Module 4: Cellular Localization Data

Objective

The objectives of this module are:

1. To determine if the protein you are annotating is located in the cytoplasm of the cell, embedded in the cytoplasmic membrane or secreted using TMHMM, SignalP, PSORTb and Phobius applications.
2. To document your search results in the Cellular Localization Module lab notebook.

Materials

To perform this activity you will need:

- Access to the internet on a computer equipped with the most recent version of Firefox (preferred), Chrome or Safari.
- To have completed the sign up for GENI-ACT described in the Signing Up for GENI-ACT section of the manual.

Background

The purpose of this module is to determine where in the bacterial cell the protein encoded by the gene you are investigating is likely to reside. You will be using multiple tools, some of which have similar functions, to analyze your protein. In the best case scenario all of the tools will agree and you will have a very straightforward result. However, on occasion some of the tools will give results that conflict with one another. In those cases you will need to use your best judgment to determine the likely location of your protein within *Kytococcus*, or you will need to conclude that there is not enough data to support a specific location.

The tools you will be using include TMHMM (to predict whether or not you have any transmembrane helixes in your protein), SignalP (to determine if you have a signal peptide in your protein), PSORTb (uses a combination of looking for transmembrane helixes and signal peptide sequences to assign the cellular localization of the protein) and Phobius (combines the output from TMHMM and SignalP in one graphical output). In order to understand how these tools work, the meaning of transmembrane helixes and signal peptides will be discussed.

The structure of a cell membrane is referred to as a bilayer and is shown in figure 4.1. Phospholipids are oriented with their hydrophilic ("water loving") components facing either the exterior of the cell or the cytoplasm, both of which are aqueous (water solvent based) compartments. The hydrophilic domains of the phospholipids are the clear circles at the top and bottom of figure. The colored lines projecting toward the interior of the bilayer are the lipid ("fat") components of the bilayer. This creates a hydrophobic ("water

fearing") domain. The lipid bilayer creates a barrier to generally prevent movement of substances from inside the cell to outside or from outside the cell to inside. Proteins in particular must have a mechanism to be moved from the cytoplasm inside the cell to the exterior if the protein is to be secreted, or to be anchored within the lipid bilayer if the protein is to reside in the membrane. Proteins that reside in the membrane may be, for example, transporters, which are involved in the specific transport of molecules from outside of the cell to in or inside of the cell to out. Proteins that reside in the membrane will characteristically have a series of approximately 20 hydrophobic amino acid residues that reside in the hydrophobic compartment of the lipid bilayer forming a transmembrane helix (Figure 4.2.1). Some proteins will pass back and forth through the membrane multiple times, having multiple hydrophobic transmembrane helixes separated by regions of hydrophilic amino acids that project to the exterior or interior of the cell (Figure 4.2). Proteins embedded in the bilayer are referred to as integral membrane proteins.

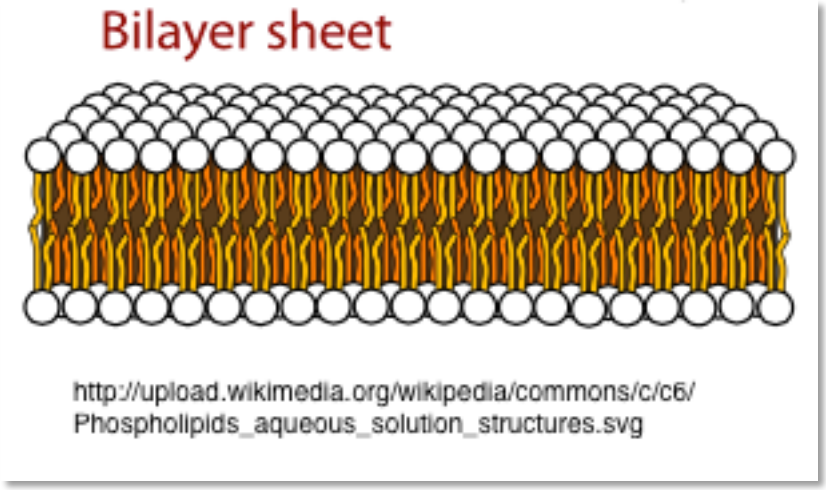


Figure 4.1. The lipid bilayer

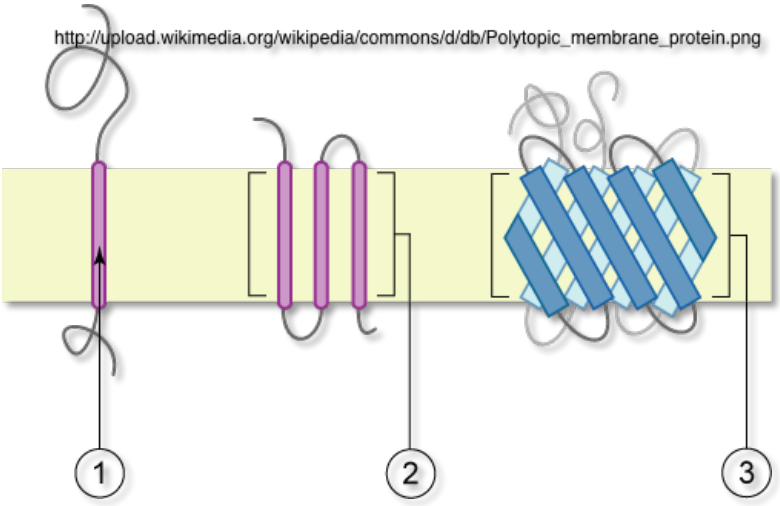


Figure 4.2. Examples of transmembrane proteins. Helixes discussed in this module are shown in 4.2.1 and 4.2.2.

A signal peptide is often present at the amino terminal end of a protein if that protein is meant to be secreted from the cell. This peptide will target the protein, either directly or by way of an intermediate, to a protein complex that facilitates the passage of the protein through a channel created in the lipid bilayer (Figure 4.3). As the protein crosses the membrane, the signal peptide is cleaved off by a specific type of protease (protein

cutting enzyme) know as a signal peptidase that recognizes a characteristic sequence of amino acids to cut. The signal peptide is thus it is not a part of the functional secreted protein.

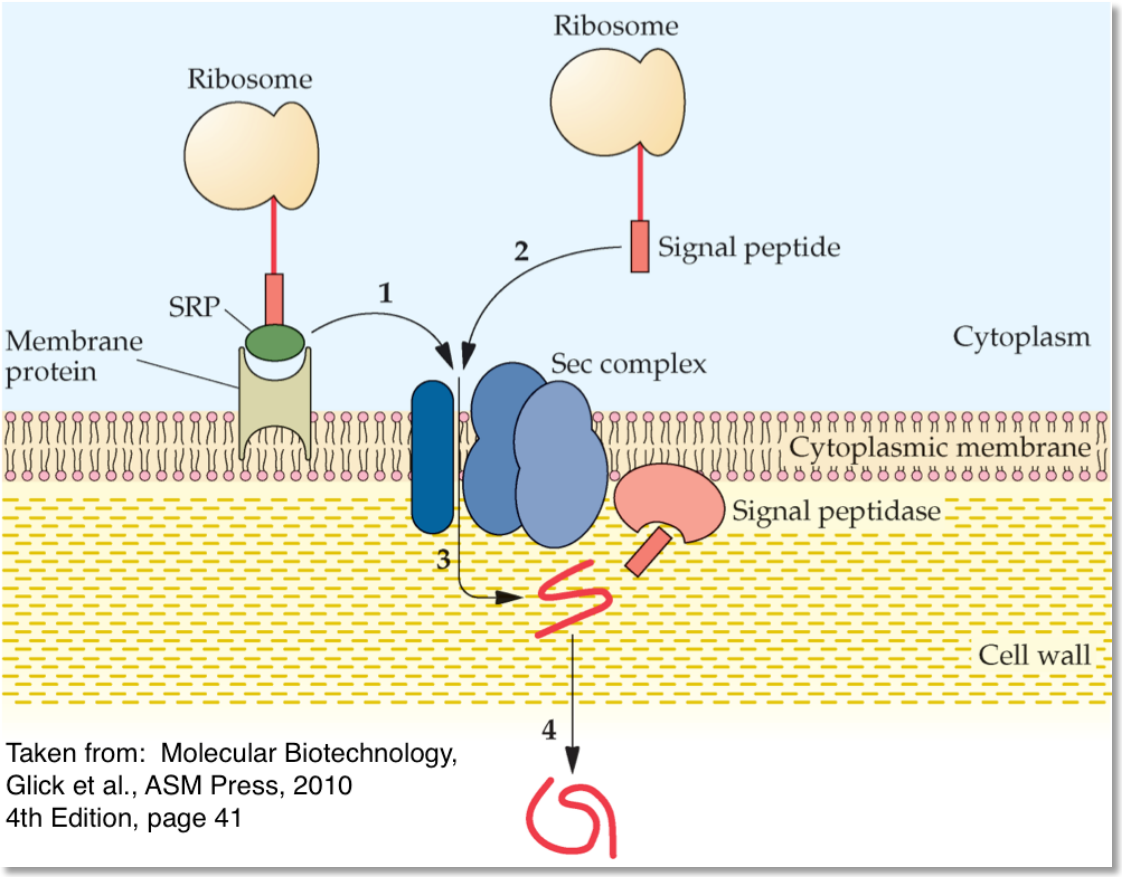


Figure 4.3. Secretion of proteins in Gram + bacteria

The tools in this module take advantage of the knowledge of transmembrane helixes, signal peptides and signal peptidase recognition sites to determine if the protein being investigated has any of these characteristics. Thus a protein that has what looks to be a signal peptide and a signal peptidase cleavage recognition sequence will likely be a secreted protein. A protein that has one or more transmembrane helix domains will likely be an integral membrane protein. A protein that has no clear indication of transmembrane helixes or a signal peptide will likely reside in the interior of the cell (located in the cytoplasm).

