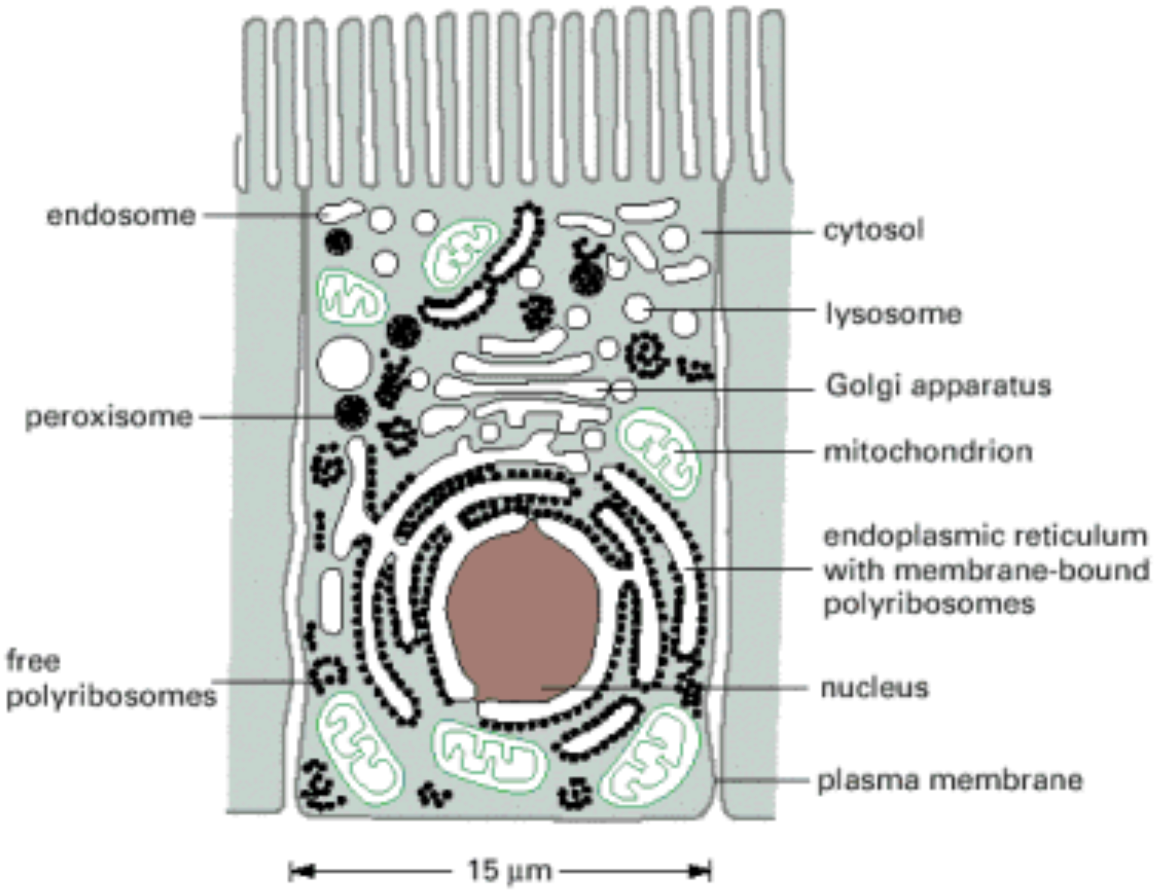
Annotation Presentation
Week 5

Cellular Localization
Data Module

Where in the cell is your protein most likely found?

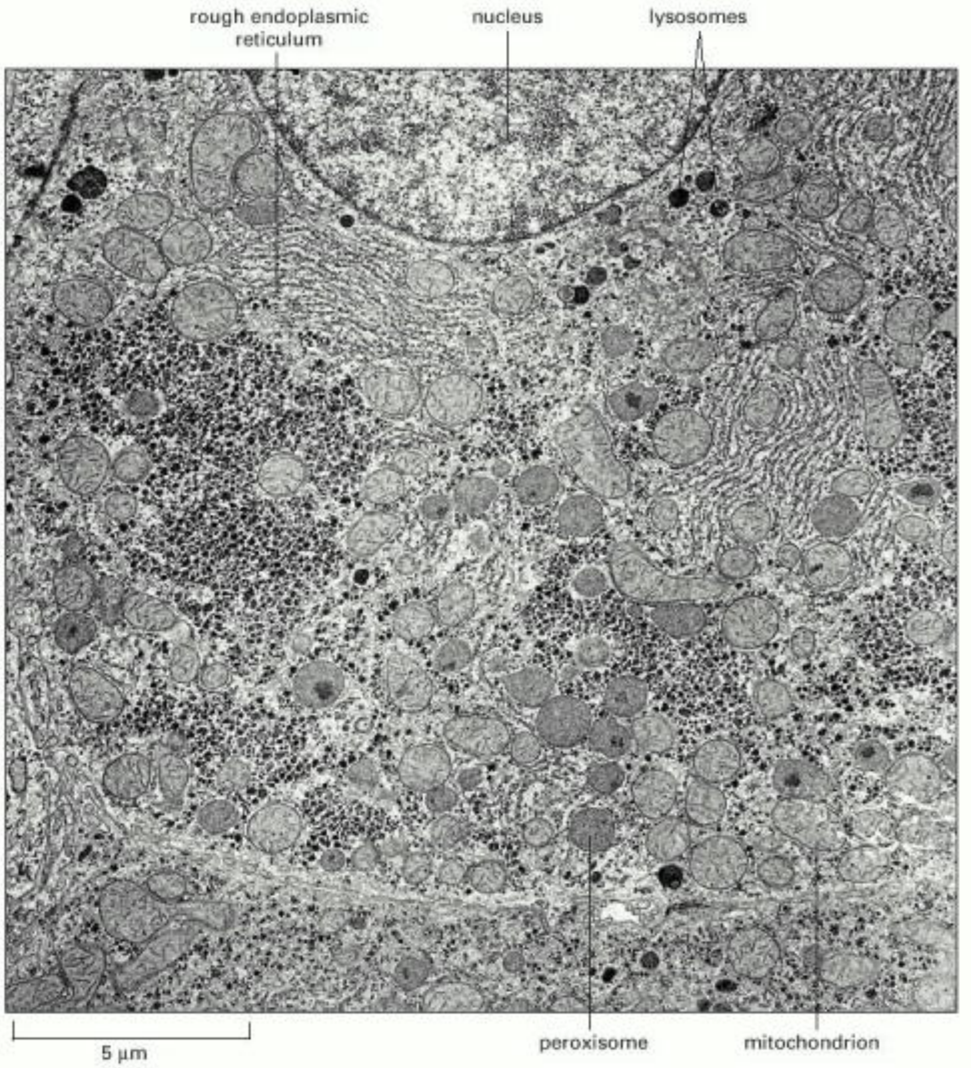
The major intracellular compartments of an animal cell



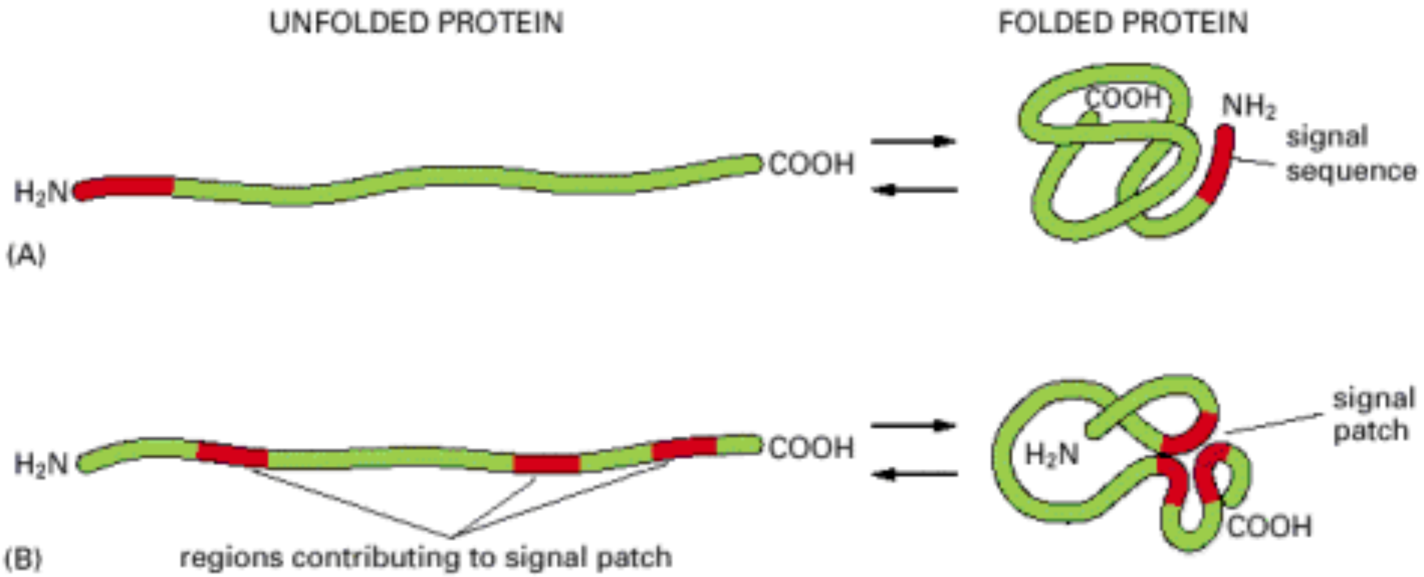
Relative Volumes Occupied by the Major Intracellular Compartments

INTRACELLULAR COMPARTMENT	PERCENTAGE OF TOTAL CELL VOLUME
Cytosol	54
Mitochondria	22
Rough ER cisternae	9
Smooth ER cisternae plus Golgi cisternae	6
Nucleus	6
Peroxisomes	1
Lysosomes	1
Endosomes	1

