Cellular localization module-guide

Introduction

Protein localization or protein targeting refers to where a protein goes after it is made, which can include specific organelles, the cytoplasm or be secreted outside of the cell. Protein localization is crucial for the proper functioning of proteins, as it allows them to execute their tasks in the appropriate cellular context. Mislocalized proteins can lead to cellular dysfunction. Proteins can have more than one localization, in fact, many proteins change their subcellular location in response to specific cellular signals, environmental conditions, or developmental stages. For example, some proteins that typically localize to the cytoplasm may localize to the nucleus under stress conditions.

Studying protein localization can offer insights into a protein's potential function. By identifying the subcellular compartments where a protein is found, researchers can infer the processes in which it is likely involved, based on the known functions of other proteins within that compartment.

Localization can be determined using experimental techniques such as fluorescence microscopy, as well as computational methods to predict potential localization. In this module we will explore various computational methods to predict potential localization of your proto-gene.

This guide will walk through investigating the localization of an example proto-gene. Follow along and answer the questions in the accompanying worksheet for your proto-gene.

Goals: In this module, participants will learn:

1) Common online informatics tools for predicting protein location in the cell
2) Limitations in localization prediction

Objectives: After completing this module, participants will be able to:

1) Use online tools to predict protein localization and assess the results
2) Use TMHMM to determine if a protein has a predicted transmembrane spanning region.

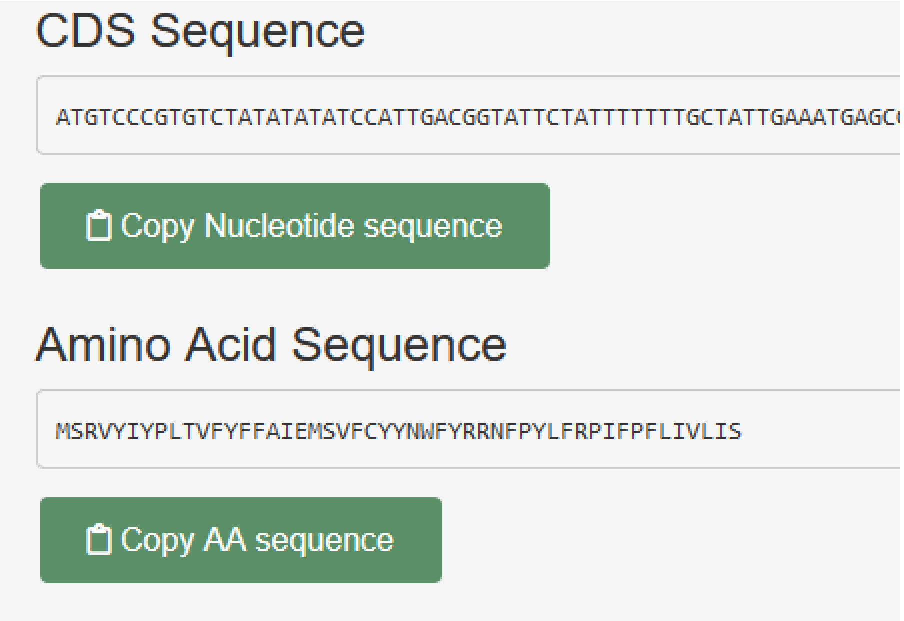
Activity

DeepLoc 2.0

DeepLoc is a type of machine learning algorithm, called a transformer. DeepLoc takes in an amino acid sequence and predicts the subcellular localization of the protein. It will classify your protein sequence as localizing to one (or more) of the following 10 locations: Nucleus, Cytoplasm, Extracellular, Mitochondrion, Cell membrane, Endoplasmic reticulum, Chloroplast, Golgi apparatus, Lysosome/Vacuole and Peroxisome.

1. Get the amino acid sequence for your proto-gene: Go to the ORF information app website: https://carvunislab.csb.pitt.edu/shiny/coexpression/

Type the ORF ID for your proto-gene in the box (remember to precede the number with ‘orf’), select ‘*Sequence*’ in the ‘*Results type*’ drop down menu. Click on the ‘*Copy AA sequence*’ button.