Procedures

PubMed – this tool will demonstrate the use of a free scientific literature database housed at the National Library of Medicine at the National Institutes of Health. You will use it to identify the Gram stain results for *Kytococcus sedentarius*. You will need Gram stain information to enter into some of the other tools used in this module. Bacteria are classified as being Gram positive (+) or Gram negative (-) and the staining procedure is one of the fundamental steps in the characterization of a bacterium. Gram staining will stain

Gram + bacteria blue and Gram – bacteria red. And example of a common Gram + bacterium is *Staphylococcus aureus*, while a common Gram – bacterium is *Escherichia coli* (also called *E.* coli for short). The distinction is important for the tools in this module because of the fact that Gram + bacterial have only a single membrane separating the cytoplasm from the exterior, whereas Gram – bacteria have an inner plasma membrane, an extracellular space called the periplasm and then a second membrane that separates the periplasm from the exterior. Thus when a protein is secreted from a Gram + bacterium it only needs to traverse a single membrane to move from the interior to the exterior. In contrast, a protein secreted by a Gram – bacterium must traverse the inner membrane, the periplasmic space and the outer membrane to reach the exterior of the cell. The mechanism of protein secretion is different for Gram + and Gram – bacteria as a result. The software tools in this module will use the Gram stain information to determine if your protein has characteristics that would allow it to be secreted by a Gram + or Gram – bacterium.

You should feel free to use PubMed to search for information about your protein as you move through the modules, similar to the way you would use Google to search for information in your everyday life. Some of the papers you will find in a PubMed search are freely downloadable for reading, while others will only have the abstract of the paper listed. The abstract of a scientific paper is a concise summary of the information contained in the body of the paper.

1. Click on the following link. http://www.ncbi.nlm.nih.gov/pubmed/.
2. You will be taken to the PubMed start page (Figure 4.4).

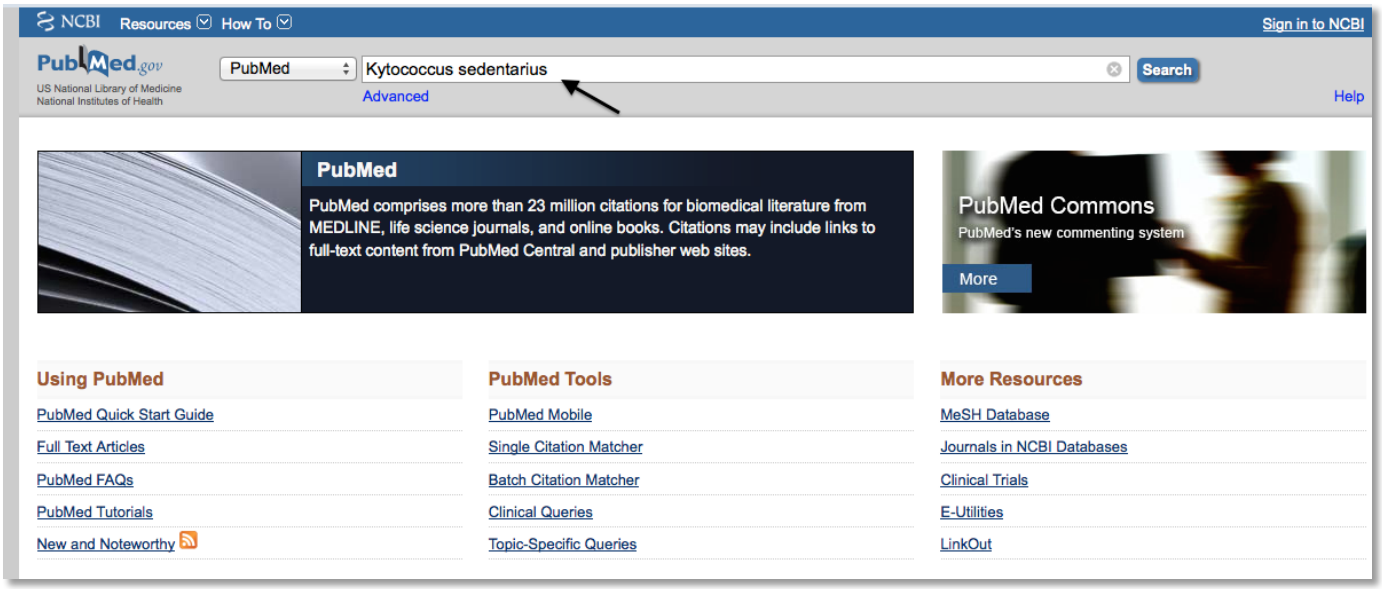


Figure 4.4. The PubMed entry page. Search terms are placed in the box indicated by the

3. Enter your organism name in the search box as indicated by the arrow in figure 4.4 (*Kytococcus sedentarius* is the example used here) and click search.
4. You will see a results window similar to the one shown in figure 4.5.
5. Scroll down the page looking for publications that have titles that seem like they could have the information you are looking for. Note the hyperlink for each article and the fact that some of the articles

will show "free article" or "free PMC article". These articles can be downloaded or read online in their entirety.

6. Scrolling down the results showing in figure 4.5, you will see an article at the bottom that was published to describe the sequencing of the *Kytococcus sedentarius* genome (arrow figure 4.5).

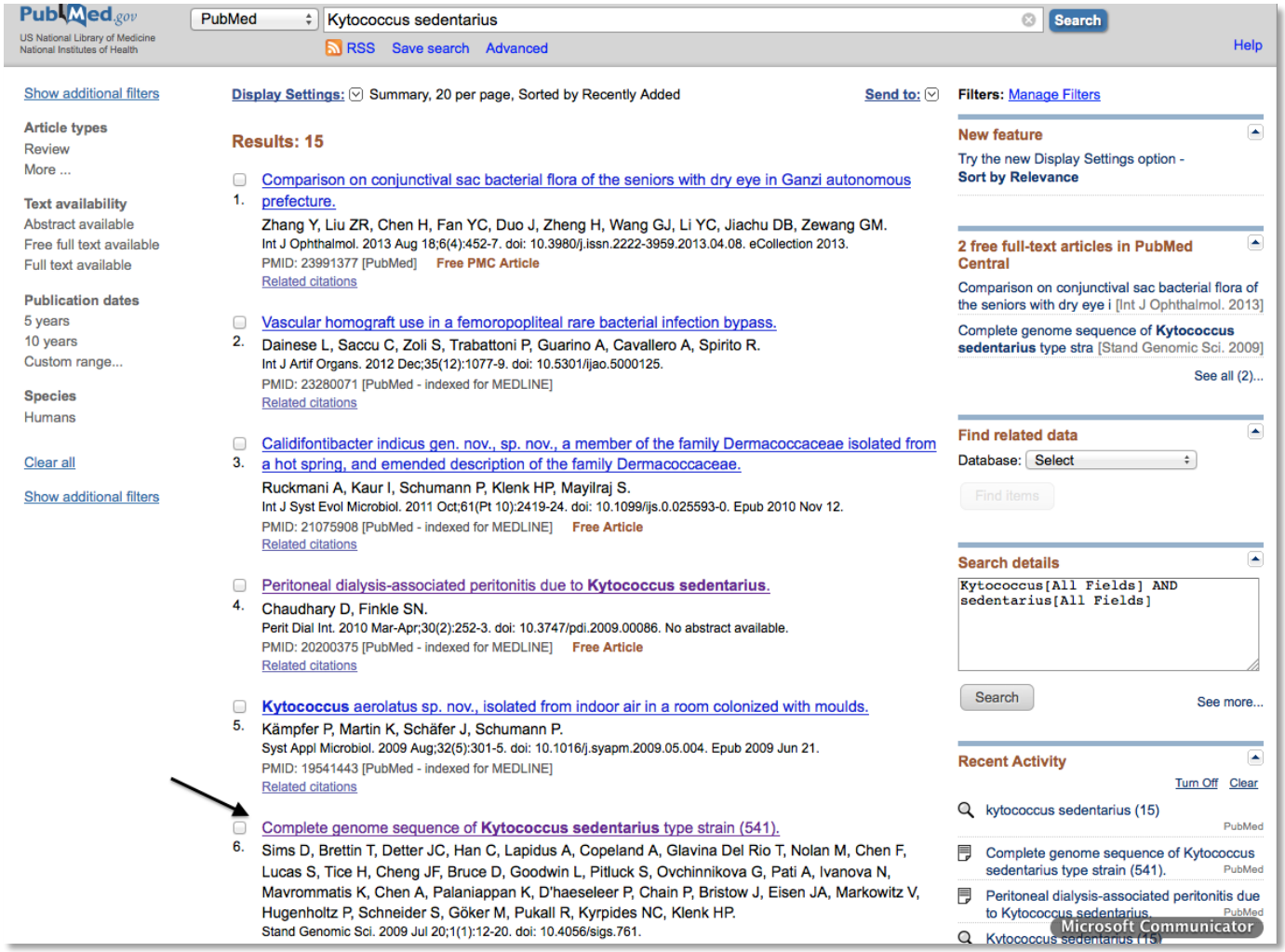


Figure 4.5. The PubMed results for a *Kytococcus* search. The arrow indicates a paper likely to be significant in determining the Gram stain characteristics of *Kytococcus*.