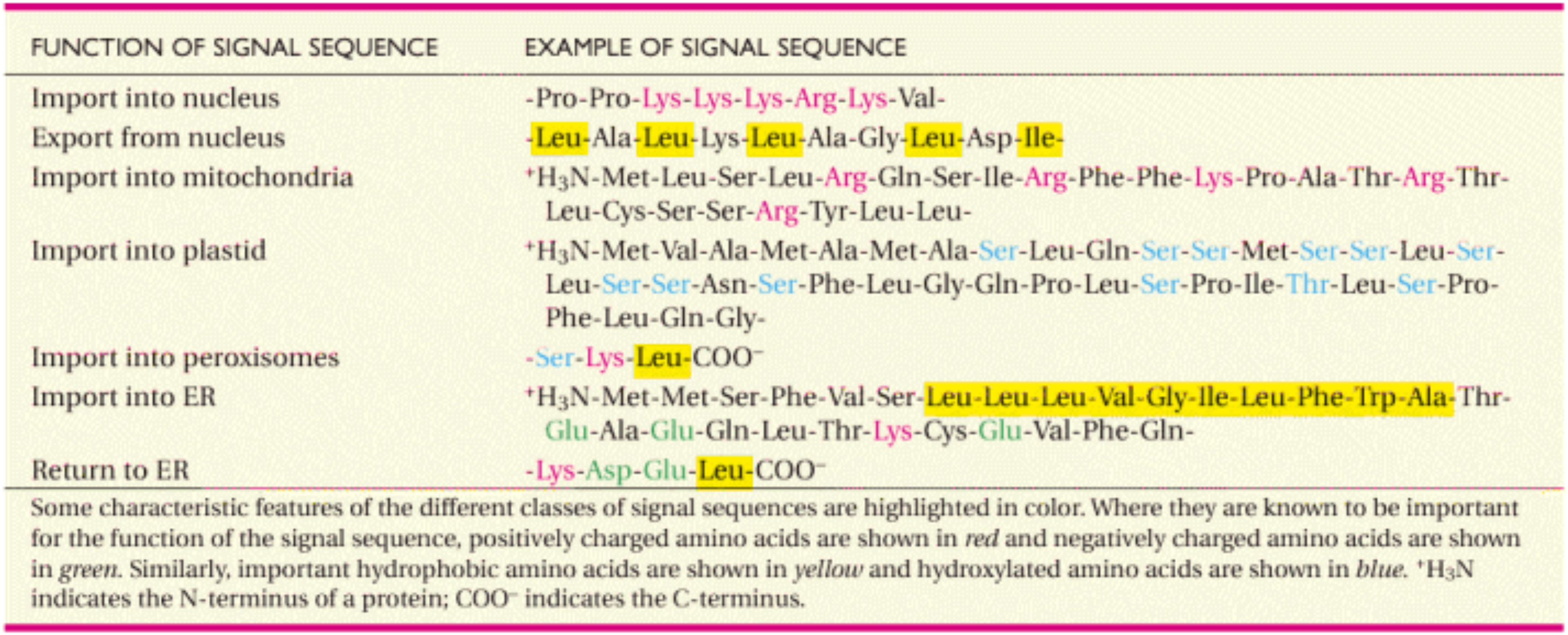
An electron micrograph



Sorting sequences



Some sorting sequences



FUNCTION OF SIGNAL SEQUENCE	EXAMPLE OF SIGNAL SEQUENCE
Import into nucleus	-Pro-Pro-Lys-Lys-Lys-Arg-Lys-Val-
Export from nucleus	-Leu-Ala-Leu-Lys-Leu-Ala-Gly-Leu-Asp-Ile-
Import into mitochondria	$^{+}H_3N$-Met-Leu-Ser-Leu-Arg-Gln-Ser-Ile-Arg-Phe-Phe-Lys-Pro-Ala-Thr-Arg-Thr-Leu-Cys-Ser-Ser-Arg-Tyr-Leu-Leu-
Import into plastid	$^{+}H_3N$-Met-Val-Ala-Met-Ala-Met-Ala-Ser-Leu-Gln-Ser-Ser-Met-Ser-Ser-Leu-Ser-Leu-Ser-Ser-Asn-Ser-Phe-Leu-Gly-Gln-Pro-Leu-Ser-Pro-Ile-Thr-Leu-Ser-Pro-Phe-Leu-Gln-Gly-
Import into peroxisomes	-Ser-Lys-Leu-COO^-
Import into ER	$^{+}H_3N$-Met-Met-Ser-Phe-Val-Ser-Leu-Leu-Leu-Val-Gly-Ile-Leu-Phe-Trp-Ala-Thr-Glu-Ala-Glu-Gln-Leu-Thr-Lys-Cys-Glu-Val-Phe-Gln-
Return to ER	-Lys-Asp-Glu-Leu-COO^-

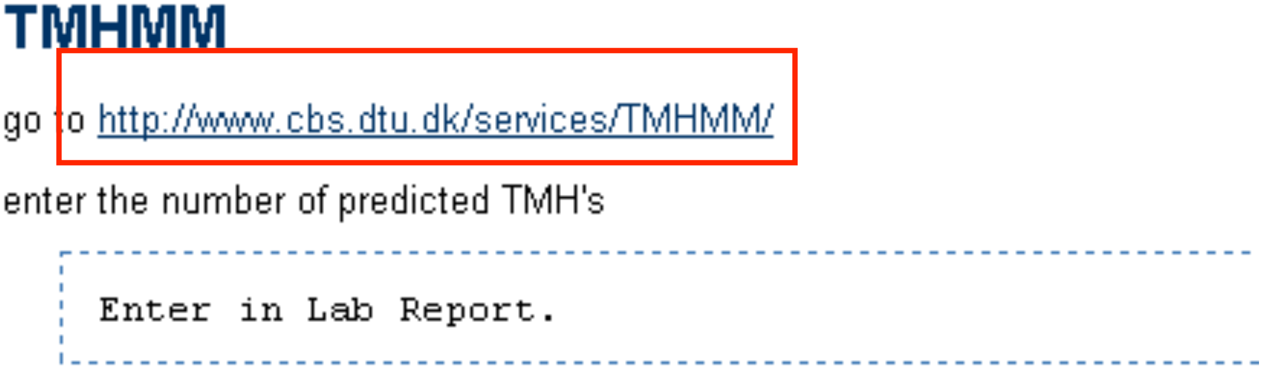
Some characteristic features of the different classes of signal sequences are highlighted in color. Where they are known to be important for the function of the signal sequence, positively charged amino acids are shown in *red* and negatively charged amino acids are shown in *green*. Similarly, important hydrophobic amino acids are shown in *yellow* and hydroxylated amino acids are shown in *blue*. $^{+}H_3N$ indicates the N-terminus of a protein; COO^- indicates the C-terminus.

How do we figure out where proteins are located?

- **Transmembrane Helices Hidden Markov Models (TMHMM)**
  - ✓ Does my protein have transmembrane helices?
- **Signal Peptide (SignalP)**
  - ✓ Does my protein have a sequence of amino acids that target it to a particular place in or outside the cell?
- **PSORT-B**
  - ✓ Where is my protein most likely located? The cytoplasm? The membrane? The periplasm? The cell wall? The extracellular space?
- **Phobius**
  - ✓ Does my protein have transmembrane helices & signal peptides? Do these results agree with TMHMM and SignalP?

Transmembrane Helices Hidden Markov Models (TMHMM)

- A Hidden Markov Model is a probabilistic model developed from observed sequences of proteins of a known function.
- TMHMM is a tool used to predict the presence of transmembrane helices in proteins. The results will indicate the segments of the protein that lie inside, outside or within the membrane.



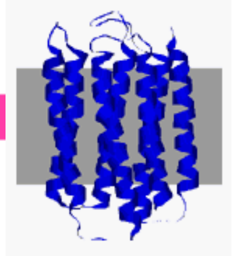
TMHMM

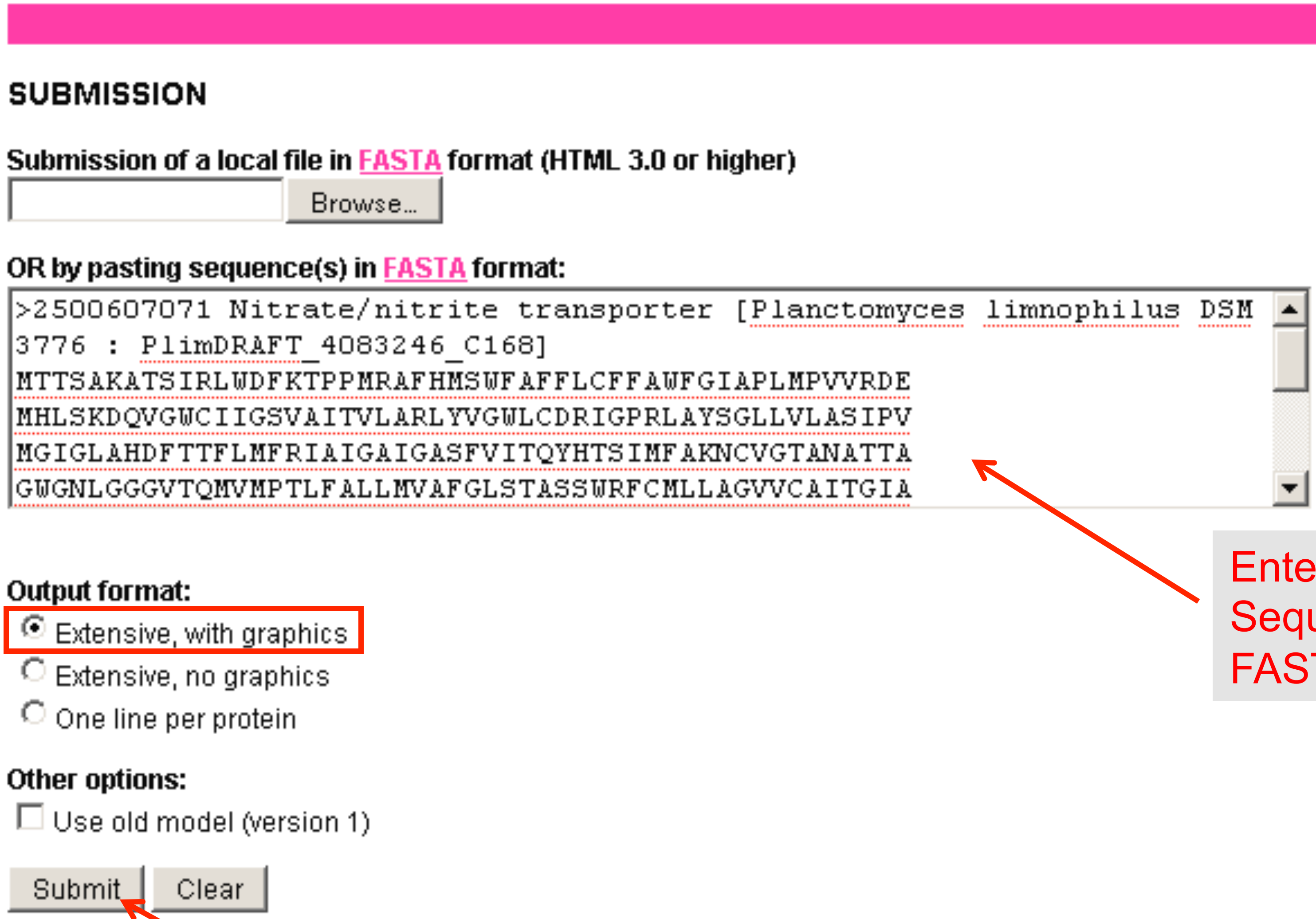
go to http://www.cbs.dtu.dk/services/TMHMM/

enter the number of predicted TMH's

```
Enter in Lab Report.
```

TMHMM Database Search





SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

OR by pasting sequence(s) in FASTA format:

```
>2500607071 Nitrate/nitrite transporter [Planctomyces limnophilus DSM
3776 : PlimDRAFT_4083246_C168]
MTTSAKATSIRLWDFKTPPMRAFHMSWFAFFLCFFAWFGIAPLMPVVRDE
MHLSKDQVGWCIIGSVAITVLARLYVGWLCDRIGPRLAYSGLLVLASIPV
MGIGLAHDFTTFLMFRIAIGAIGASFVITQYHTSIMFAKNCVGTANATTA
GWGNLGGGVTQMVMPTLFALLMVAFGLSTASSWRFCMLLAGVVCAITGIA
```

Output format:

● Extensive, with graphics

○ Extensive, no graphics

○ One line per protein

Other options:

