Cellular Localization Data

Another important component to learn about your protein is where it physically is in the cell. This information will be critical to helping you discern between possible functions you have uncovered so far and in proposing a final protein function or action based on where it is localized and what else is there for the protein to act on. The tools described below will help you predict where your gene's product is most likely to be found in the cell based on its sequence patterns. Each tool adds an additional layer of analysis to the others. *Saccharomyces cerevisiae* is a single-celled eukaryote; thus, it is important that you understand the structure of the cellular membranes including those of membrane-bound organelles and the plasma membrane, in you need a refresher basic eukaryotic cell structure can be reviewed in any cell biology textbook or at reputable websites.

Transmembrane Helices Hidden Markov Models (TMHMM)

The TMHMM tool is used to assess the likelihood that some portion of a protein is embedded in a cellular membrane, and to predict the path that the amino acid chain follows if the protein IS membrane-associated. TMHMM compares an amino acid sequence to a database of hidden Markov models or "HMMs" (you have already seen some of these in assessing conserved domains and regions in your predicted protein). ***These were created on the basis of known transmembrane (TM) helical sequences; that is, amino acid sequences demonstrated by experimental means to form α-helical secondary structures that cross cell membranes. If portions of a query sequence appear to be similar to any TMHMM's, then your gene product – assuming it is a protein and not a functional RNA molecule – may be localized to one of the cellular membranes. The search results will include a prediction of which segments of the protein are most likely to lie within the cytoplasm, to span a membrane, or to lie on the external side of the plasma membrane.***

Note that predictions of "inside" or "outside" location for a protein segment should be viewed cautiously, since it can be difficult to determine if a potential helical structure completely crosses a membrane without experimental evidence. Also be aware that these predictions are not valid at all if your protein sequence does not include any membrane-spanning helices.

Navigate to TMHMM at http://www.cbs.dtu.dk/services/TMHMM/ and enter the FASTA-formatted protein sequence in the search box. [*Recall this can be downloaded from your protein data on your gene page of SGD*] Click the "Submit" button to begin your search.

You will be presented with a graph that shows the positions of amino acids in the primary sequence on the X axis (these are numbered in the N-terminal to C-terminal direction) and the probability of being in a particular location on the Y axis. Vertical red lines beneath the curves indicate portions of the sequence that match a TMHMM and are likely to enter or cross a membrane. The blue line represents the probability that a given portion of the sequence lies inside the cytoplasm and the pink line represents the probability of being external to the plasma membrane or outer membrane. **REMEMBER THAT THE "INSIDE" AND "OUTSIDE" PREDICTIONS SHOULD BE VIEWED CAUTIOUSLY AT BEST AND COMPLETELY IGNORED IF THE SEQUENCE DOES NOT INCLUDE ANY REGIONS THAT ARE LIKELY TO CROSS A MEMBRANE COMPLETELY.**

Be alert for predicted transmembrane segments at the far N terminus of the protein. These could be **signal peptides** (see next section) rather than functional membrane-spanning regions of a mature protein.

Record the number of predicted transmembrane helices (Number of predicted TMHs:) in your Module 4 Worksheet. Also record the amino acid sequence ranges covered by these TMHs.

Save the graph generated by TMHMM and upload it to the designated space in your Module 4 Worksheet or print the diagram separately and include this document in your notebook at the appropriate point.

You will need to examine information obtained in the following sections before settling on a probable cellular location for your protein.

SignalP

A signal peptide (SP) is an amino acid sequence found at the N terminus of newly-translated proteins that serves as a ticket of sorts to tell the cell where the protein needs to go. They can be used to signal for proteins to be secreted or for them to move into a particular organelle or so that they can become integrated into a membrane. This works because the signal peptide directs the ribosome synthesizing the protein to a transport complex on the cytoplasmic face of a membrane, where the emerging polypeptide is threaded into or through the membrane. In most cases, the signal peptide is later cleaved from the protein as part of a maturation process; the enzyme responsible for this cleavage is called a "signal peptidase". SignalP is a tool that attempts to predict SP regions on the basis of amino acid sequence similarity to known signal peptides.

Navigate to http://www.cbs.dtu.dk/services/SignalP. Enter the entire amino acid sequence in FASTA format. Select "Eukaryotes" under "Organism group" and click "Submit".

The graphical output from SignalP (neural network) comprises three different scores: *C*, *S* and *Y*. Two additional scores, the *S-mean* and the *D-score*, are reported in the SignalP3-NN output but these are only stated as numerical values.
For each class of organism, two different neural networks are used: one for predicting the actual signal peptide and one for predicting the position of the **signal peptidase I** (SPase I) cleavage site. **The *S-score* for the signal peptide prediction is reported for each individual amino acid position in the submitted sequence, with high scores indicating that the corresponding amino acid is likely to be part of a signal peptide and low scores indicating that the amino acid is part of a mature protein. The *C-score* is the "cleavage site" score. This is reported for each position in the submitted sequence and should only be significantly elevated at the cleavage site.**

The hidden Markov model calculates the probability that the submitted amino acid sequence contains a signal peptide, and the cleavage site (if applicable) is also assigned on the basis of a probability score. If a probable signal peptide is identified, scores indicating the most likely locations of the n-region, h-region, and c-region are reported. The SignalP value is considered significant if it is >0.45.

Upload the HMM graph to the Module 4 Worksheet.

Report the Signal Peptide Probability in the Module 4 Worksheet.