7. Clicking on the hyperlink for that article will take you to the abstract of the article, as well as the opportunity to download or read the full article (Figure 4.6). Within the abstract you will see a statement of the fact that *Kytococcus sedentarius* is a Gram + bacterium (oval, figure 4.6). Follow the same procedure to find a paper describing the Gram stain of your bacterium.

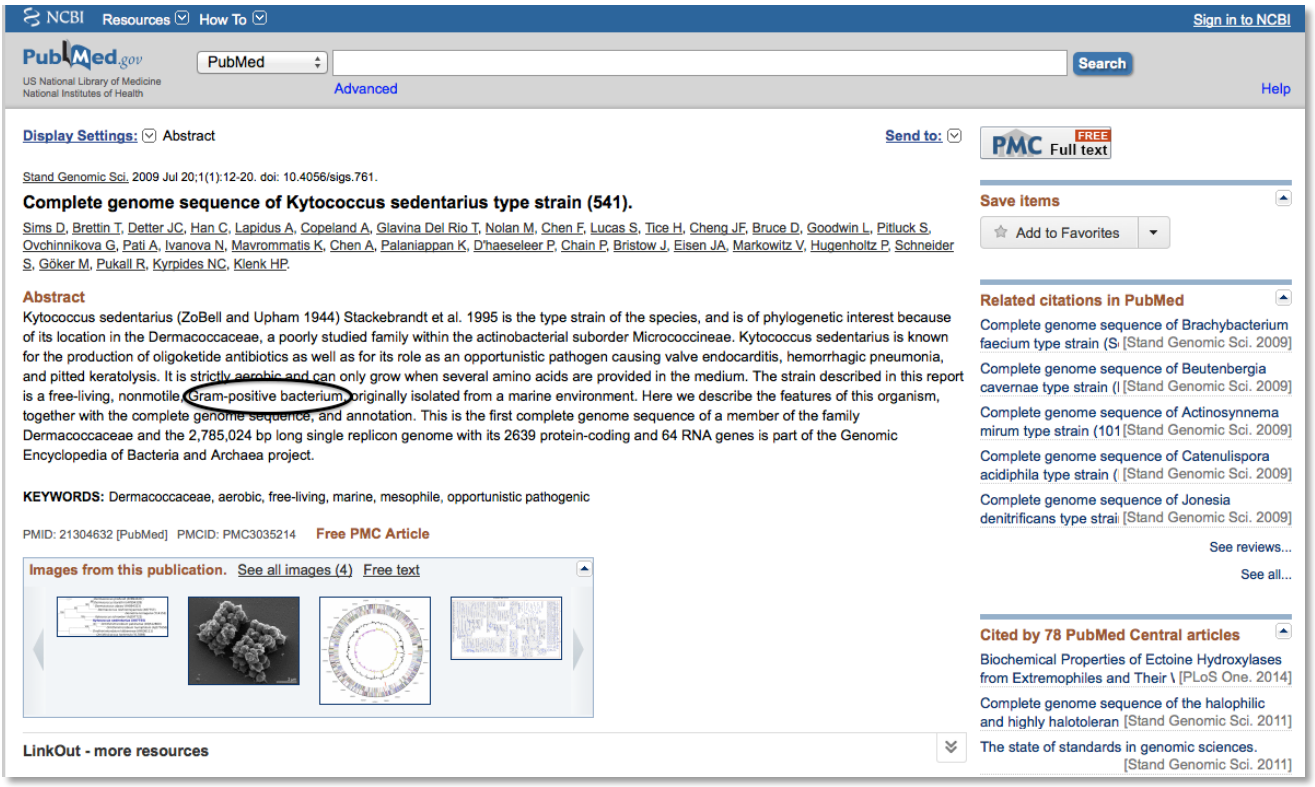


Figure 4.6. The abstract from the article indicated in Figure 4.5. A statement that *Kytococcus* is Gram + is indicated by the oval.

8. Go to the notebook for your gene and record the Gram stain results for your organism (*Kytococcus sedentarius* again used as an example)(Figure 4.7).

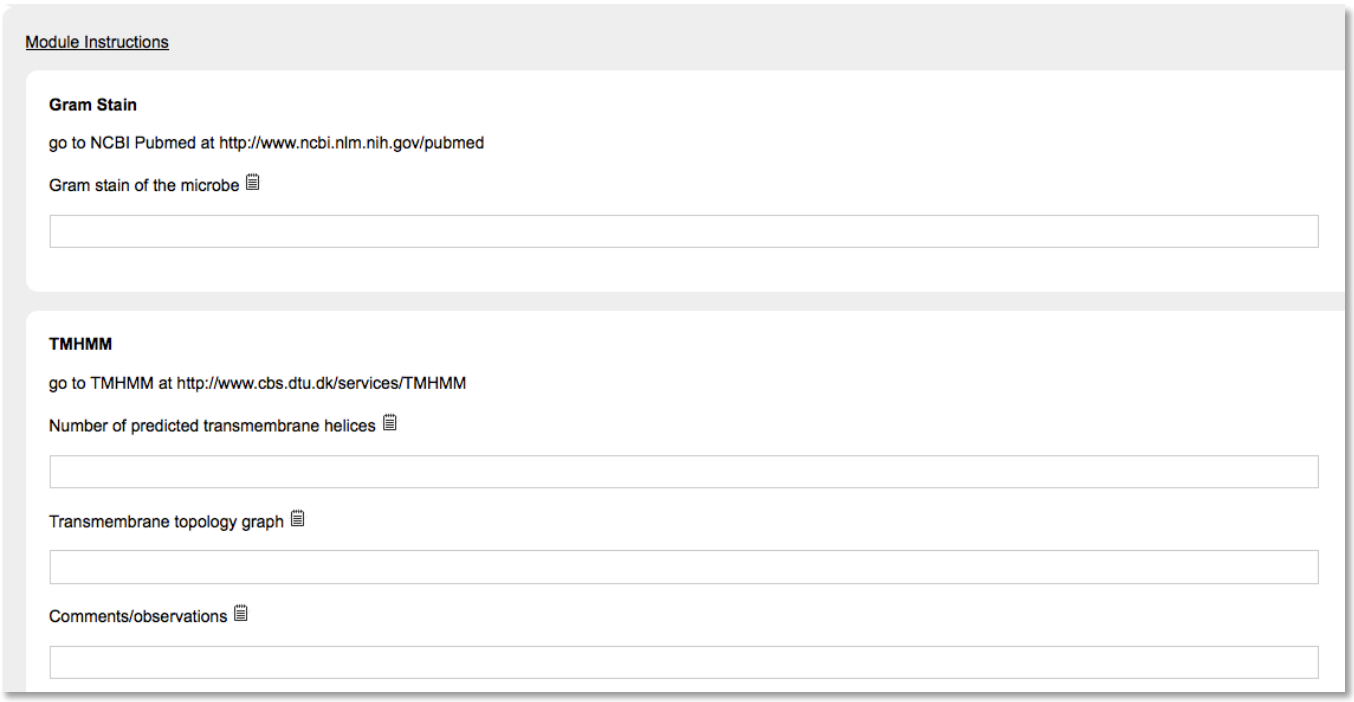


Figure 4.7. The cellular localization module notebook showing Gram stain and TMHMM areas to be filled in.

TMHMM – as noted in the introduction to this module, TMHMM will determine if any transmembrane helix domains exist in your protein. The presence of helixes will suggest your protein is an integral membrane protein. The lack of helixes would indicate that your protein is NOT an integral membrane protein.