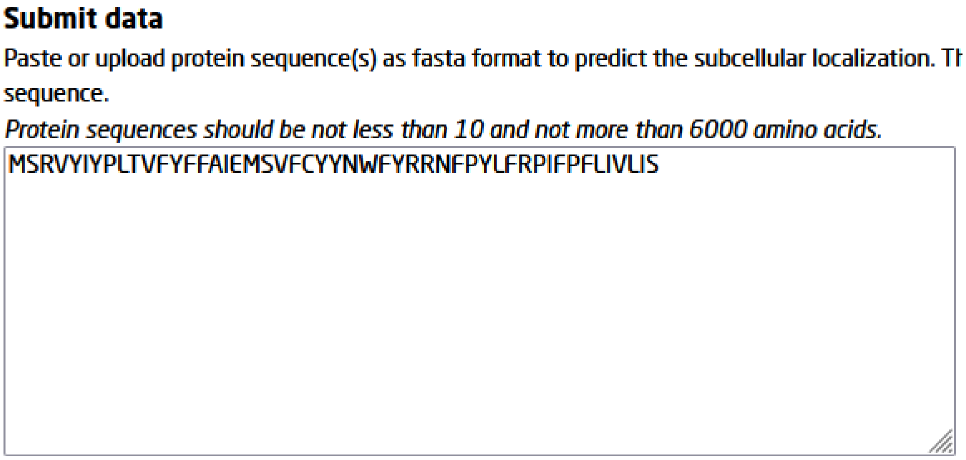
CDS Sequence

```
ATGTCCCGTGTCTATATATATCCATTGACGGTATTCTATTTTTTTGCTATTGAAATGAGC
```

Amino Acid Sequence

```
MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS
```

2. Go to DeepLoc 2.0’s website: https://services.healthtech.dtu.dk/services/DeepLoc-2.0/
3. Paste the amino acid sequence of your proto-gene in the box and click the ‘*Submit*’ button. The results should be ready in ~3-5 minutes*.



Submit data

Paste or upload protein sequence(s) as fasta format to predict the subcellular localization. Th
sequence.

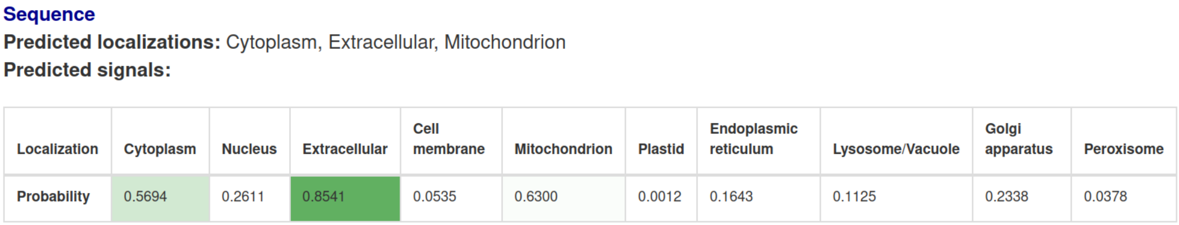
Protein sequences should be not less than 10 and not more than 6000 amino acids.

```
MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS
```

*While you wait for the results from DeepLoc, you can open a new web browser tab and begin on the next section in this module, predicting transmembrane domains with TMHMM.

4. When the results are ready you will see a table that contains the probability of your proto-gene localizing to that cellular location. The values range from 0-1, where 0 means

low probability and 1 means high probability that your proto-gene localizes there. Above the table the predicted localizations are listed.

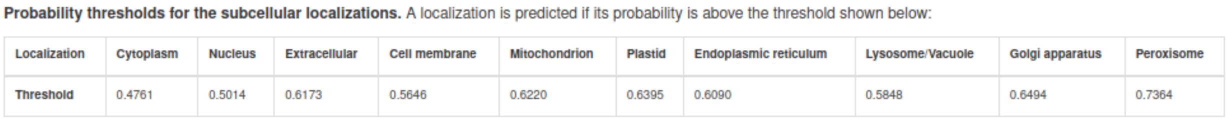


Sequence
Predicted localizations: Cytoplasm, Extracellular, Mitochondrion
Predicted signals:

Localization	Cytoplasm	Nucleus	Extracellular	Cell membrane	Mitochondrion	Plastid	Endoplasmic reticulum	Lysosome/Vacuole	Golgi apparatus	Peroxisome
Probability	0.5694	0.2611	0.8541	0.0535	0.6300	0.0012	0.1643	0.1125	0.2338	0.0378

Because some localizations are more common than others, the algorithm accounts for this by requiring a different threshold for each location prior to calling something localized there, i.e. for a protein to be predicted as going to the nucleus, the probability needs to be above 0.5014. Therefore the location with the highest probability is not necessarily the one the algorithm predicts as the localization, instead it depends on whether the probability is above the threshold for that compartment. Localizations with a probability above the threshold are shaded in green and the intensity of the green color increases with higher probability.