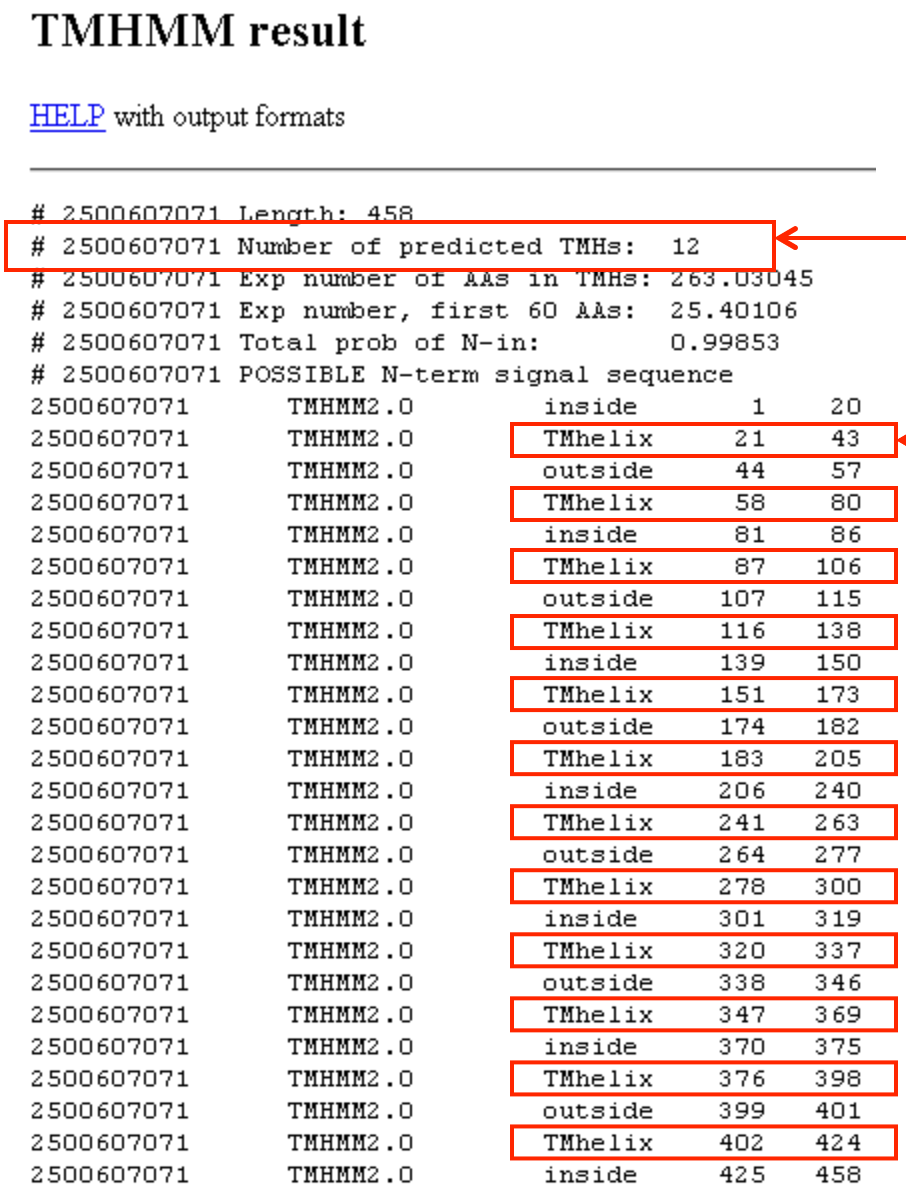
□ Use old model (version 1)

Enter “Protein Sequence” in FASTA format

“CLICK
”

**Make sure Javascript is enabled on your computer to read output

TMHMM result



TMHMM result

HELP with output formats

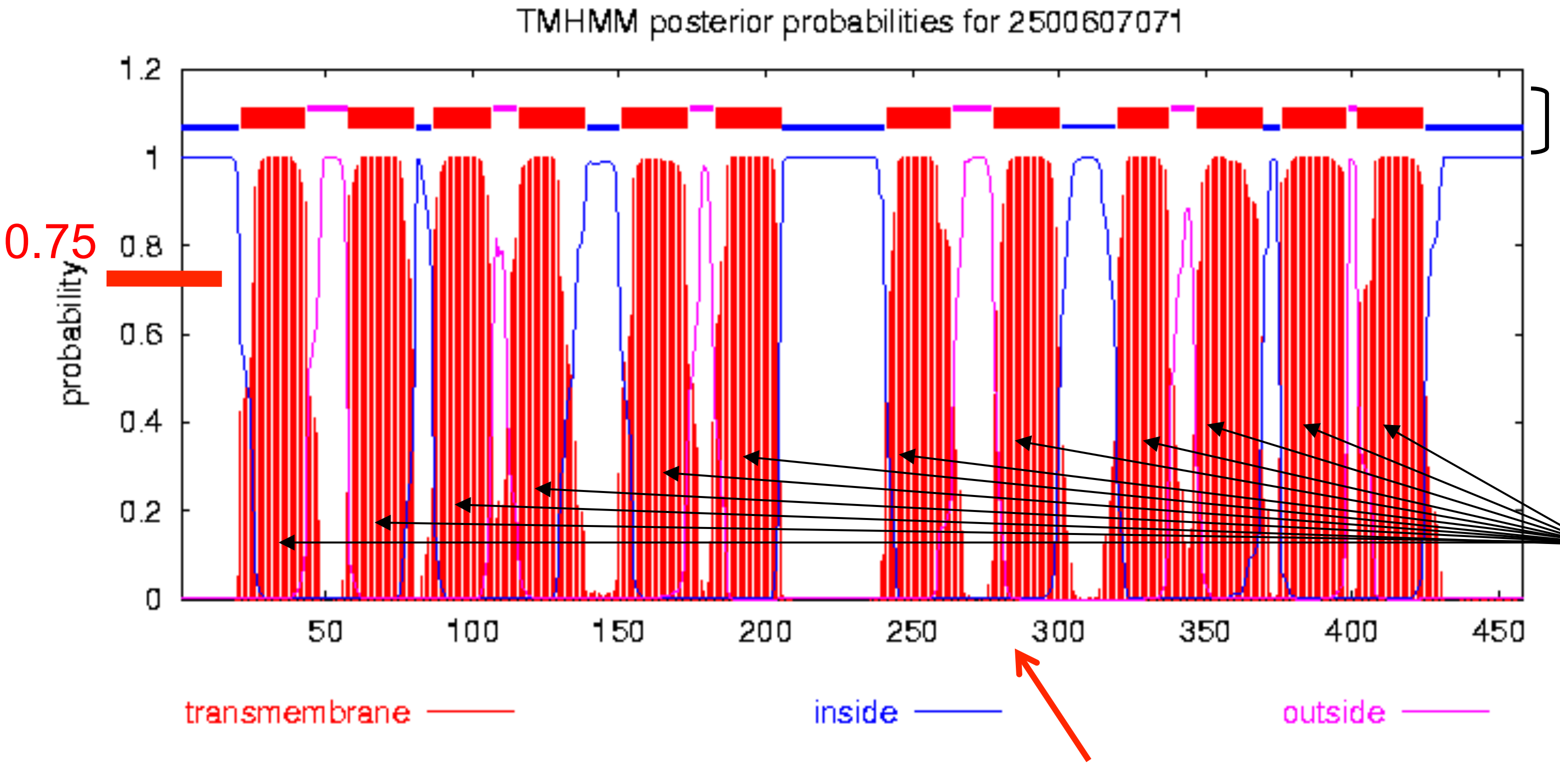
```
# 2500607071 Length: 458
# 2500607071 Number of predicted TMHs:  12
# 2500607071 Exp number of AAs in TMHs: 263.03045
# 2500607071 Exp number, first 60 AAs:  25.40106
# 2500607071 Total prob of N-in:        0.99853
# 2500607071 POSSIBLE N-term signal sequence
2500607071      TMHMM2.0        inside       1     20
2500607071      TMHMM2.0        TMhelix     21     43
2500607071      TMHMM2.0        outside     44     57
2500607071      TMHMM2.0        TMhelix     58     80
2500607071      TMHMM2.0        inside      81     86
2500607071      TMHMM2.0        TMhelix     87    106
2500607071      TMHMM2.0        outside    107    115
2500607071      TMHMM2.0        TMhelix    116    138
2500607071      TMHMM2.0        inside     139    150
2500607071      TMHMM2.0        TMhelix    151    173
2500607071      TMHMM2.0        outside    174    182
2500607071      TMHMM2.0        TMhelix    183    205
2500607071      TMHMM2.0        inside     206    240
2500607071      TMHMM2.0        TMhelix    241    263
2500607071      TMHMM2.0        outside    264    277
2500607071      TMHMM2.0        TMhelix    278    300
2500607071      TMHMM2.0        inside     301    319
2500607071      TMHMM2.0        TMhelix    320    337
2500607071      TMHMM2.0        outside    338    346
2500607071      TMHMM2.0        TMhelix    347    369
2500607071      TMHMM2.0        inside     370    375
2500607071      TMHMM2.0        TMhelix    376    398
2500607071      TMHMM2.0        outside    399    401
2500607071      TMHMM2.0        TMhelix    402    424
2500607071      TMHMM2.0        inside     425    458
```

Predicted number of TMHs (transmembrane helices)

Boundaries for THM amino acids

Copy/paste this information into the box in your lab notebook

Interpreting the TMHMM plot



Schematic that summarizes discrete regions within the protein; not probability.