1. Go to the TMHMM start page by clicking on the following link: http://www.cbs.dtu.dk/services/TMHMM/ (Figure 4.8).

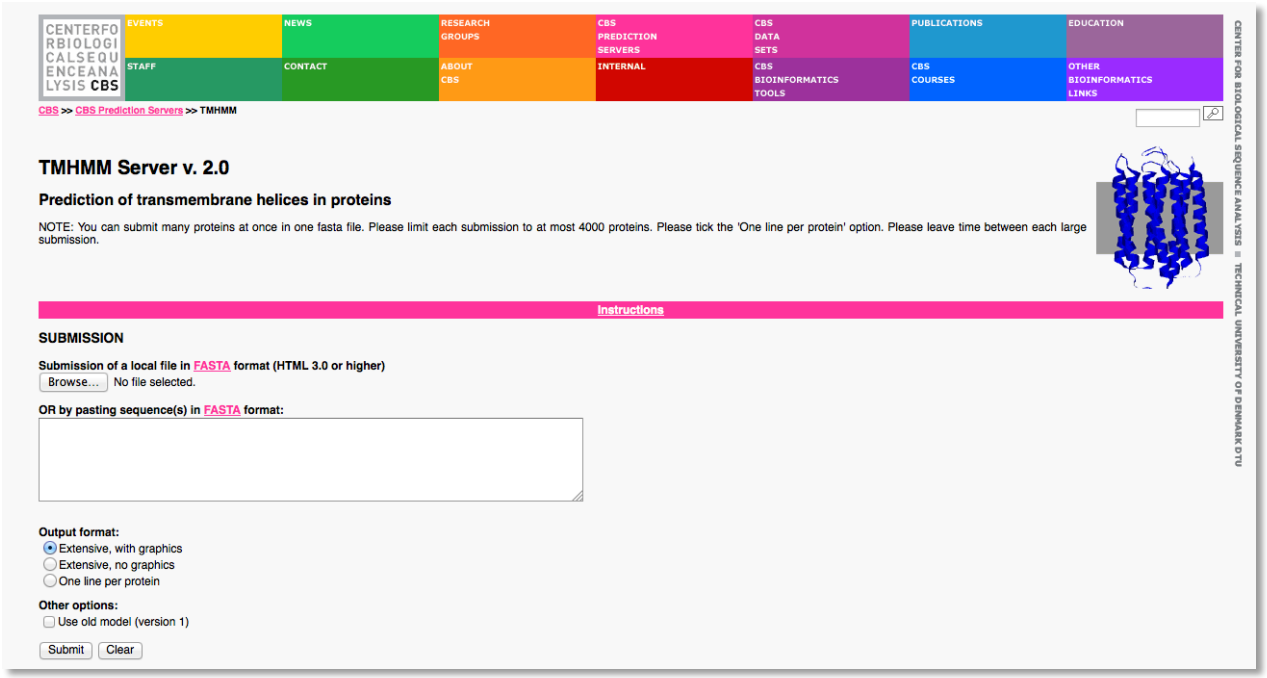


Figure 4.8. The TMHMM start

2. Paste the sequence in FASTA format in the submission box. Make sure the Extensive with graphics button is highlighted below the submission box and click on submit.

3. You will see one of two general forms of results for your protein as indicated in figures 4.9 and 4.10. In the event that your protein does NOT have any predicted transmembrane helixes, the results will look essentially as indicated in figure 4.9. No helixes are indicated in the text output (arrow in figure 4.9) and the graphic shows 3 parallel lines. **Remember, this tool ONLY predicts the presence of transmembrane helixes. It DOES NOT tell you anything else about the location of your protein. Sometimes students tend to interpret negative results such as those shown in Figure 4.10 as predicting the protein is secreted, which is an incorrect interpretation.** If your protein does have a transmembrane helix or multiple helixes, you will see results similar to those shown in figure 4.10. The text output will indicate the number of predicted helixes, and the graphic will show the position(s) of the helix or helixes in your protein (thick red lines), as well as a prediction of what portions of the protein are inside (blue line) vs. outside (red line) of the membrane. Note that the program will sometimes label a signal peptide as a transmembrane helix, which may also adversely affect the prediction of which portions of the protein are inside or outside of the cell.

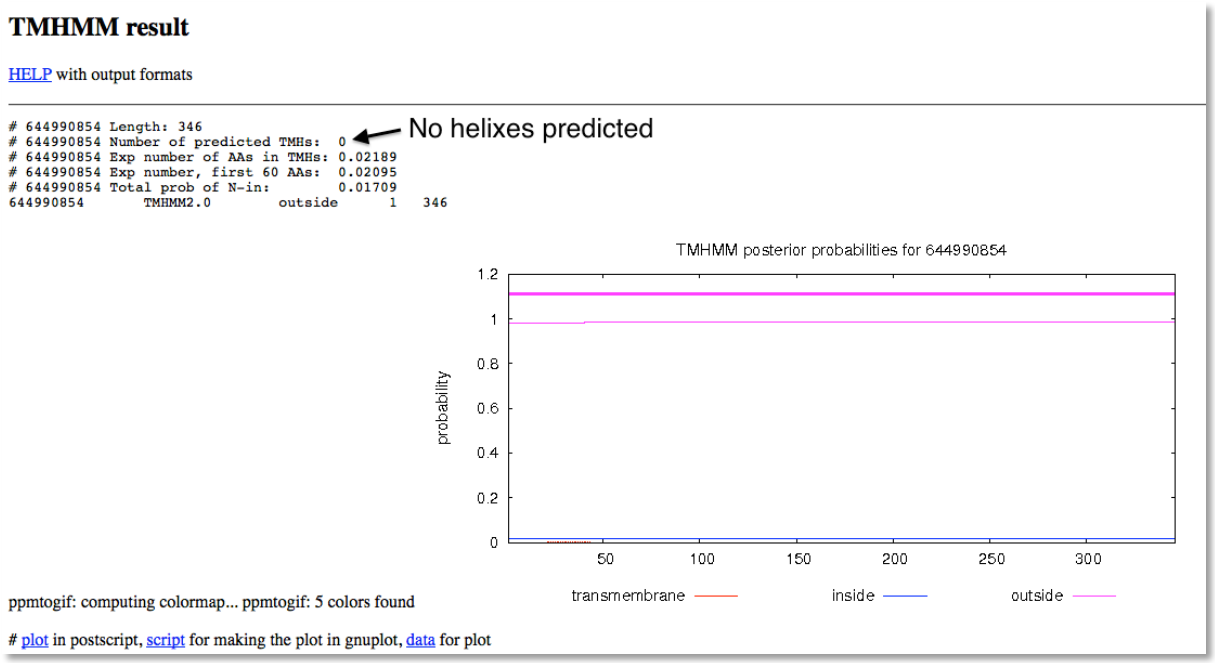


Figure 4.9. A TMHMM results page. No transmembrane helixes are identified.

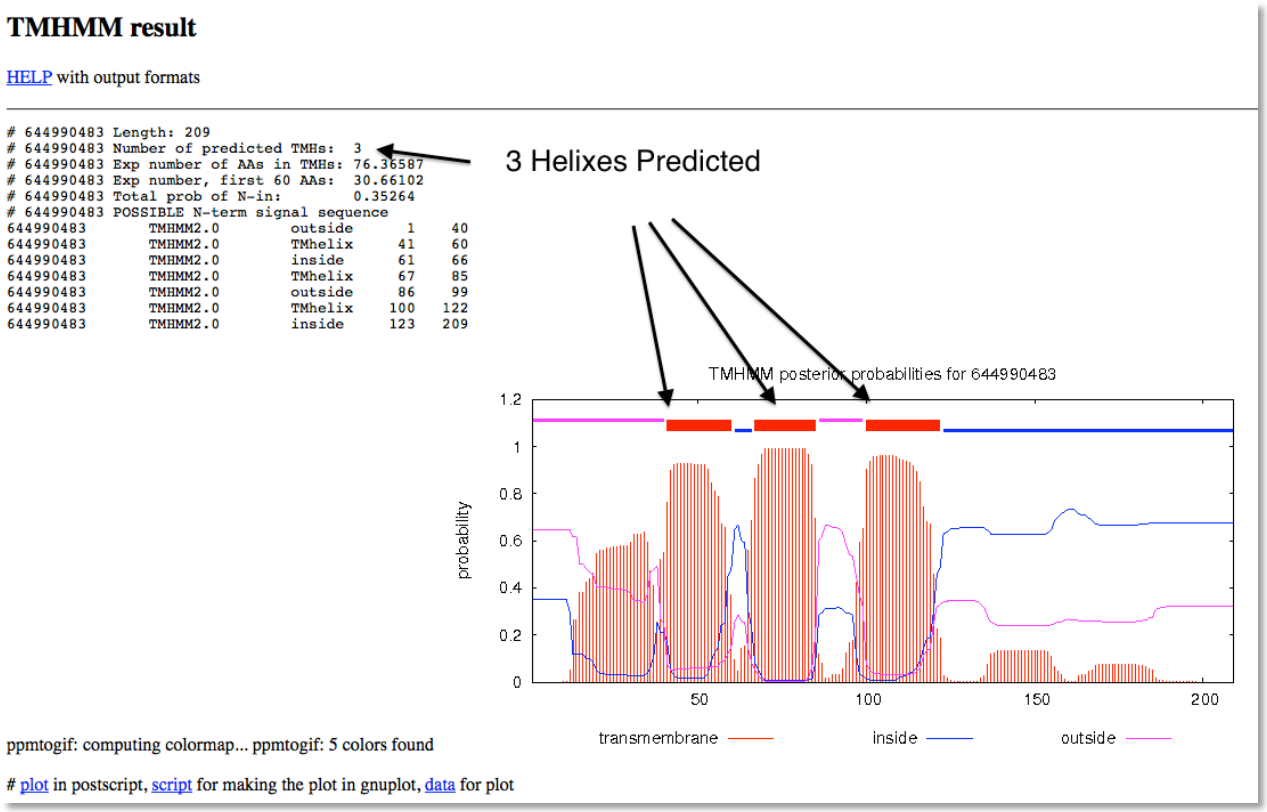


Figure 4.10. A TMHMM result predicting 3 transmembrane helixes.

4. Once the results are obtained and they have been analyzed, Snip or Capture the image and associated text file and save it as a jpg.

5. Enter the image in your lab notebook under the TMHMM topology graph section (Figure 4.7) and enter your interpretation of whether or not transmembrane helixes are present in the comments box.

SignalP - as indicated in the introduction to this module, SIGNALP will analyze your protein sequence the predict whether or not a signal peptide is present. The presence of a signal peptide will suggest that your protein is secreted from your bacterium into the surrounding medium.

6. Click on the following link. http://www.cbs.dtu.dk/services/SignalP/ .

7. This will take you to the SignalP start page (Figure 4.11) where you will past the FASTA formatted amino acid sequence of your protein into the submission box. Be sure to select the appropriate Gram stain option as well (arrow, Figure 4.11). And then click submit.

