TMHs: 68.6888999999999
# COX2_BACSU Exp number, first 60 AAs:  39.8875
# COX2_BACSU Total prob of N-in:        0.99950
# COX2_BACSU POSSIBLE N-term signal sequence
COX2_BACSU        TMHMM2.0        inside       1     6
COX2_BACSU        TMHMM2.0        TMhelix      7    29
COX2_BACSU        TMHMM2.0        outside     30    43
COX2_BACSU        TMHMM2.0        TMhelix     44    66
COX2_BACSU        TMHMM2.0        inside      67    86
COX2_BACSU        TMHMM2.0        TMhelix     87   109
COX2_BACSU        TMHMM2.0        outside    110   278
```

If the whole sequence is labeled as inside or outside, the prediction is that it contains no membrane helices. *It is probably not wise to interpret it as a prediction of location*. The prediction gives the most probable location and orientation of transmembrane helices in the sequence. It is found by an algorithm called N-best (or 1-best in this case) that sums over all paths through the model with the same location and direction of the helices.

The first few lines gives some statistics:

- Length: the length of the protein sequence.
- Number of predicted TMHs: The number of predicted transmembrane helices.
- Exp number of AAs in TMHs: The expected number of amino acids intransmembrane 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: 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.

Plot of probabilities

The plot shows the posterior probabilities of inside/outside/TM helix. Here one can see possible weak TM helices that were not predicted, and one can get an idea of the certainty of each segment in the prediction.

At the top of the plot (between 1 and 1.2) the N-best prediction is shown.

The plot is obtained by calculating the total probability that a residue sits in helix, inside, or outside summed over all possible paths through the model. Sometimes it seems like the plot and the prediction are contradictory, but that is because the plot shows probabilities for each residue, whereas the prediction is the over-all most probable structure. Therefore the plot should be seen as a complementary source of information.

Below the plot there are links to

- The plot in encapsulated postscript
- A script for making the plot in gnuplot.
- The data for the plot.

Short output format

In the short output format one line is produced for each protein with no graphics. Each line starts with the sequence identifier and then these fields:

- "len=": the length of the protein sequence.
- "ExpAA=": The expected number of amino acids intransmembrane helices (see above).
- `"First60="`: The expected number of amino acids in transmembrane helices in the first 60 amino acids of the protein (see above).
- `"PredHel="`: The number of predicted transmembrane helices by N-best.
- "Topology=": The topology predicted by N-best.

For the example above the short output would be (except that it would be on one line):

```
COX2_BACSU
len=278
ExpAA=68.69
First60=39.89
PredHel=3
Topology=i7-29o44-66i87-109o
```

The topology is given as the position of the transmembrane helices separated by 'i' if the loop is on the inside or 'o' if it is on the outside. The above example '`i7-29o44-66i87-109o' means that it starts on the inside, has a predicted TMH at position 7 to 29, the outside, then a TMH at position 44-66 etc.`

Final remarks

Predicted TM segments in the n-terminal region sometime turn out to be signal peptides.

One of the most common mistakes by the program is to reverse the direction of proteins with one TM segment.

Do not use the program to predict whether a non-membrane protein is cytoplasmic or not.