You can view these cutoffs by clicking the red 'Probability thresholds' button at the top of the page.



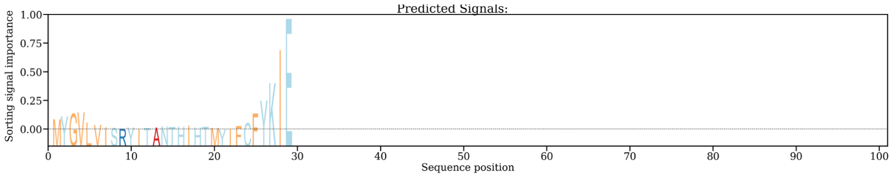
Probability thresholds for the subcellular localizations. A localization is predicted if its probability is above the threshold shown below:

Localization	Cytoplasm	Nucleus	Extracellular	Cell membrane	Mitochondrion	Plastid	Endoplasmic reticulum	Lysosome/Vacuole	Golgi apparatus	Peroxisome
Threshold	0.4761	0.5014	0.6173	0.5646	0.6220	0.6395	0.6090	0.5848	0.6494	0.7364

Also be careful, if none of the probabilities are above the threshold the algorithm will output the compartment with the highest probability as the predicted localization but really
no prediction is made.

Answer question 1 in the worksheet: Where does DeepLoc predict your proto-gene will localize? What is the probability of your proto-gene localizing to that localization? What is the probability threshold for that compartment?

5. The graph at the bottom of the results page shows which part of the protein sequence the algorithm focused on to make its prediction, where the x axis represents the position along the protein sequence and the y axis indicates how important that region was for the prediction.



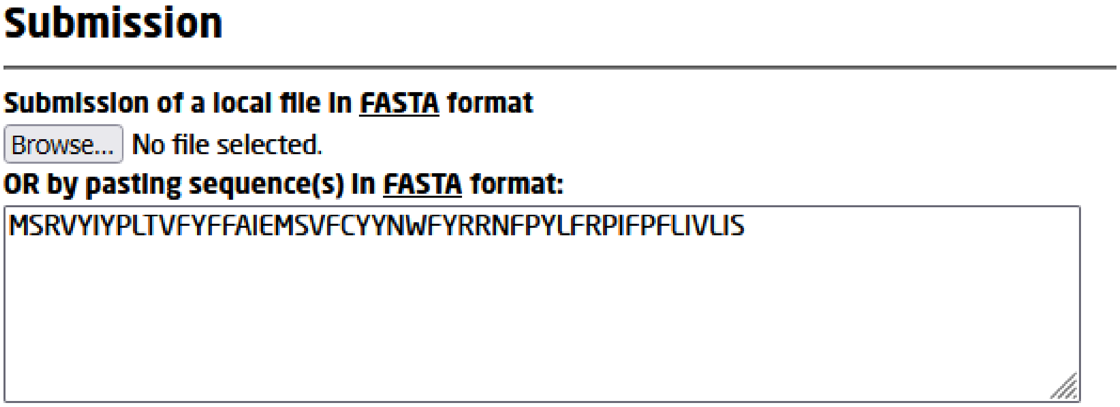
Answer question 2 in the worksheet: Was there a certain region (ie beginning, middle or end) of the sequence that the algorithm paid more attention to?

TMHMM 2.0

TMHMM is a hidden markov model that predicts regions of a protein that are likely to be within a membrane, also called transmembrane helices (TMH).

1. Go to TMHMM's website: https://services.healthtech.dtu.dk/services/TMHMM-2.0/

2. Paste your proto-gene's amino acid sequence into the box and click 'Submit'



Submission

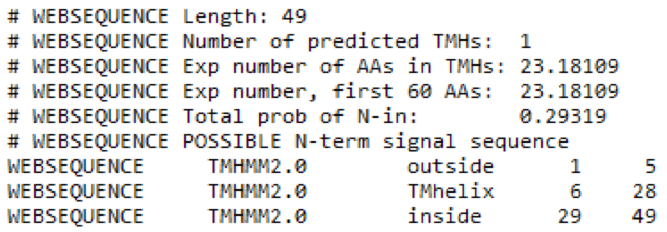
Submission of a local file in FASTA format