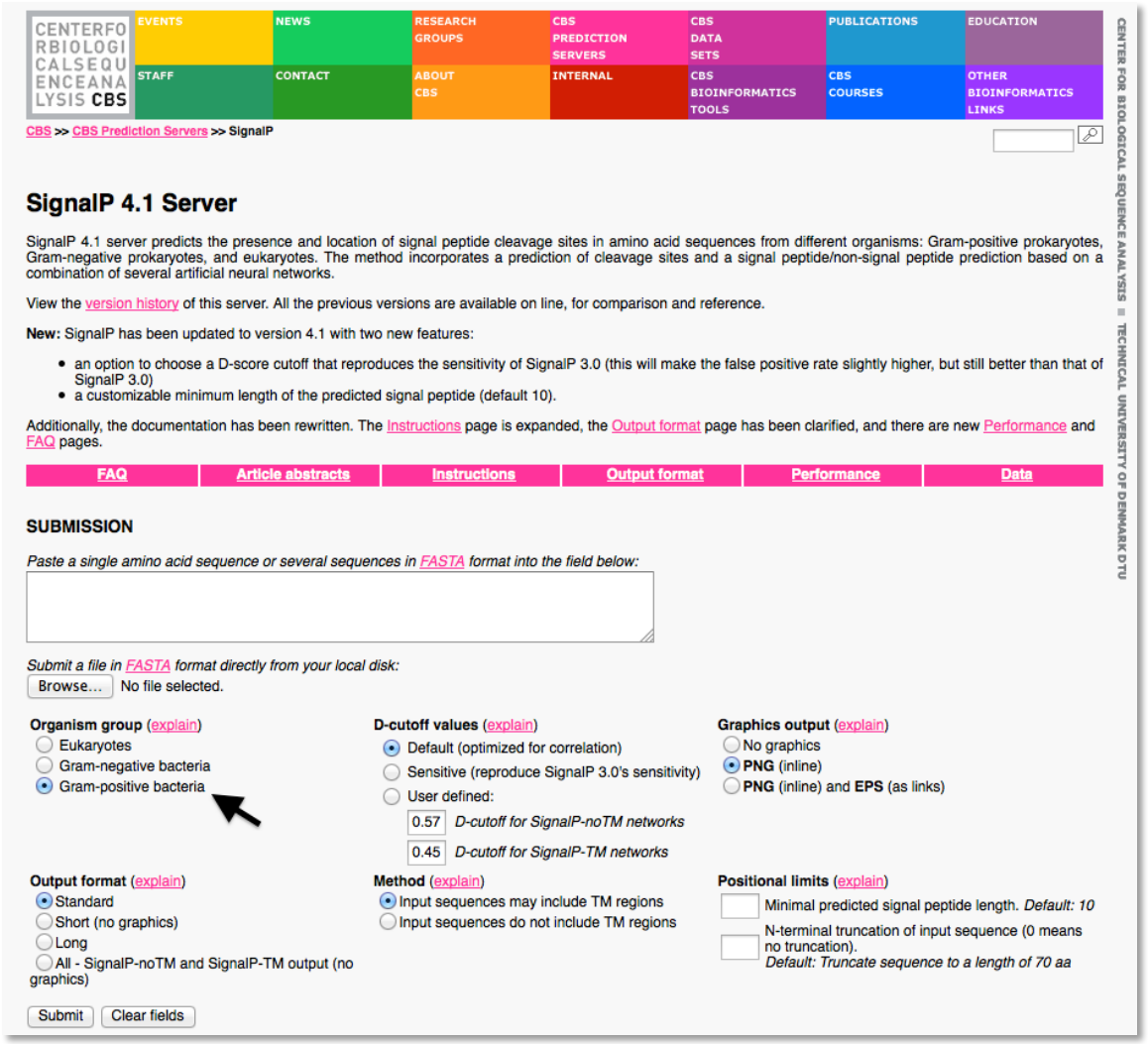


Figure 4.11. The SignalP start page.

8. You will see a graphical output for your results as well as a table of score values. The scores and their meaning are described below (taken directly from the SignalP output description page at http://www.cbs.dtu.dk/services/SignalP/output.php).

9. DESCRIPTION OF THE SCORES

   A. The neural networks in SignalP produce three output scores for each position in the input sequence:

   1. **C-score** (raw cleavage site score)
      1. The output from the CS networks, which are trained to distinguish signal peptide cleavage sites from everything else.
      2. Note the position numbering of the cleavage site: the C-score is trained to be high at the position immediately after the cleavage site (the first residue in the mature protein).
   2. **S-score** (signal peptide score)
      1. The output from the SP networks, which are trained to distinguish positions within signal peptides from positions in the mature part of the proteins and from proteins without signal peptides.
   3. **Y-score** (combined cleavage site score) A combination (geometric average) of the C-score and the slope of the S-score, resulting in a better cleavage site prediction than the raw C-score alone. This is due to the fact that multiple high-peaking C-scores can be found in one sequence, where only one is the true cleavage site. The Y-score distinguishes between C-score peaks by choosing the one where the slope of the S-score is steep.

10. The graphical output from SignalP shows the three different scores, C, S and Y, for each amino acid position in the sequence. Note that the entire amino acid sequence for your protein will not be shown in the graphic. As a signal peptides reside at the amino terminus of proteins, only the most amino terminal amino acids in your sequence will be shown.

11. In the summary below the plot, the maximal values of the three scores are reported. In addition, the following two scores are shown:

    mean S- The average S-score of the possible signal peptide (from position 1 to the position immediately before the maximal Y-score).

    D-score (discrimination score)- A weighted average of the mean S and the max. Y scores. This is the score that is used to discriminate signal peptides from non-signal peptides, and must reach a threshold value of at least 0.45 to be significant.

    There will also be a statement (Yes or No) next to the D score as to whether there is a signal peptide present. If one is present there will also be text describing between which amino acid residues the cleavage by the signal peptidase is predicted to occur.

12. For non-secretory proteins all the scores represented in the SignalP output should ideally be very low (close to the negative target value of 0.1)

13. In practice, a positive result for a signal peptide will show the S-score high at the amino terminal portion of your protein and the Y and C scores low in the same region. You will then see the S score drop abruptly (indicating the amino acid position at which the signal peptide domain no longer exists). You

will then see an abrupt rise in the Y and C scores at the nearly the same amino acid residues that you see the sharp drop in the S score.

14. Figure 4.12 shows an example output from a protein where a signal peptide **IS NOT** predicted to exist, and Figure 4.13 shows an example output from a protein where a signal peptide **IS** predicted to exist.

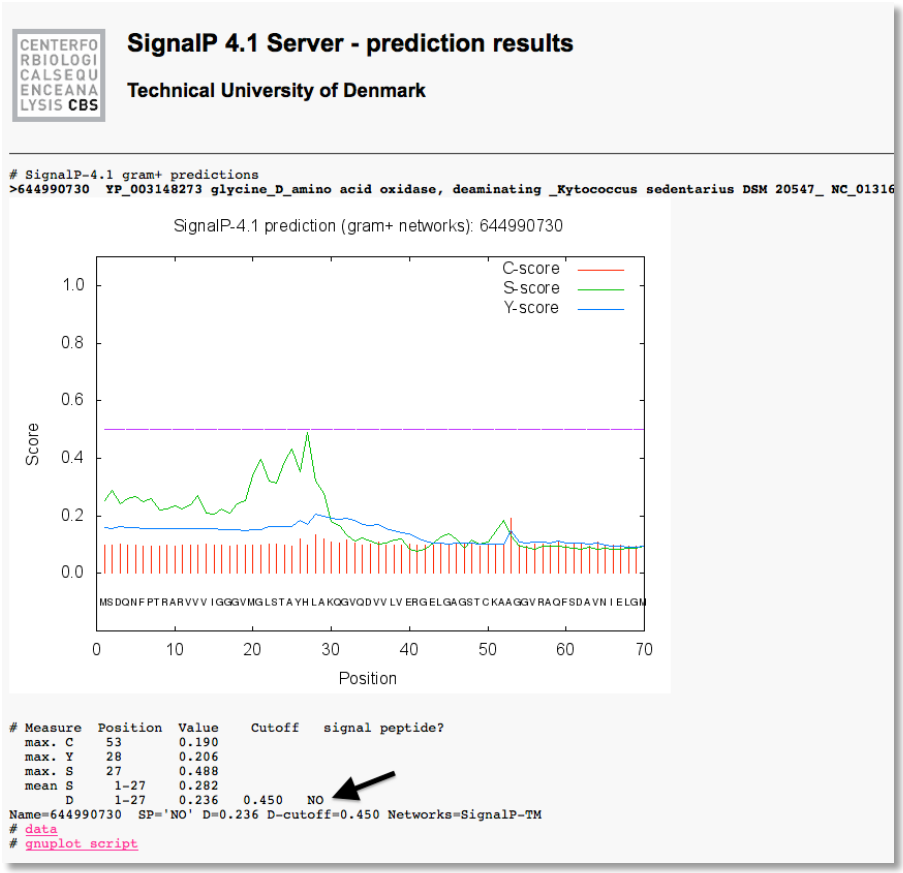
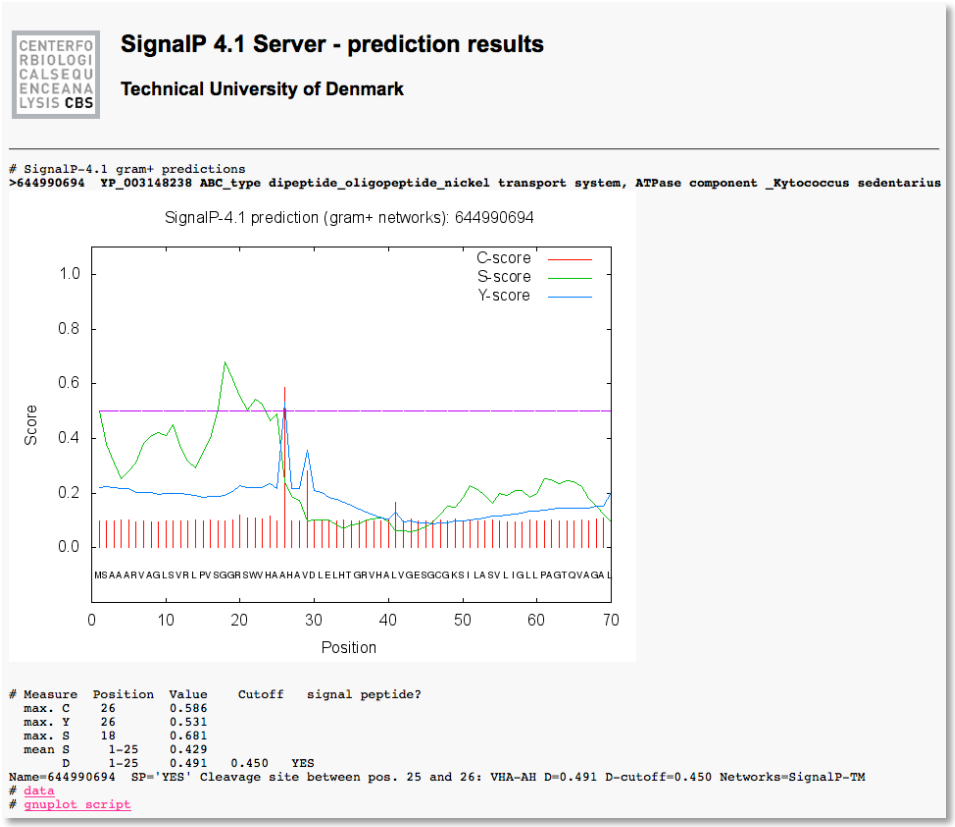