The 12 predicted TMHs

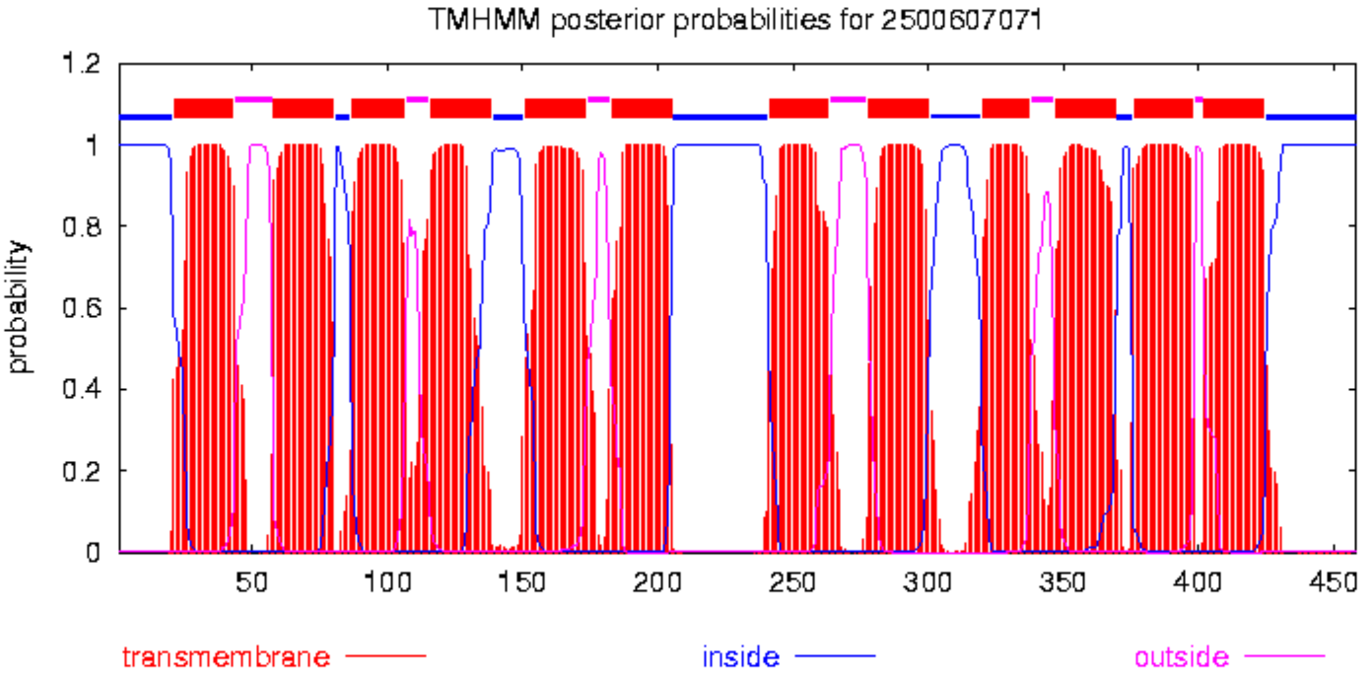
Residue number

X-axis: the amino acid number

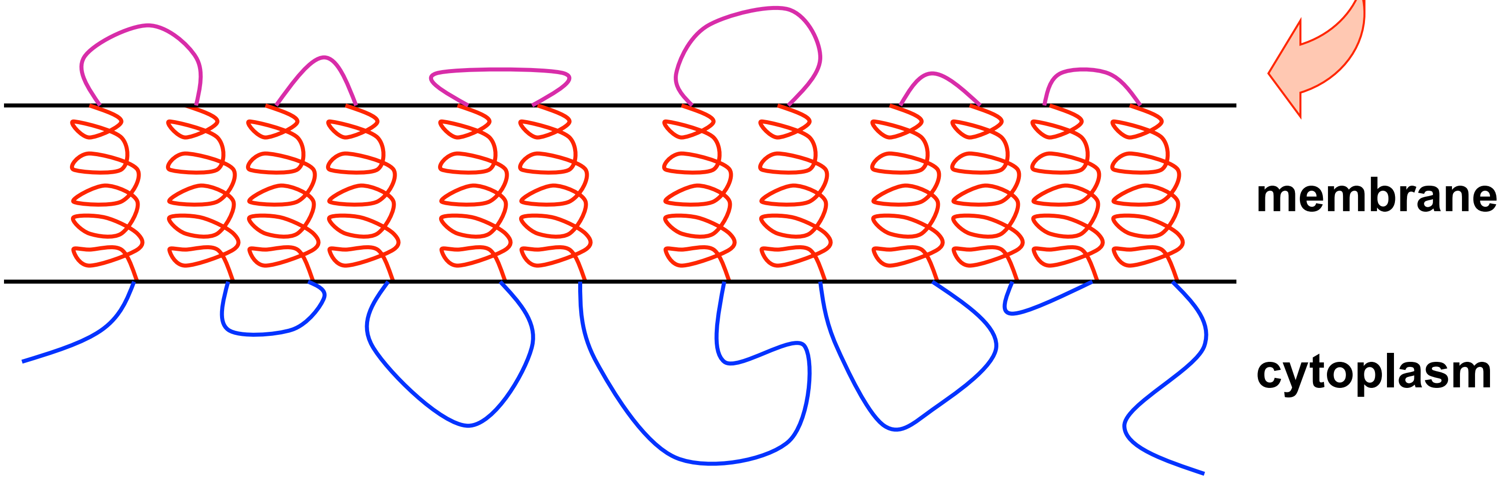
Y-axis: the probability that the amino acid is located within the membrane, outside the cell, or in the cytoplasm

Ex: If probability >0.75, then result is significant. The maximum probability is 1, so the probability that amino acids #1-#20 are "inside" is 100%

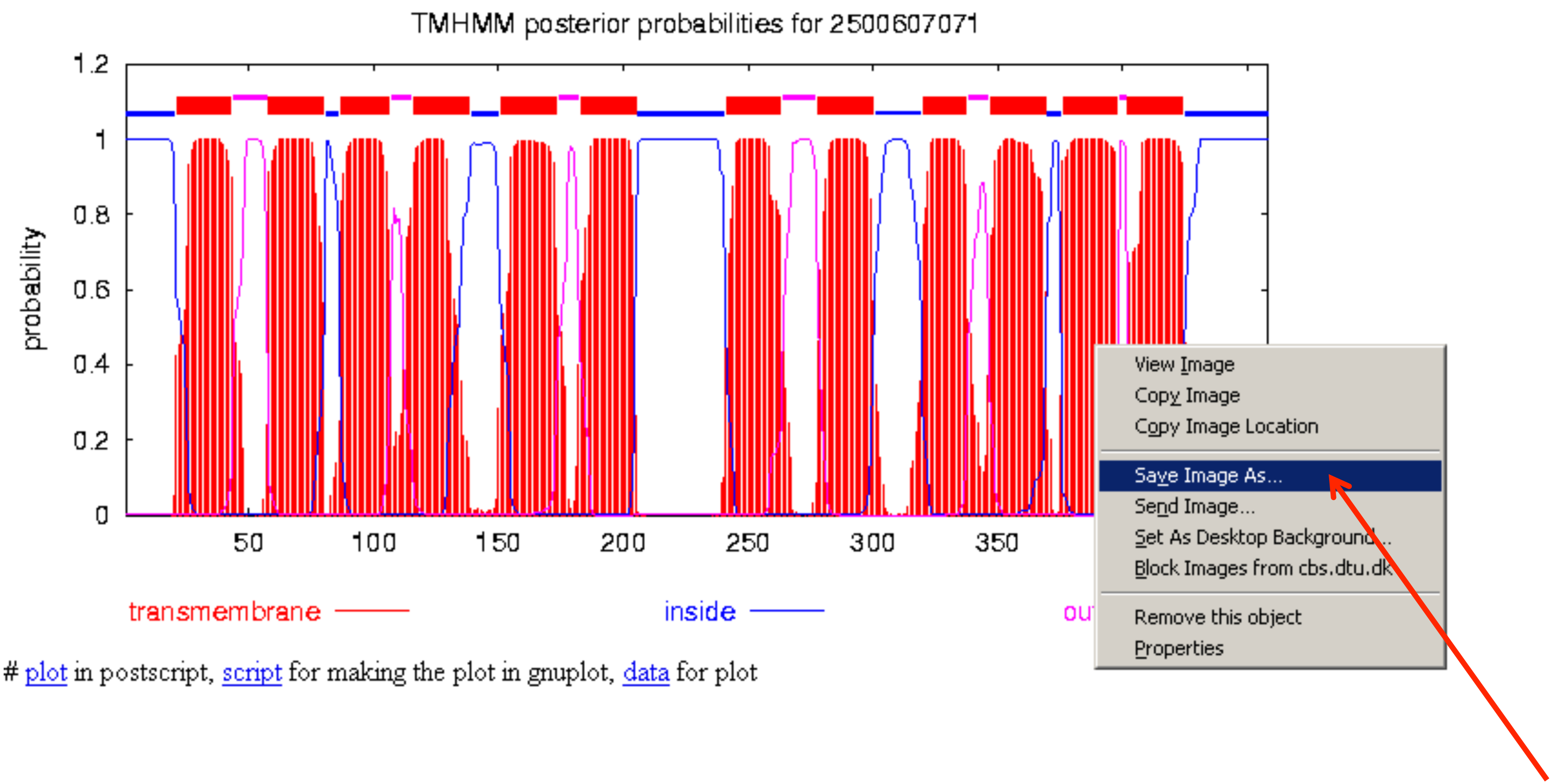
Transmembrane	🟥
Inside (cytoplasm)	🟦
Outside (extracellular, periplasm)	🟪



By analyzing the probabilities shown on the plot, you can determine where segments within the protein are located.



Inserting the TMHMM plot into your notebook

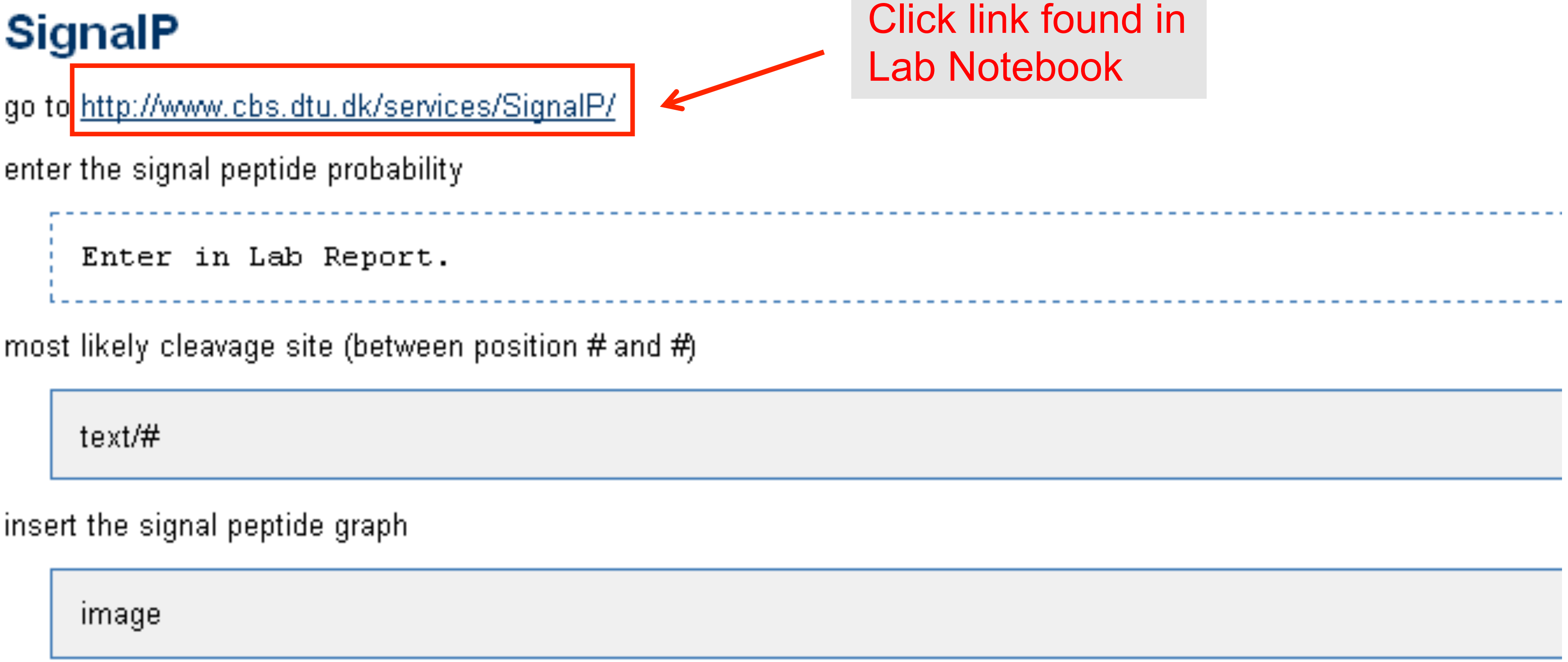


plot in postscript, script for making the plot in gnuplot, data for plot

Save image in GIF format to your computer and insert into Lab Notebook

SignalP

- A Signal Peptide (SignalP) is a series of amino acids in the polypeptide that directs the protein to its proper cellular location
    - **Ex:** Single TMH at N-terminus of protein that gets cleaved by proteases once inserted into membrane