OR by pasting sequence(s) in FASTA format:

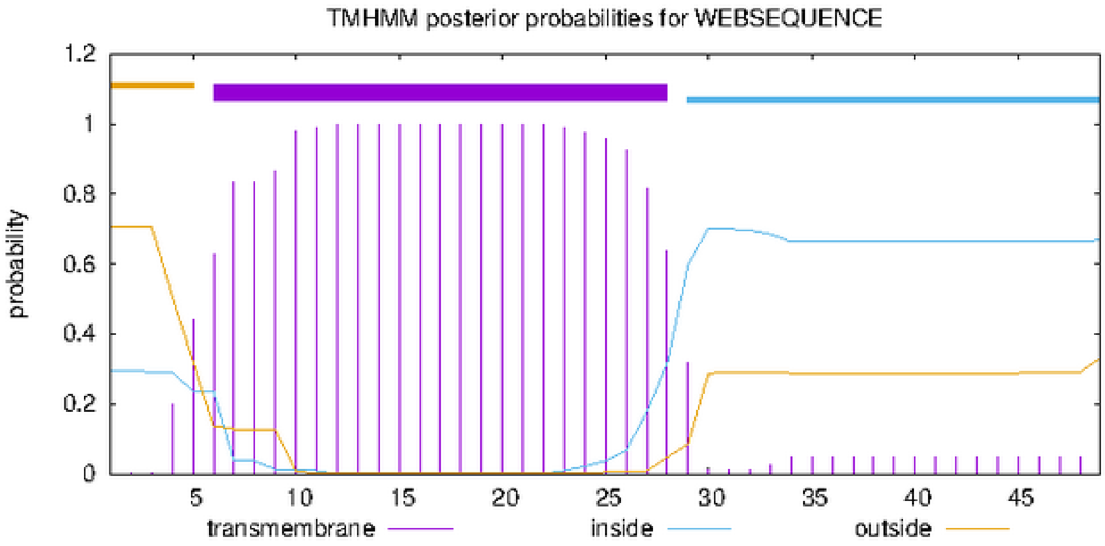
MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS

3. When the results are ready, look to see how many predicted transmembrane helices (TMHs) are in your proto-gene. In this example there is one predicted TMH. The graph below shows which region of the protein contains a TMH (in purple) as well as which regions are predicted to be inside the lumen of the organelle/in the cytoplasm (in blue) and which regions of the protein are predicted to be outside of the organelle/outside of the cell (in yellow). The x-axis of the graph is the amino acid indexes (ie 1 through the length of your protein sequence). The y-axis of the graph is the probability (ranging from 0-1) for each amino acid being either inside, transmembrane or outside.

   For this example, TMHMM predicts that the N' terminus of the proto-gene is located outside, then a transmembrane helix spans about a third of the proto-gene and then the C terminus is located inside the organelle/cell.



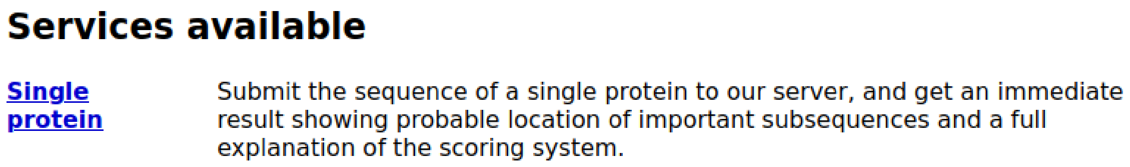
```
# WEBSEQUENCE Length: 49
# WEBSEQUENCE Number of predicted TMHs:  1
# WEBSEQUENCE Exp number of AAs in TMHs: 23.18109
# WEBSEQUENCE Exp number, first 60 AAs:  23.18109
# WEBSEQUENCE Total prob of N-in:        0.29319
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        outside      1     5
WEBSEQUENCE     TMHMM2.0        TMhelix      6    28
WEBSEQUENCE     TMHMM2.0        inside      29    49
```



Answer question 3 in the worksheet: How many transmembrane helices (TMH) does TMHMM predict your proto-gene has? What is meant by “inside” and “outside”? Include a copy of the TMHMM graph.