Figure 4.12. A SignalP output indicating that no signal peptide is present



SignalP 4.1 Server - prediction results

Technical University of Denmark

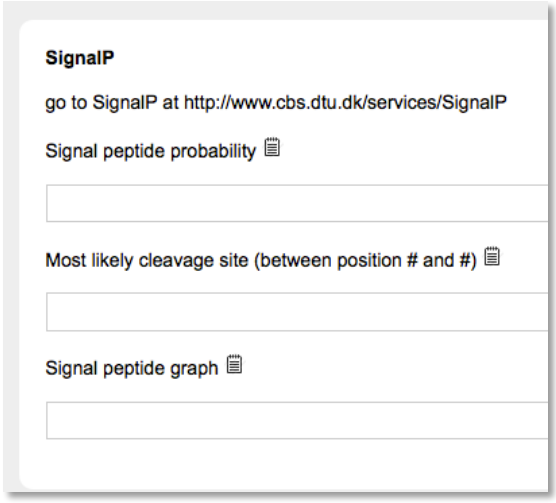
```
# SignalP-4.1 gram+ predictions
>644990694  YP_003148238 ABC_type dipeptide_oligopeptide_nickel transport system, ATPase component _Kytococcus sedentarius
```

SignalP-4.1 prediction (gram+ networks): 644990694


```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    26       0.586
  max. Y    26       0.531
  max. S    18       0.681
  mean S     1-25    0.429
       D     1-25    0.491   0.450   YES
Name=644990694  SP='YES' Cleavage site between pos. 25 and 26: VHA-AH D=0.491 D-cutoff=0.450 Networks=SignalP-TM
# data
# gnuplot script
```

Figure 4.13. A SignalP output showing the presence of a signal peptide. See text for discussion.

15. Determine whether your protein has a signal peptide by analysis of the graphical output and add the results to the lab notebook (Figure 4.14). If no signal peptide is predicted, type " no signal peptide predicted" in the signal peptide probability box. If one is predicted, write the probability of there being one in the box (the value is 0.491 in the example shown in figure 4.13; found immediately after the statement of where the likely cleavage site is located and in the table after D score). The probability must be greater than 0.450 for a signal peptide to be predicted.
16. Once the results are obtained and they have been analyzed, Snip or Capture the image and associated text file and save it as png.
17. Be sure to save your notebook periodically by clicking on the save button at the bottom of the module section.



SignalP

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

Signal peptide probability

Most likely cleavage site (between position # and #)

Signal peptide graph

Figure 4.14. The SignalP notebook.

PSORTb – this tool will analyze your sequence and attempt to predict whether your protein is soluble in the cytoplasm, an integral membrane protein or a secreted protein.

1. Go to http://www.psort.org/psortb/ . This will take you to the PSORTb start page (Figure 4.15).

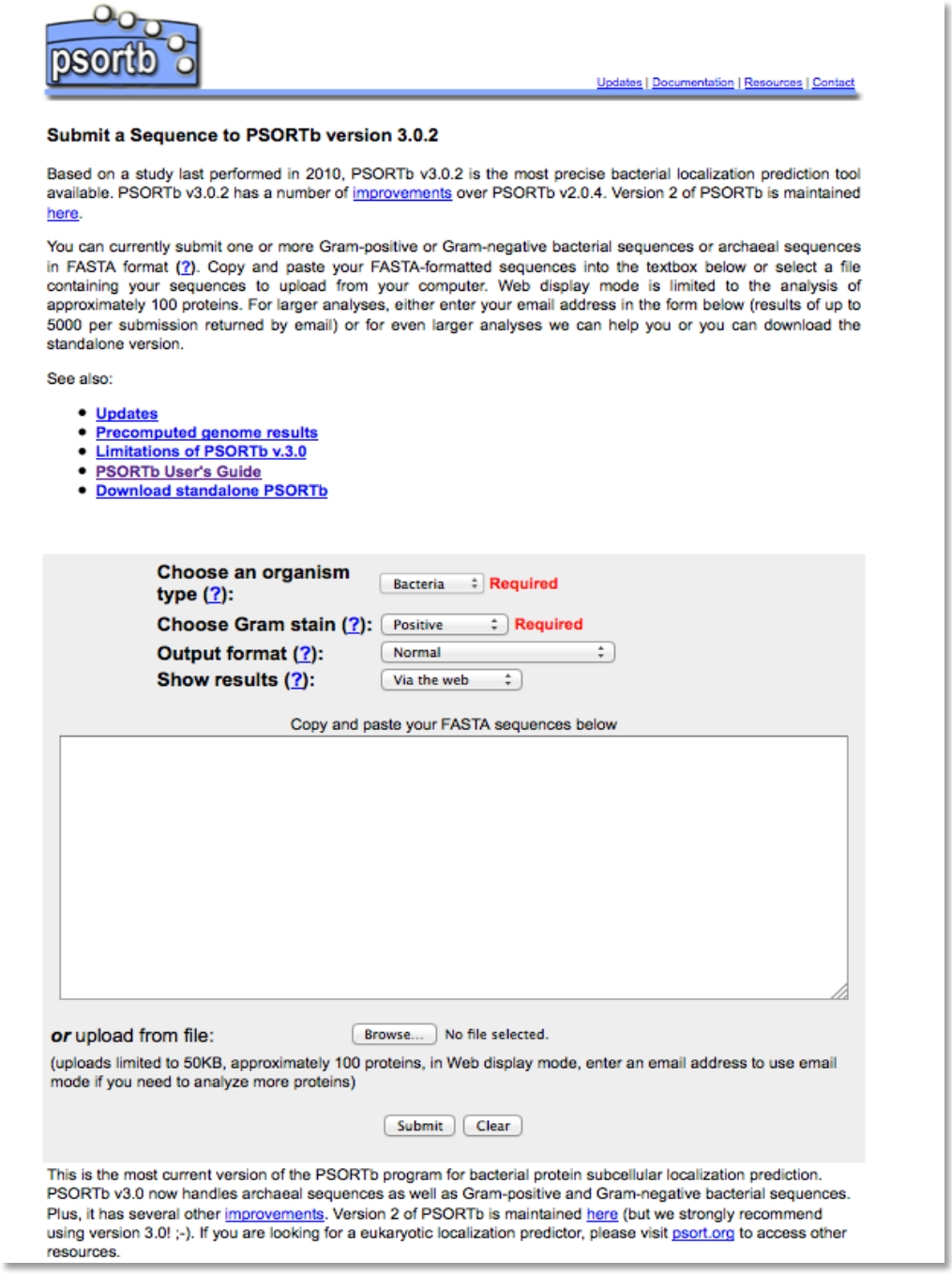


Figure 4.15. The PSORTb start page.