SignalP

go to http://www.cbs.dtu.dk/services/SignalP/

enter the signal peptide probability

```
Enter in Lab Report.
```

most likely cleavage site (between position # and #)

text/#

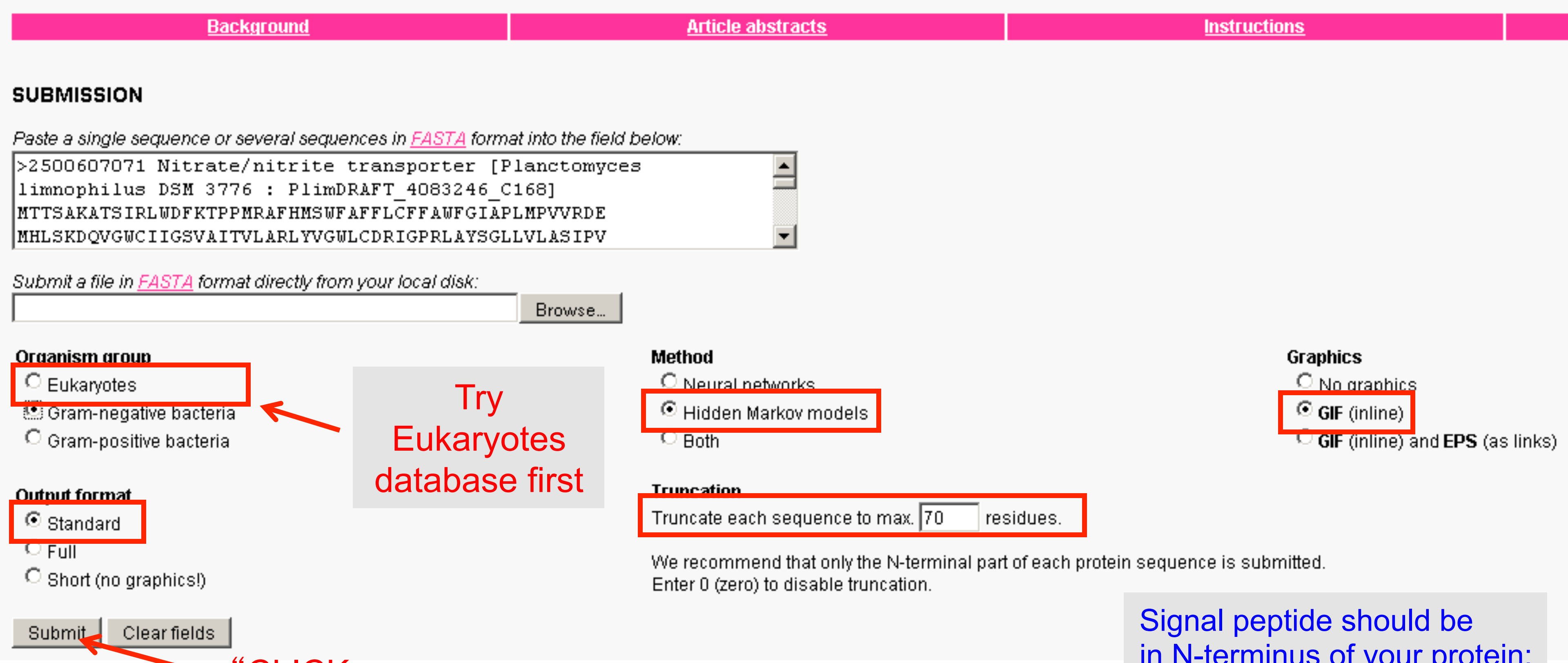
insert the signal peptide graph

image

Click link found in Lab Notebook

Locating proteins in the cell using TargetP, SignalP, and related tools
Olof Emanuelsson, Søren Brunak, Gunnar von Heijne, Henrik Nielsen
Nature Protocols **2**, 953-971 (2007).

SignalP Database Search



SUBMISSION

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

```
>2500607071 Nitrate/nitrite transporter [Planctomyces
limnophilus DSM 3776 : PlimDRAFT_4083246_C168]
MTTSAKATSIRLWDFKTPPMRAFHMSWFAFFLCFFAWFGIAPLMPVVRDE
MHLSKDQVGWCIIGSVAITVLARLYVGWLCDRIGPRLAYSGLLVLASIPV
```

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

Organism group

○ Eukaryotes
● Gram-negative bacteria
○ Gram-positive bacteria

Method

○ Neural networks
● Hidden Markov models
○ Both

Graphics

○ No graphics
● **GIF** (inline)
○ **GIF** (inline) and **EPS** (as links)

Output format

● Standard
○ Full
○ Short (no graphics!)

Truncation

Truncate each sequence to max. 70 residues.

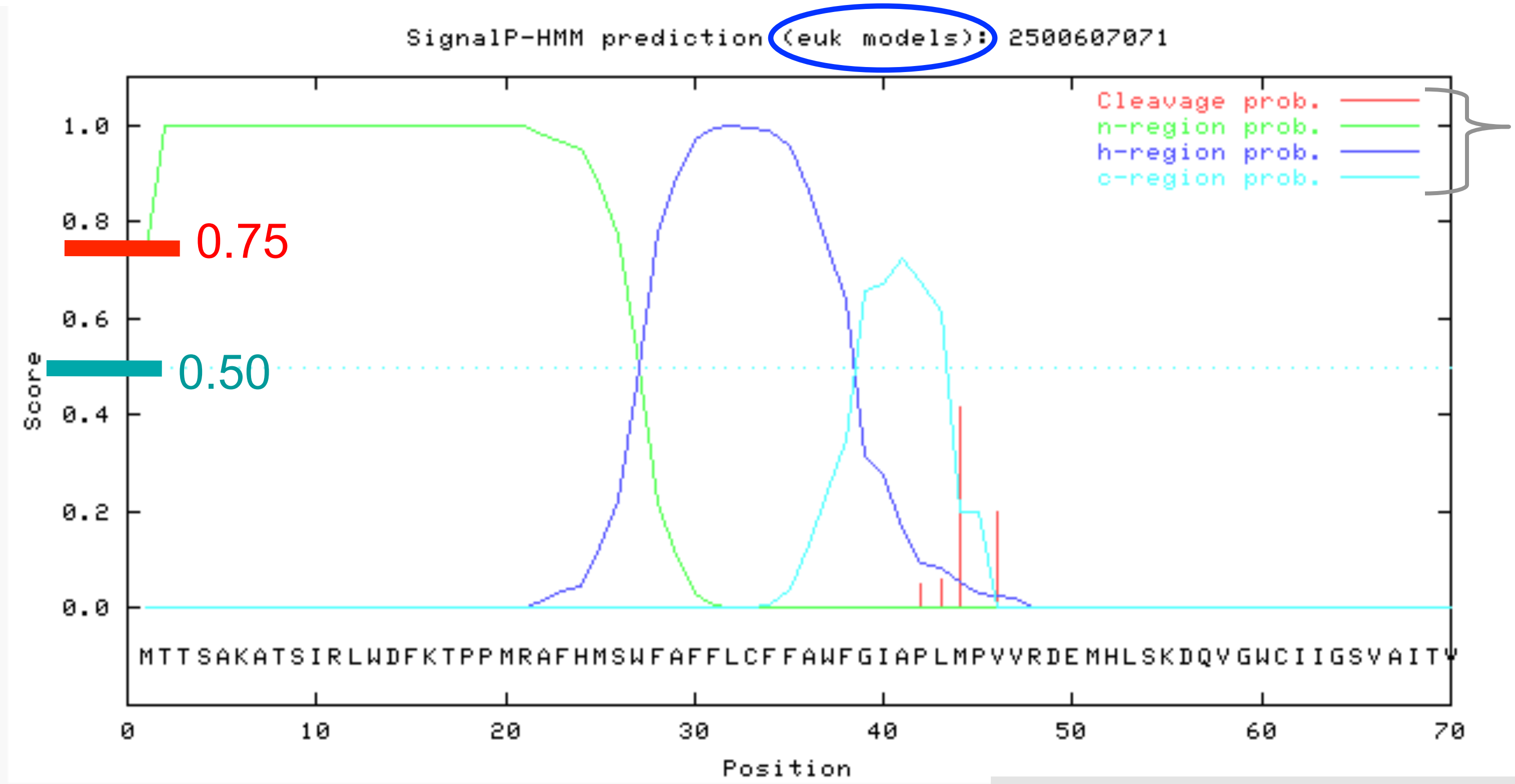
We recommend that only the N-terminal part of each protein sequence is submitted.
Enter 0 (zero) to disable truncation.

Try Eukaryotes database first

"CLICK"

Signal peptide should be in N-terminus of your protein;
No need to scan full length

Signal P (Eukaryote)



- Signal peptide cleaved by proteolytic enzymes
- N-terminus of signal peptide
- Hydrophobic Region (TMH)
- C-terminus of signal peptide

What would you conclude for this protein?

If the probability is >0.50, then the results suggest that your gene encodes a signal peptide. Higher confidence in probability score if >0.75