NucPred

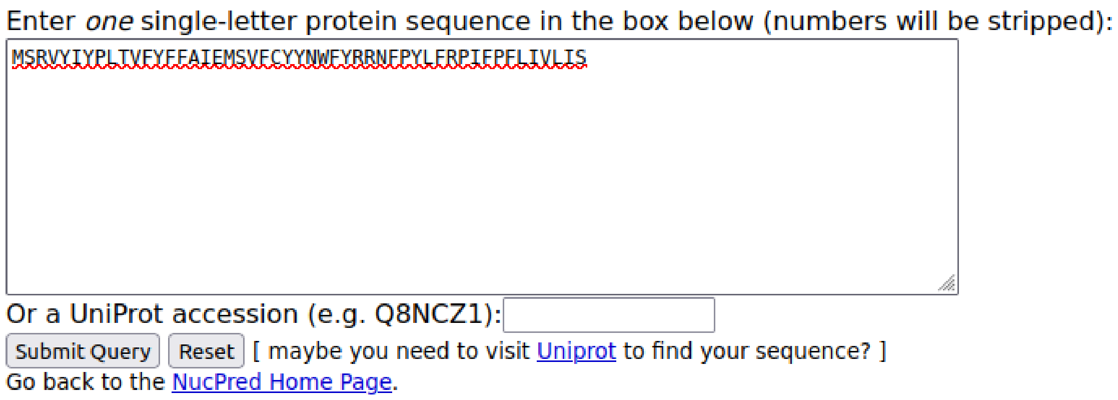
NucPred is an algorithm that predicts the likelihood that your protein localizes to the nucleus.

1. Go to NucPred’s website: https://nucpred.bioinfo.se/nucpred/
2. Click the ‘Single protein’ link to submit a single protein sequence for prediction.
3. 

Services available

Single protein	Submit the sequence of a single protein to our server, and get an immediate result showing probable location of important subsequences and a full explanation of the scoring system.

4. Paste your amino acid sequence into the text box and click ‘Submit Query’



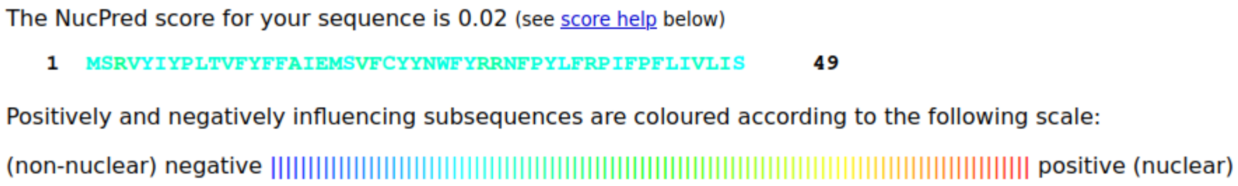
Enter *one* single-letter protein sequence in the box below (numbers will be stripped):

MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS

Or a UniProt accession (e.g. Q8NCZ1):

[maybe you need to visit Uniprot to find your sequence?]
Go back to the NucPred Home Page.

5. When the results have loaded look at the prediction. The algorithm predicts an overall score between 0 and 1 for how likely your protein is to localize to the nucleus, a value closer to 1 means a higher probability and a value towards 0 means a lower probability your protein goes to the nucleus. In this example the NucPred score is 0.02 which is very low. Also the algorithm colors each amino acid to indicate how much that region of the protein contributes to its predicted localization, where yellow and red mean higher confidence regions. In this example, all parts of the proto-gene are unlikely to influence the protein to be nuclear located. For more information about scoring, click the ‘score help’ link.



The NucPred score for your sequence is 0.02 (see score help below)

```
1   MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS      49
```

Positively and negatively influencing subsequences are coloured according to the following scale:

(non-nuclear) negative ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| positive (nuclear)