2. Paste the amino acid sequence in FASTA format into the space provided on the webpage.
3. Select the type of organism your amino acid sequence belongs to.
4. Select the appropriate Gram stain for your bacterium (Gram positive or Gram negative).
5. Select normal output.
6. Select show results via the web.
7. Click submit.
8. Record all the data into the notebook (cytoplasmic, cytoplasmic membrane, cell wall, and extracellular scores) (Figure 4.16). **If any one score is greater than 7.5 it will be the final predicted location for your protein. If no score is greater than 7.5 the prediction may be that the protein can exist in multiple sites. If all scores are equal then PSORTb is unable to assign a final predicted location.**

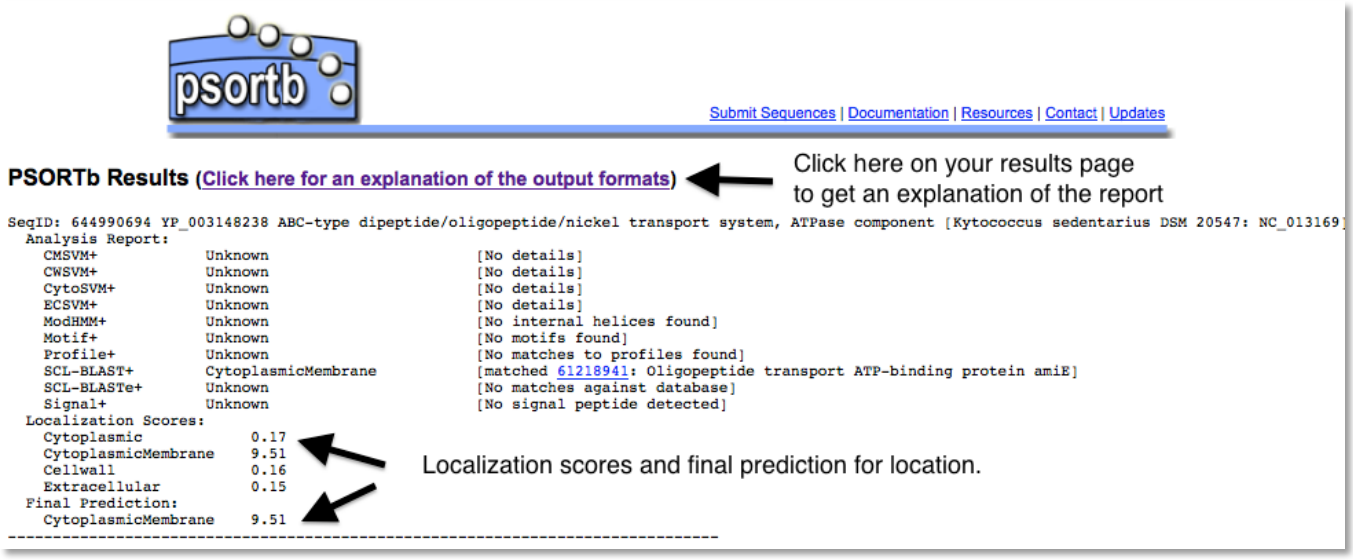


Figure 4.16. A PSORTb results page.

9. Record the PSORT final prediction in the notebook (Figure 4.17). Note that if you have a Gram + bacterium, that you will not get an outer membrane score or a periplamsic score (recall that only Gram - bacteria have an outer membrane and periplasmic space). Simply write "NA" for Not Applicable in the notebook for OuterMembrane and Periplasmic scores for Gram + organisms.

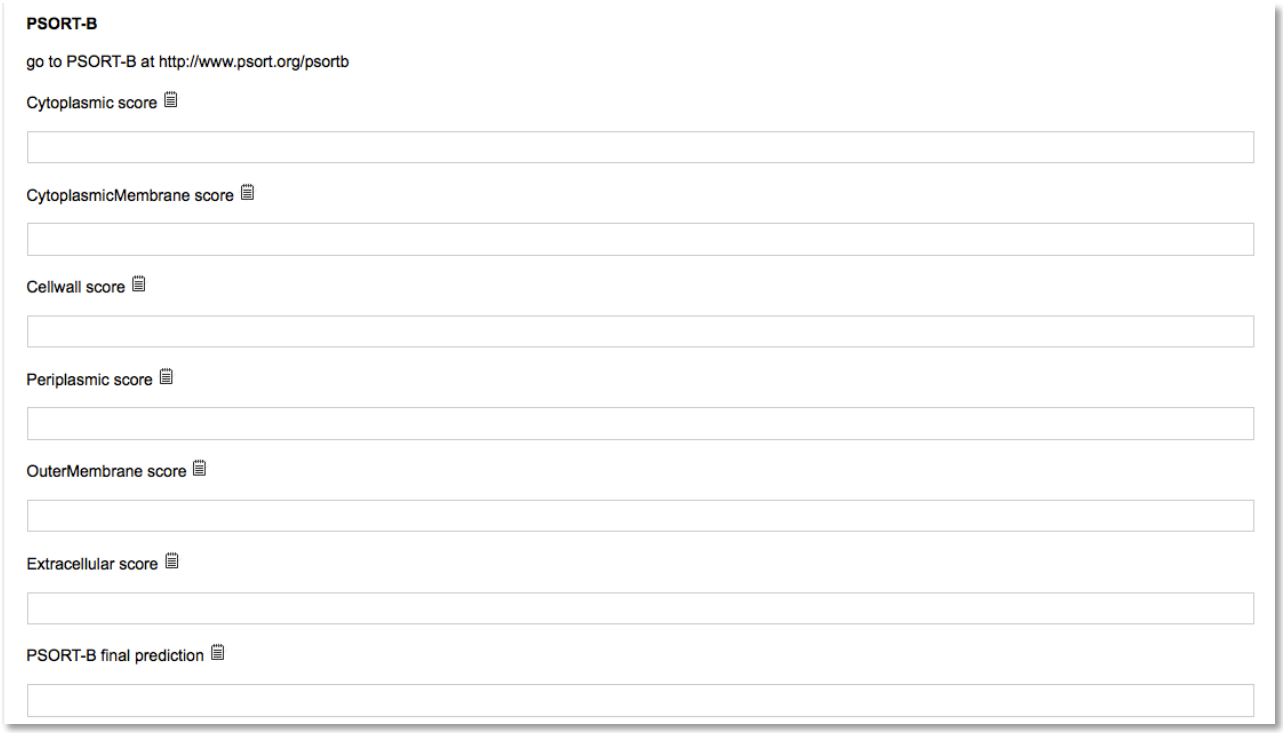


Figure 4.17. The PSORTb notebook page. You will not obtain an outer membrane or periplasmic score for Gram + bacteria (see text).

Phobius – this tool will combine the results of TMHMM and signalP in one output and acts as a confirmatory test for transmembrane helixes and signal peptides.

1. Go to http://phobius.sbc.su.se/. This will take you to the Phobius start page (Figure 4.18).

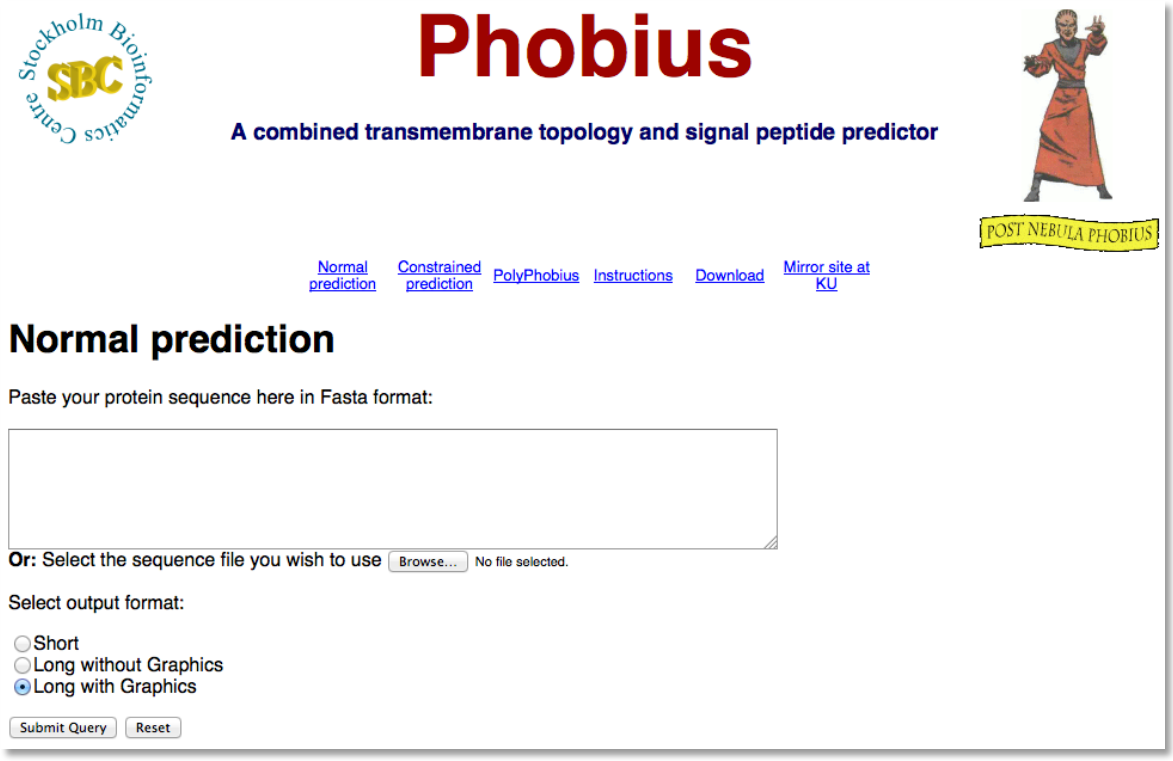


Figure 4.18. The Phobius start page.