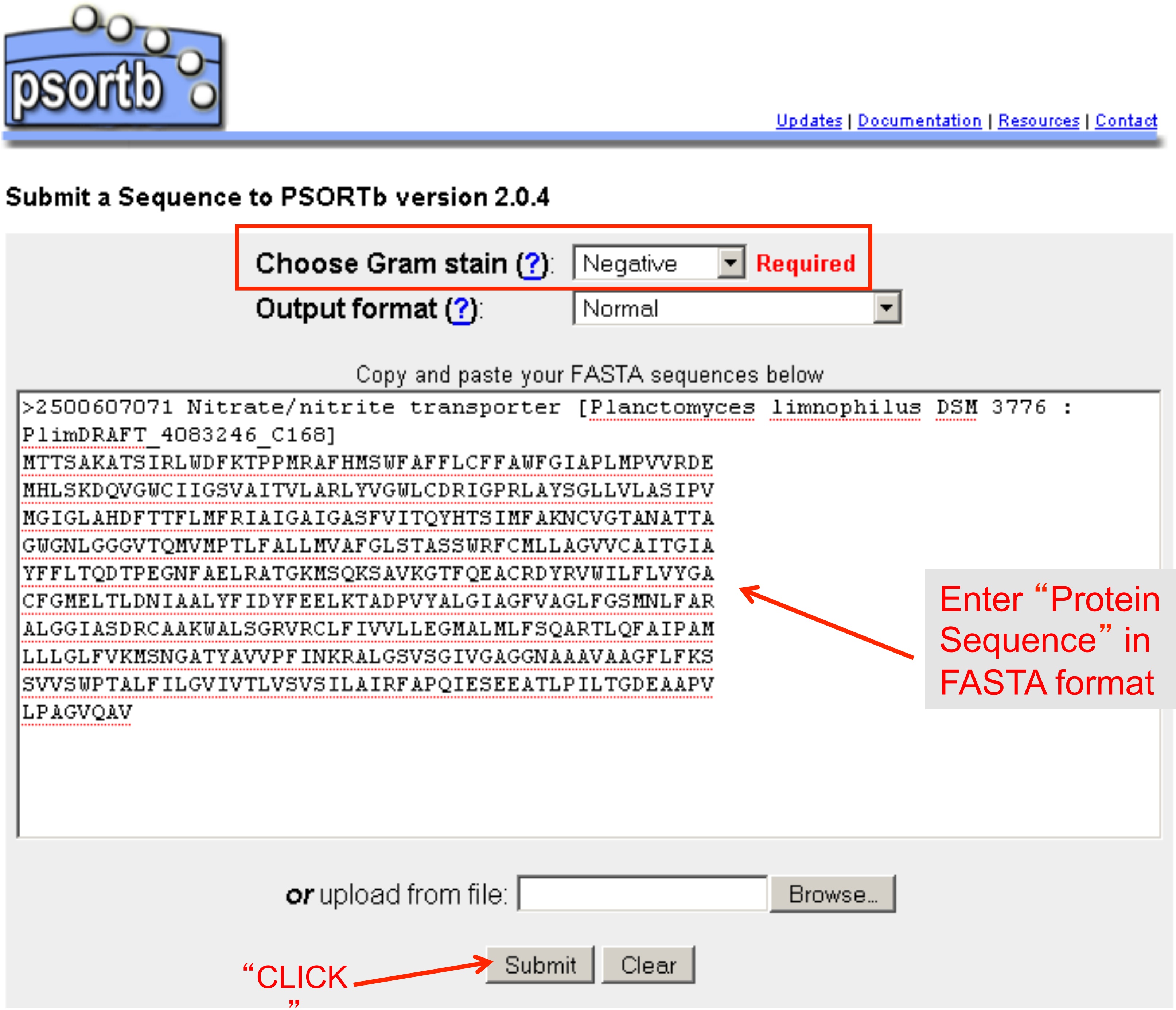
```
# data

>2500607071
Prediction: Signal peptide
Signal peptide probability: 0.732
Signal anchor probability: 0.267
Max cleavage site probability: 0.417 between pos. 43 and 44
```

Possible protease cleavage site if probability > 0.75

Prediction of protein sorting

- Psort web server: http://psort.nibb.ac.jp/
  - prediction of protein localization sites in cells from their primary amino acid sequence



Submit a Sequence to PSORTb version 2.0.4

Choose Gram stain (?): Negative ▾ **Required**

Output format (?): Normal ▾

Copy and paste your FASTA sequences below

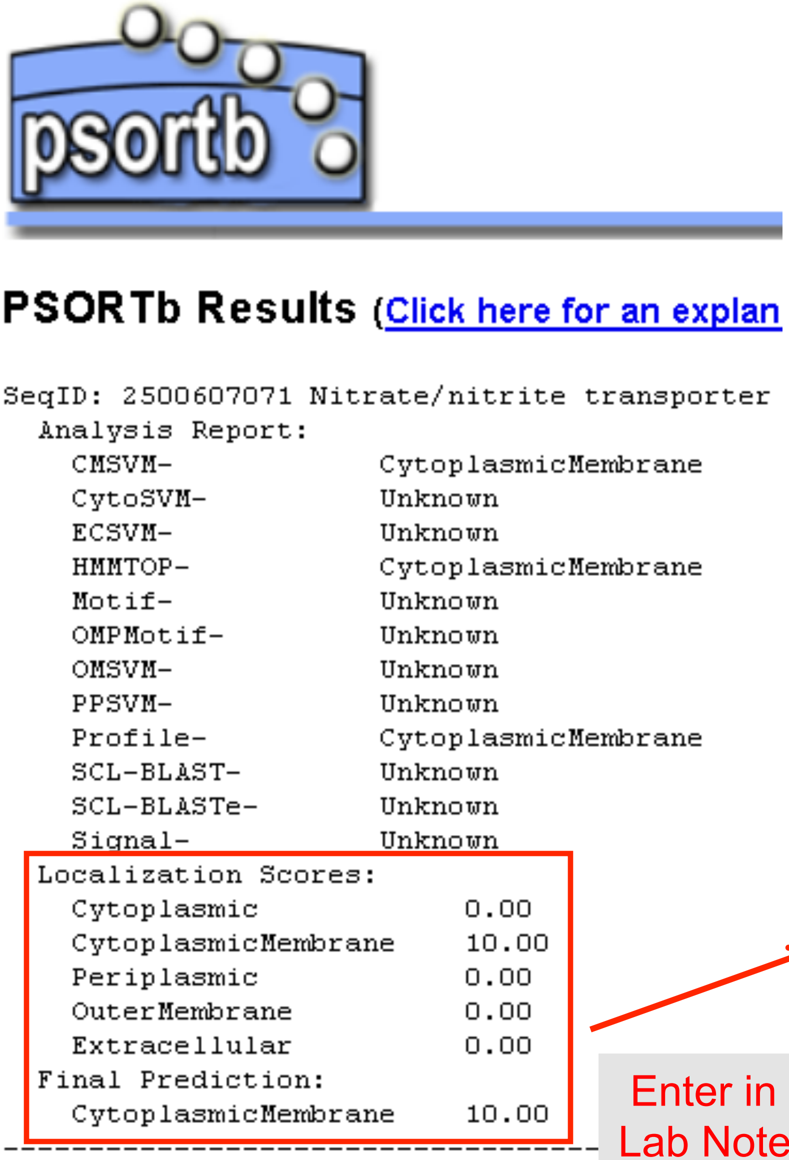
```
>2500607071 Nitrate/nitrite transporter [Planctomyces limnophilus DSM 3776 :
PlimDRAFT_4083246_C168]
MTTSAKATSIRLWDFKTPPMRAFHMSWFAFFLCFFAWFGIAPLMPVVRDE
MHLSKDQVGWCIIGSVAITVLARLYVGWLCDRIGPRLAYSGLLVLASIPV
MGIGLAHDFTTFLMFRIAIGAIGASFVITQYHTSIMFAKNCVGTANATTA
GWGNLGGGVTQMVMPTLFALLMVAFGLSTASSWRFCMLLAGVVCAITGIA
YFFLTQDTPEGNFAELRATGKMSQKSAVKGTFQEACRDYRVWILFLVYGA
CFGMELTLDNIAALYFIDYFEELKTADPVYALGIAGFVAGLFGSMNLFAR
ALGGIASDRCAAKWALSGRVRCLFIVVLLEGMALMLFSQARTLQFAIPAM
LLLGLFVKMSNGATYAVVPFINKRALGSVSGIVGAGGNAAAVAAGFLFKS
SVVSWPTALFILGVIVTLVSVSILAIRFAPQIESEEATLPILTGDEAAPV
LPAGVQAV
```

or upload from file: Browse...

Enter "Protein Sequence" in FASTA format

"CLICK"

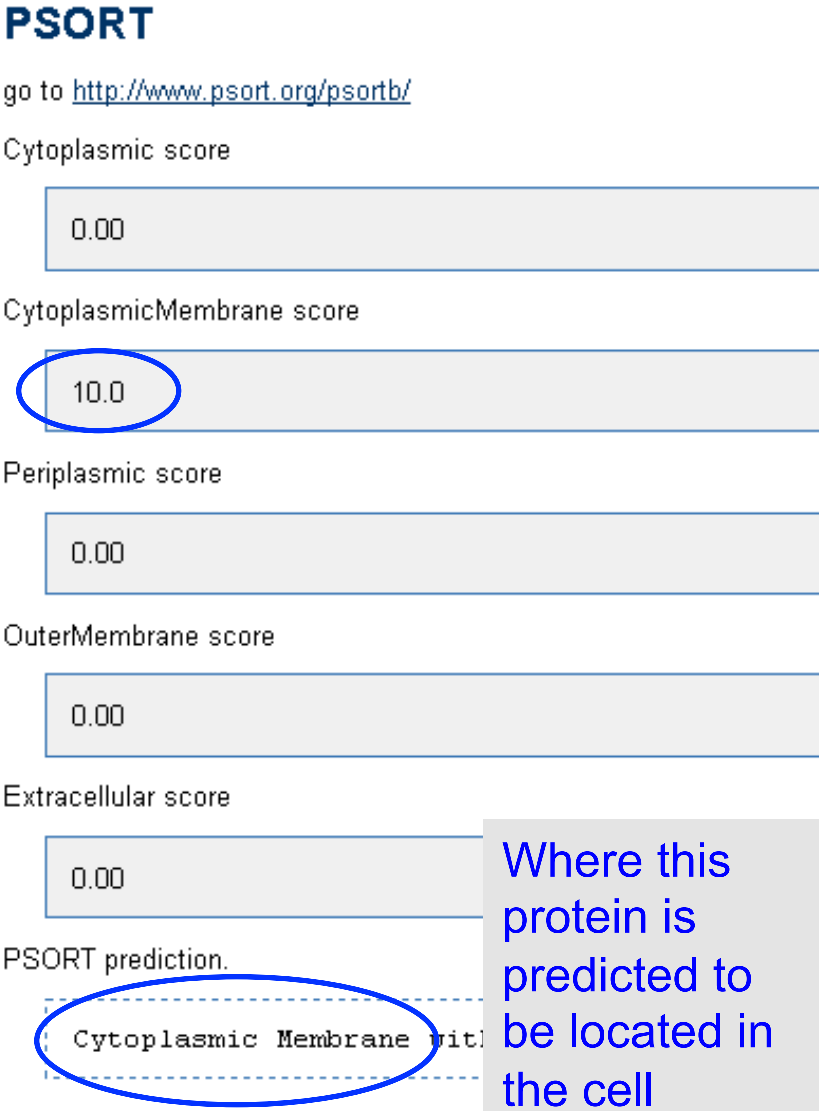
Recording results in your Lab Notebook



PSORTb Results (Click here for an explan

```
SeqID: 2500607071 Nitrate/nitrite transporter
  Analysis Report:
    CMSVM-              CytoplasmicMembrane
    CytoSVM-            Unknown
    ECSVM-              Unknown
    HMMTOP-             CytoplasmicMembrane
    Motif-              Unknown
    OMPMotif-           Unknown
    OMSVM-              Unknown
    PPSVM-              Unknown
    Profile-            CytoplasmicMembrane
    SCL-BLAST-          Unknown
    SCL-BLASTe-         Unknown
    Signal-             Unknown
  Localization Scores:
    Cytoplasmic                 0.00
    CytoplasmicMembrane         10.00
    Periplasmic                 0.00
    OuterMembrane               0.00
    Extracellular               0.00
  Final Prediction:
    CytoplasmicMembrane         10.00
-----------------------------------
```

Enter in your Lab Notebook



PSORT

go to http://www.psort.org/psortb/

Cytoplasmic score

0.00

CytoplasmicMembrane score

10.0

Periplasmic score

0.00

OuterMembrane score

0.00

Extracellular score

0.00

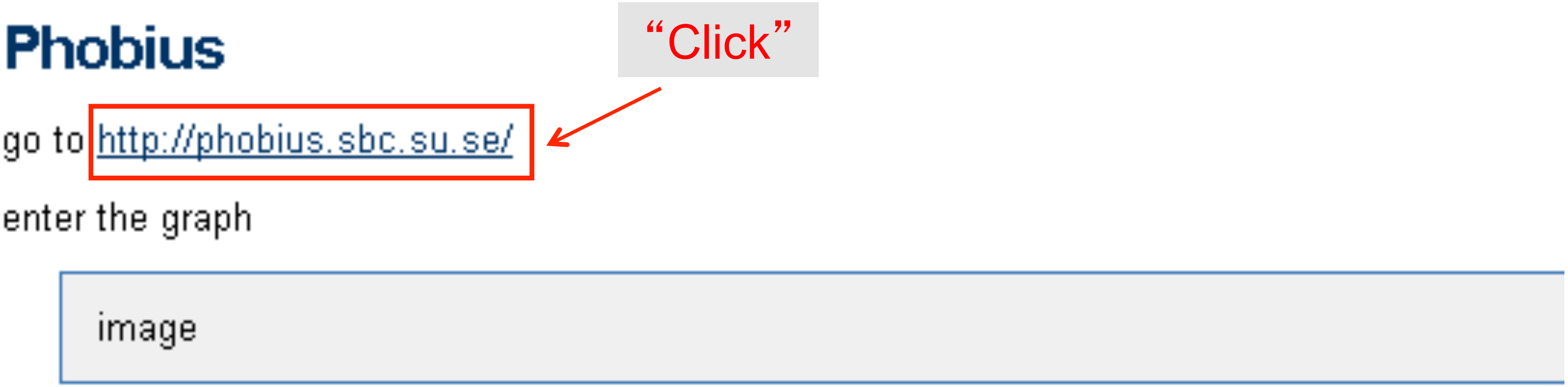
PSORT prediction.