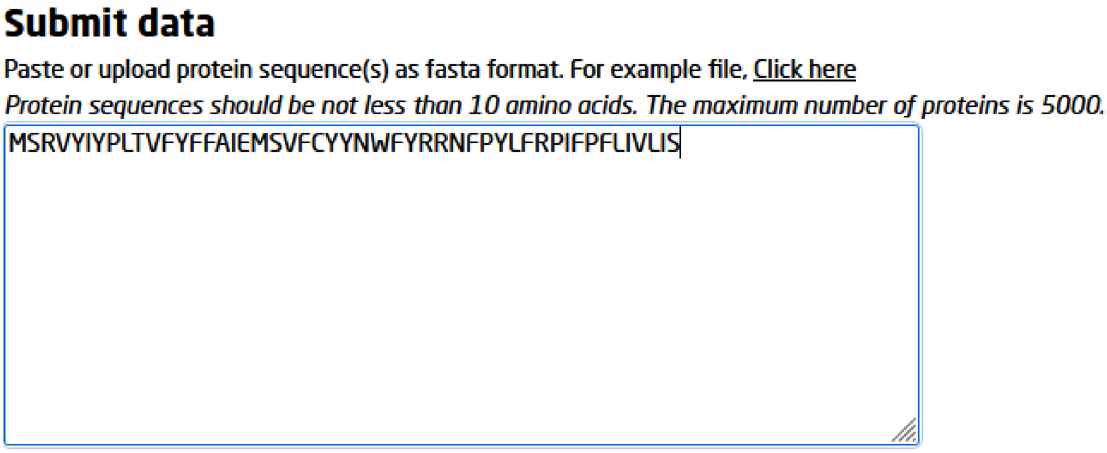
Answer question 4 in the worksheet: What was your proto-gene's NucPred score? From this score- do you think your proto-gene localizes to the nucleus? Does this result fit in the context of the DeepLoc result?

Answer question 5 in the worksheet: Look at the color-coded sequence in the NucPred output. Was there any region/motif in your proto-gene that influenced NucPred’s prediction?

TargetP

TargetP is an algorithm that detects if protein sequences have either mitochondrial targeting peptide sequences (mTP) or secretory pathway signal peptides (SP). Mitochondrial peptide sequences (mTPs) are sequence motifs used to localize proteins to the mitochondria and secretory pathway signal peptides (SP) are sequence motifs that direct proteins to the ER to then enter the secretory pathway to be secreted outside of the cell.

1. Go to TargetP's website: https://services.healthtech.dtu.dk/services/TargetP-2.0/
2. Paste your proto-gene's amino acid sequence into the box and click 'Submit'



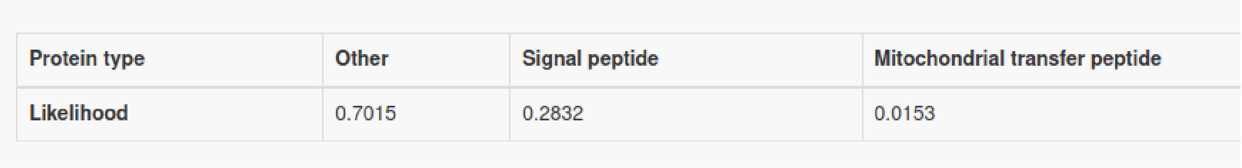
Submit data

Paste or upload protein sequence(s) as fasta format. For example file, Click here

Protein sequences should be not less than 10 amino acids. The maximum number of proteins is 5000.

```
MSRVYIYPLTVFYFFAIEMSVFCYYNWFYRRNFPYLFRPIFPFLIVLIS
```

3. Look at the results table, which contains the likelihood of your proto-gene containing a signal peptide or mitochondrial peptide sequence. The likelihood values range from 0 to 1, where a value of 0 means low likelihood and a value of 1 means very high likelihood. For this example, this proto-gene is unlikely to have a signal peptide sequence and is therefore unlikely to be secreted outside of the cell. This proto-gene is also unlikely to have a mitochondrial transfer peptide and is therefore unlikely to localize to the mitochondria. Other just means the algorithm predicts there are no signal or mitochondrial peptides present.



Protein type	Other	Signal peptide	Mitochondrial transfer peptide
Likelihood	0.7015	0.2832	0.0153

Answer question 6 in the worksheet: What is the likelihood of your proto-gene having a signal peptide or mitochondrial transfer peptide? Does this result align with what DeepLoc predicted?

Answer question 7 in the worksheet: Considering all the information from the various algorithms, where do you hypothesize your proto-gene localizes?

Sources

- José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, Ole Winther. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btx431
- Markus Brameier, Andrea Krings, Robert M. MacCallum. (2007). NucPred—Predicting nuclear localization of proteins. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btm066
- Anders Krogh, Björn Larsson, Gunnar von Heijne, Erik L.L Sonnhammer. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes, *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.2000.4315
- José Juan Almagro Armenteros, Marco Salvatore, Ole Winther, Olof Emanuelsson, Gunnar von Heijne, Arne Elofsson, and Henrik Nielsen. (2019). Detecting sequence signals in targeting peptides using deep learning, *Life Science Alliance*, https://doi.org/10.26508/lsa.201900429