2. Paste the amino acid sequence in FASTA format into the space provided.
3. Select long with graphics.
4. Click submit. You will see an output as illustrated in Figure 4.19. The red line indicates the probability of a signal peptide and the blue and green lines indicate portions of the protein that are non-cytoplasmic (extracellular) or cytoplasmic (intracellular), respectively. Figure 4.19 is the same protein used to demonstrate the presence of a signal peptide with the signalP tool above. Note that Phobius also identifies the signal peptide and predicts that the protein will be non-cytoplasmic (extracellular or secreted).

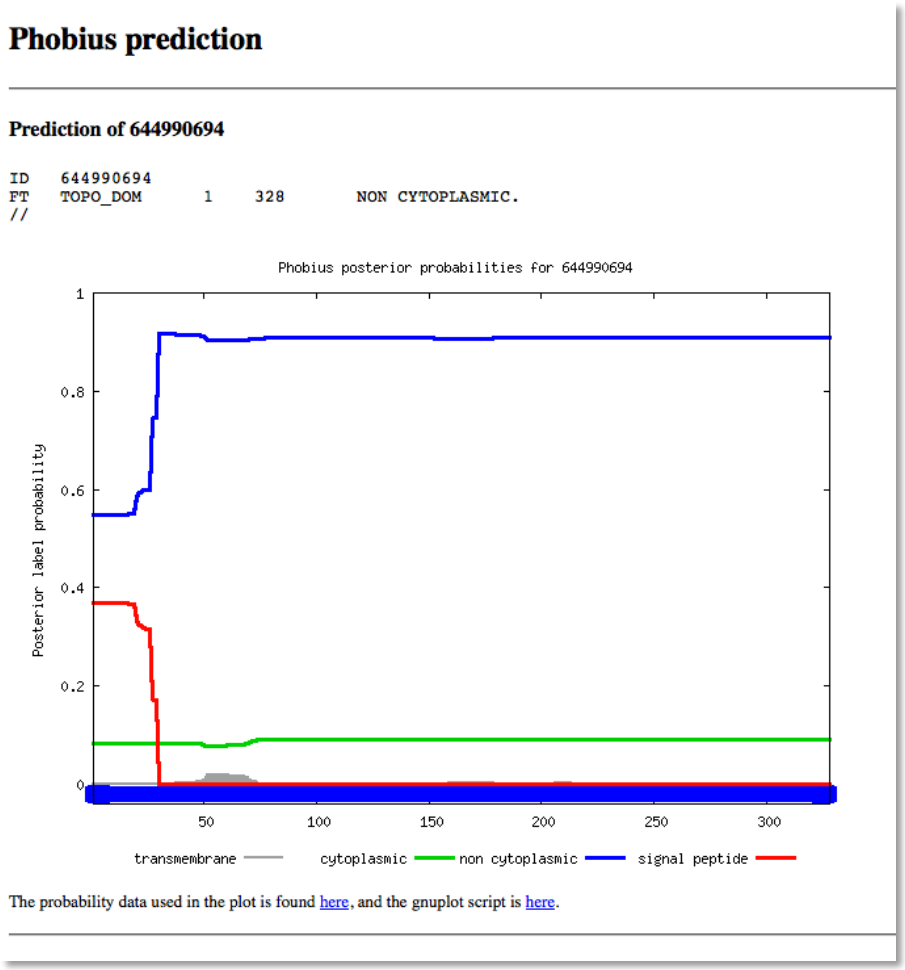


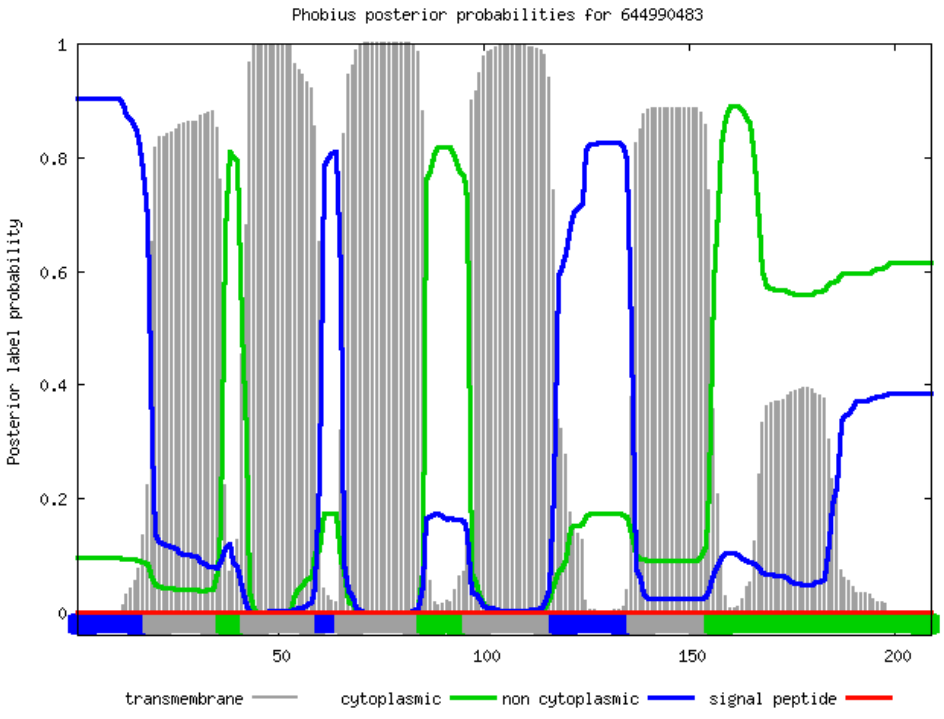
Figure 4.19. A Phobius output showing the presence of a signal peptide. See the text for an explanation.

5. Figure 4.20 shows the Phobius results for the same protein used to demonstrate what transmembrane helixes look like in TMHMM (see figure 4.10). In this example we see how the two tools can disagree. Phobius identifies 5 helixes instead of the 3 identified by TMHMM. If you compare the images you will see that two helixes that did not meet the cutoff in TMHMM are the ones predicted to be helixes by Phobius.

Phobius prediction

Prediction of 644990483

```
ID   644990483
FT   TOPO_DOM      1     18       NON CYTOPLASMIC.
FT   TRANSMEM     19     36
FT   TOPO_DOM     37     42       CYTOPLASMIC.
FT   TRANSMEM     43     60
FT   TOPO_DOM     61     65       NON CYTOPLASMIC.
FT   TRANSMEM     66     85
FT   TOPO_DOM     86     96       CYTOPLASMIC.
FT   TRANSMEM     97    117
FT   TOPO_DOM    118    136       NON CYTOPLASMIC.
FT   TRANSMEM    137    155
FT   TOPO_DOM    156    209       CYTOPLASMIC.
//
```



The probability data used in the plot is found here, and the gnuplot script is here.

Figure 4.20. A Phobius output showing the presence of 5 transmembrane helixes. See the text for details.

6. Capture the graphical output in an image and save it to your notebook (Figure 4.21).

Module Summary – the last entry in your notebook, as shown in figure 4.21, is where you will state your hypothesis of where your protein is located. Look at all the results from the different tools to come to your conclusion. You can point out if there are differences in the results between the tools (for instance the difference in the number of helixes for the protein above in TMHMM vs. Phobius) here as well. Do not simply write that it is cytoplasmic or non-cytoplasmic, but rather write a brief explanation of why you think it is in that location. If you are unable to come up with a location for your protein, you can state that as well, along with an explanation of why you cannot determine the location.

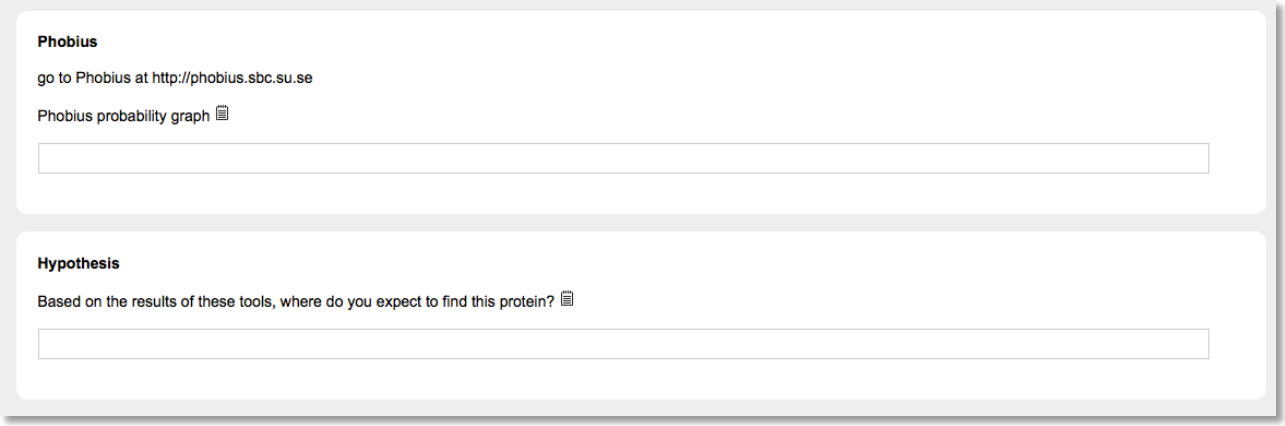


Figure 4.21. The Phobius and final hypothesis notebook sections