Cytoplasmic Membrane wit

Where this protein is predicted to be located in the cell

Phobius

- Graphical output
- Combination of **transmembrane topology** (TMHMM) and **signal peptide predictor** (SignalP)



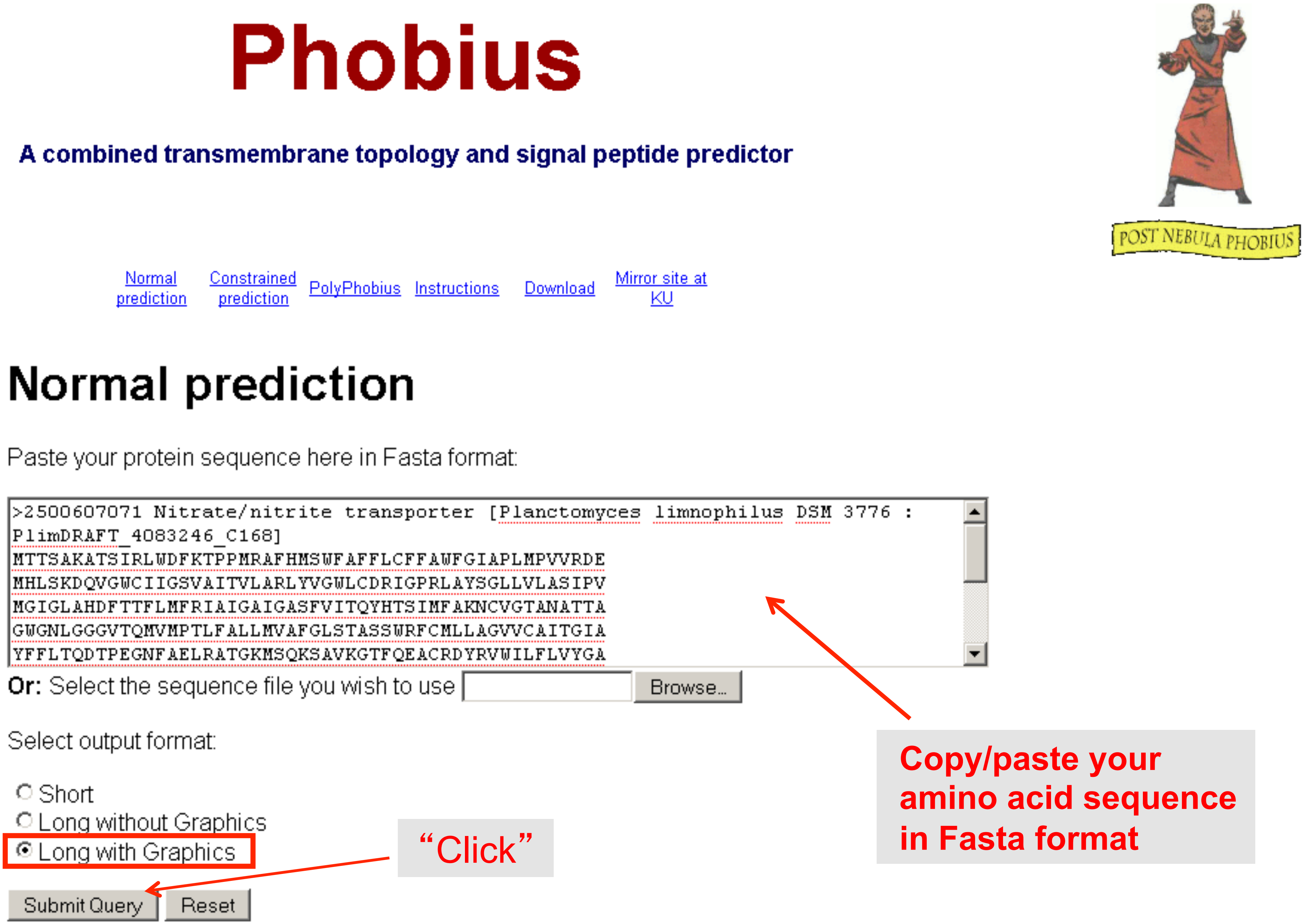
Phobius

go to http://phobius.sbc.su.se/

enter the graph

image

“Click”



Phobius

A combined transmembrane topology and signal peptide predictor

Normal prediction

Paste your protein sequence here in Fasta format:

```
>2500607071 Nitrate/nitrite transporter [Planctomyces limnophilus DSM 3776 :
PlimDRAFT_4083246_C168]
MTTSAKATSIRLWDFKTPPMRAFHMSWFAFFLCFFAWFGIAPLMPVVRDE
MHLSKDQVGWCIIGSVAITVLARLYVGWLCDRIGPRLAYSGLLVLASIPV
MGIGLAHDFTTFLMFRIAIGAIGASFVITQYHTSIMFAKNCVGTANATTA
GWGNLGGGVTQMVMPTLFALLMVAFGLSTASSWRFCMLLAGVVCAITGIA
YFFLTQDTPEGNFAELRATGKMSQKSAVKGTFQEACRDYRVWILFLVYGA
```

Or: Select the sequence file you wish to use

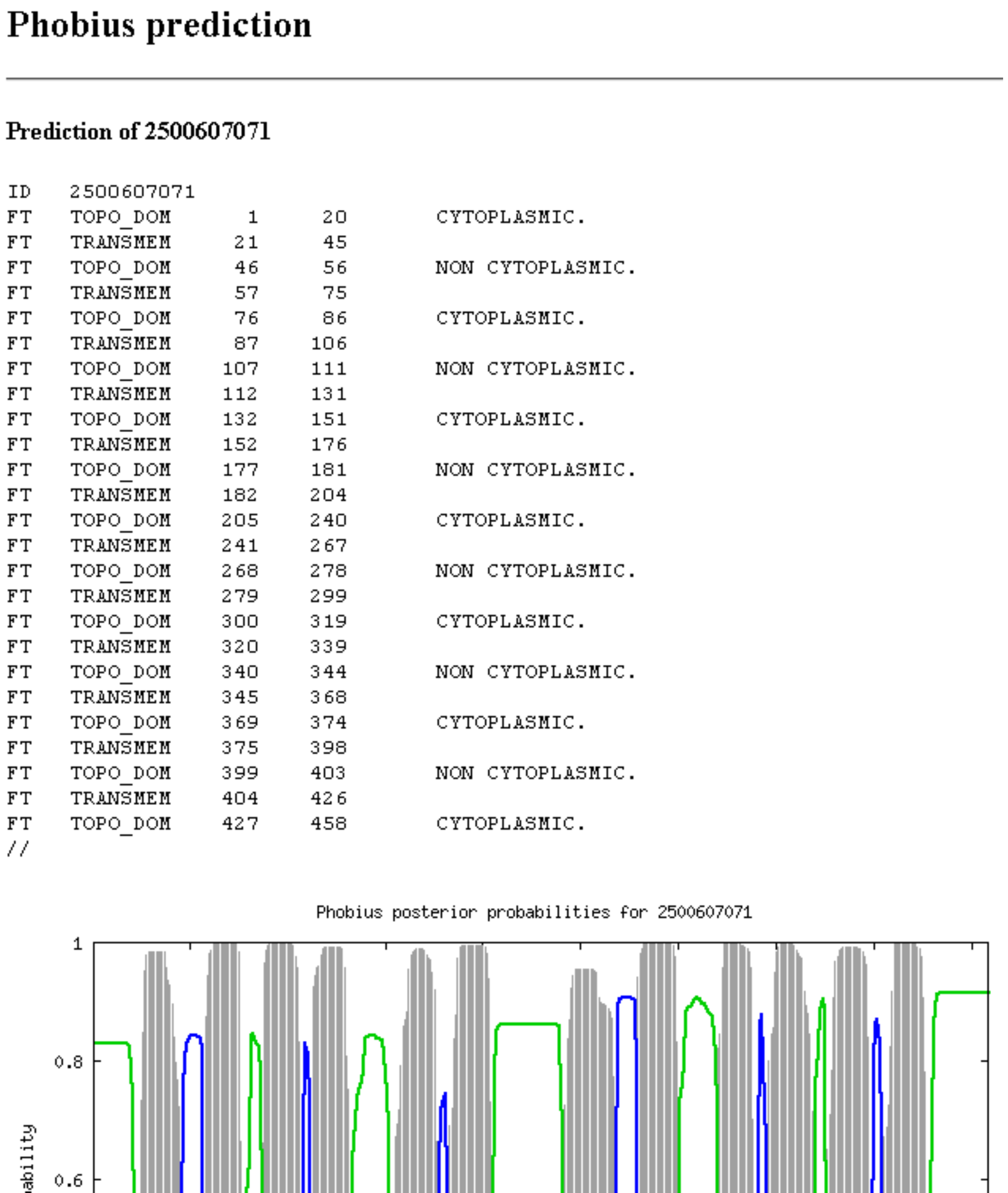
Select output format:

○ Short
○ Long without Graphics
● Long with Graphics

“Click”

Copy/paste your amino acid sequence in Fasta format

Query Results



Phobius prediction

Prediction of 2500607071

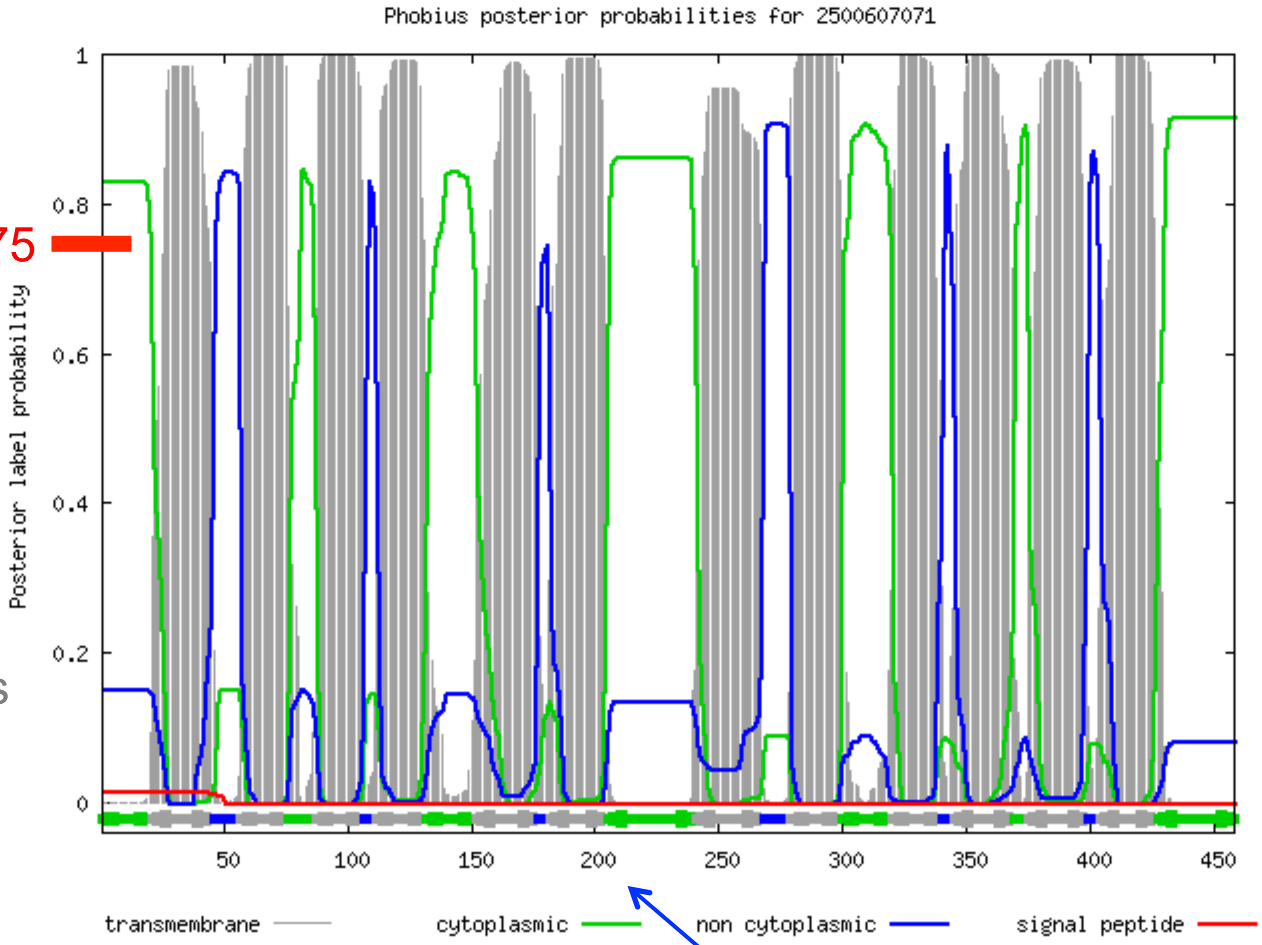
```
ID   2500607071
FT   TOPO_DOM      1     20       CYTOPLASMIC.
FT   TRANSMEM     21     45
FT   TOPO_DOM     46     56       NON CYTOPLASMIC.
FT   TRANSMEM     57     75
FT   TOPO_DOM     76     86       CYTOPLASMIC.
FT   TRANSMEM     87    106
FT   TOPO_DOM    107    111       NON CYTOPLASMIC.
FT   TRANSMEM    112    131
FT   TOPO_DOM    132    151       CYTOPLASMIC.
FT   TRANSMEM    152    176
FT   TOPO_DOM    177    181       NON CYTOPLASMIC.
FT   TRANSMEM    182    204
FT   TOPO_DOM    205    240       CYTOPLASMIC.
FT   TRANSMEM    241    267
FT   TOPO_DOM    268    278       NON CYTOPLASMIC.
FT   TRANSMEM    279    299
FT   TOPO_DOM    300    319       CYTOPLASMIC.
FT   TRANSMEM    320    339
FT   TOPO_DOM    340    344       NON CYTOPLASMIC.
FT   TRANSMEM    345    368
FT   TOPO_DOM    369    374       CYTOPLASMIC.
FT   TRANSMEM    375    398
FT   TOPO_DOM    399    403       NON CYTOPLASMIC.
FT   TRANSMEM    404    426
FT   TOPO_DOM    427    458       CYTOPLASMIC.
//
```

Text listing predicted locations of TMHs, intervening loops, and signal peptide

Graphical summary

Interpreting the Phobius Plot

- Y axis shows probability



0.75

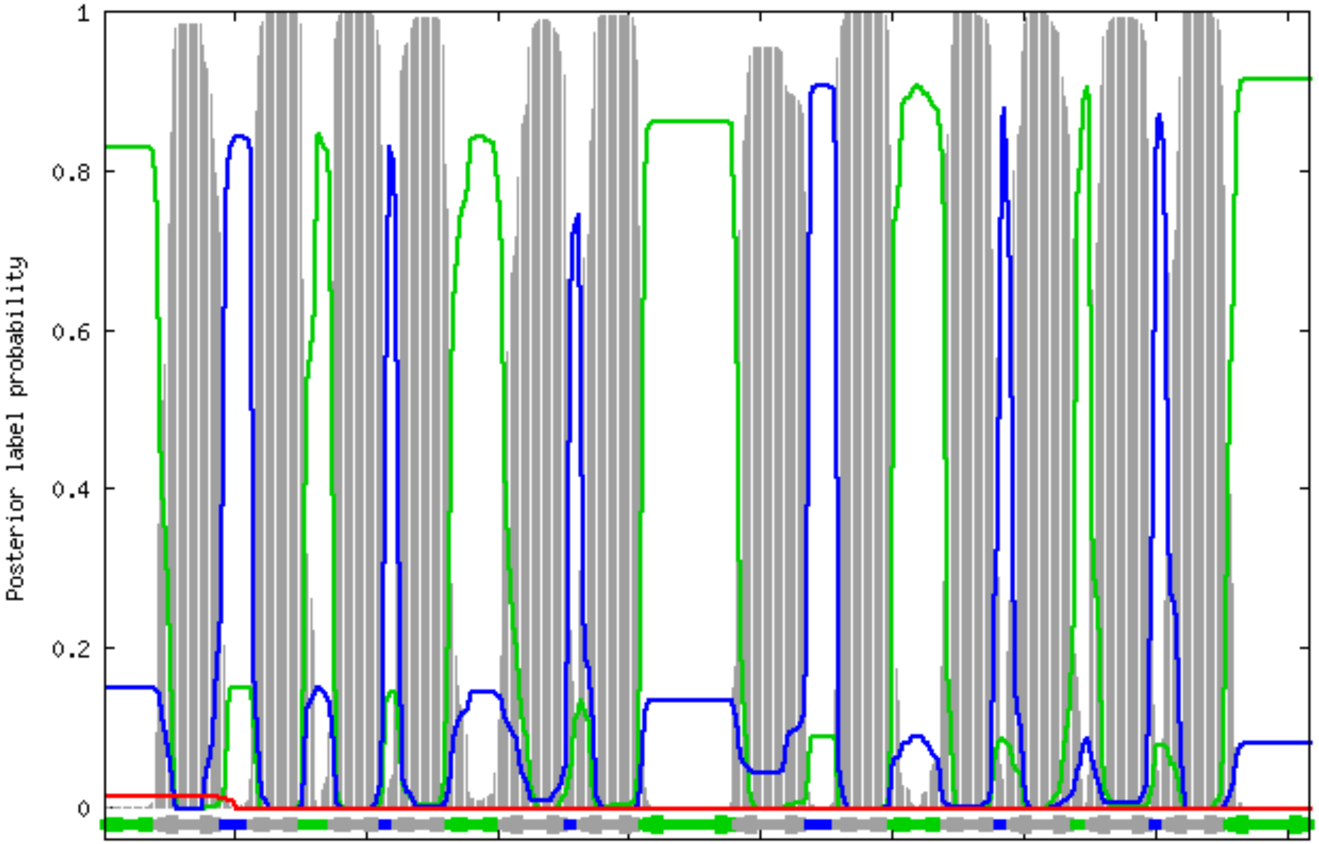
- X axis shows amino acid position

■ GRAY regions = transmembrane helices

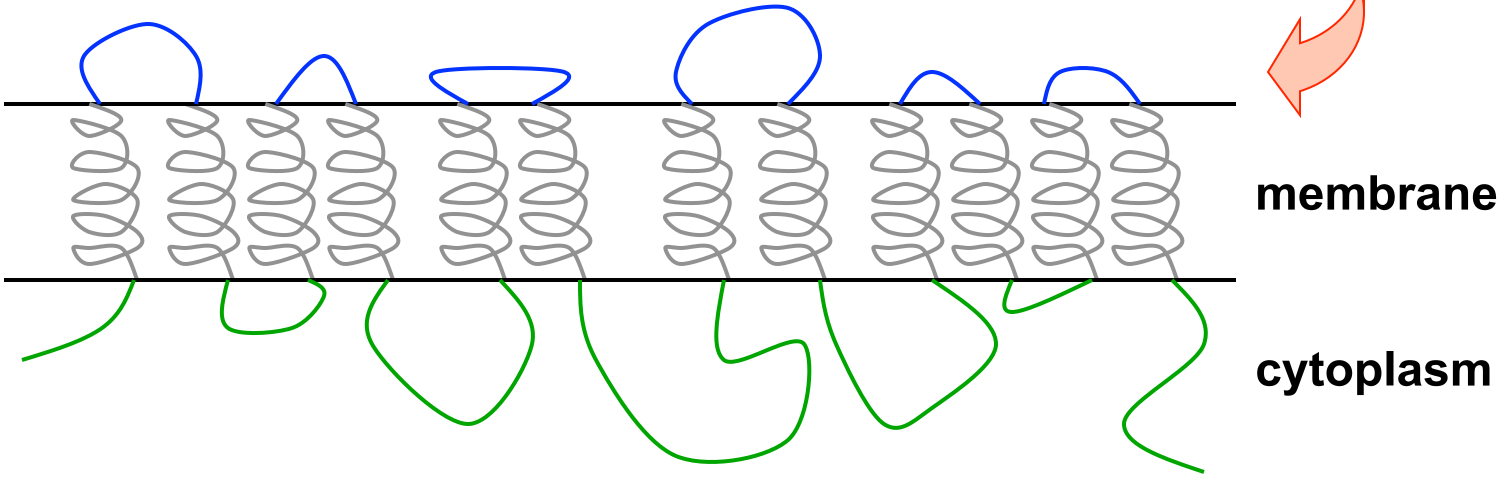
▬ Green lines = cytoplasmic regions

▬ Blue lines = non-cytoplasmic regions

▬ Red lines = signal peptides



By analyzing the probabilities shown on the plot, you can determine where segments within the protein are located.


membrane
cytoplasm