Y-max is a derivative of the C-score combined with the S-score, and is a better predictor of cleavage-site location than the raw C-score alone. This is because multiple high-peaking C-scores can be found in one sequence, even though only one of them is the actual cleavage site. The cleavage site is assigned from the Y-score where the slope of the S-score is steep and a significant C-score is found. *Note: Position numbering of the cleavage site can often be a source of confusion. When a cleavage position is shown as a single number, it refers to the first residue in the mature protein after removal of the signal peptide. [Reminder: the term "residue" refers to an amino acid that is part of a protein or peptide. Dehydration synthesis results in removal of a water molecule, and what remains of each amino acid is called an amino-acid residue]. If a cleavage site is reported as being between two amino acid residues – e.g. "amino acids 26-27" – this means that the mature protein begins with the second residue shown (number 27 in the example) and that the signal* *peptide terminates with the first.* The *S-mean* is the average of the S-scores, ranging from the N-terminal amino acid to the amino acid that is assigned the highest Y-max score. Thus, the S-mean score is calculated for the entire length of the predicted signal peptide. The S-mean score was used in SignalP versions 1.0 and 2.0 as the criterion for discrimination of secretory and non-secretory proteins. The *D-score* was introduced with SignalP version 3.0 and is a simple average of the S-mean and Y-max score. The D-score is superior to the S-mean score in discriminating secretory from non-secretory polypeptides. Note that for non-secretory proteins, all the scores presented in the SignalP3-NN output should be very low, at least in theory.

Philius

Hidden Markov models (HMM) have been successfully applied to the tasks of transmembrane protein topology prediction and signal peptide prediction. Philius, is inspired by a previously published HMM, Phobius, and combines a signal peptide sub-model with a transmembrane sub-model. We are running this algorithm in addition to TMHMM because the graphics are in a different output format and may help you with visualization of the different parts of your protein.

Navigate to Philius at **http://www.yeastrc.org/philius/pages/philius/runPhilius.jsp** and enter the FASTA-formatted protein sequence in the search box. [*Recall this can be downloaded from your protein data on your gene page of SGD*] Click the "Run Philius" button to begin your search.

You will be presented with a Predication Overview, and possibly a Predication Image Map and Prediction Sequence Map; look over these three pieces of data and enter the results in your Module 4 Worksheet.

Within the Prediction Overview you will be provided with a summary based on the combined data from the Philius algorithm. Record the Predicted Protein Type – this will indicate if the system predicts the protein is globular (not in a membrane) or a transmembrane protein. You should also record the Type and Topology Confidence values.

The Prediction Image Map gives a cartoon display of where any transmembrane, cytoplasmic, non-cytoplasmic and signal peptide region are located within your protein sequence. **If present, copy this image and include it in your Module 4 Worksheet.**

The Prediction Sequence Map provides colorimetric data about the localization of these regions and the confidence values of these predictions. **If present, copy this image with the color detail and include it in your Module 4 Worksheet.** Scroll your mouse over of the colored sections of the sequence and record the confidence values for these predictions.

TargetP

TargetP 1.1 predicts the subcellular location of eukaryotic proteins. The location assignment is based on the predicted presence of any of the N-terminal presequences: chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP) or secretory pathway signal peptide (SP). For the sequences predicted to contain an N-terminal presequence a potential cleavage site can also be predicted.

Navigate to TargetP at http://www.cbs.dtu.dk/services/TargetP/ and enter the FASTA-formatted protein sequence in the search box. [*Recall this can be downloaded from your protein data on your gene page of SGD*] Under the organism group – click "Non-plant". Under the cutoff category – click "no cutoffs; winner-takes-all (default)". Click Submit to begin your search.

When your job has finished you will either be directed to the results page or a new line of text will appear that says “The job has finished, press here to show results” Click the here hyperlink **on the webpage** if the results do not load automatically.

Since we are utilizing the non-plant database your results will return three relevant numbers: 1) mTP 2) SP 3) other. These numbers are reported as probabilities out of 1 that your protein contains a mitochondrial targeting peptide or a secretory pathway signal peptide or neither of these. Target has the following to say about the scores it reports and how you interpret the data:

Reliability classes, from 1 to 5, are determined, where 1 indicates the strongest prediction. RC is a measure of the size of the difference ('diff') between the highest (winning) and the second highest output scores. There are 5 reliability classes, defined as follows:

1 : diff > 0.800
2 : 0.800 > diff > 0.600
3 : 0.600 > diff > 0.400
4 : 0.400 > diff > 0.200
5 : 0.200 > diff

Thus, the lower the value of RC the safer the prediction.

Record your values and the RC score on your Module 4 Worksheet and comment about whether you believe your protein contains either of these peptide, both or neither and how confident you are in that prediction.

NucPred

NucPred (pronounced newk-pred) analyzes a eukaryotic protein sequence and predicts if the protein spends at least some time in the nucleus or spends no time in the nucleus. NucPred is an ensemble (or jury) of 100 sequence based predictors. Each is given the sequence of interest and provides a "yes" or "no" answer to the question "does the protein spend some time in the nucleus?" If the fraction of predictors giving a "yes" answer (also known as the NucPred score) exceeds some prior agreed threshold, then the protein is predicted to have a nuclear role. Don't forget that proteins can have multiple functions and/or multiple subcellular locations. However, if a protein is already known to be secreted or is an integral membrane protein, a second role as a nuclear protein is not likely. NucPred will make a small number of confident but contradictory predictions like this. So please use all sources of biological information (both real and predicted) when interpreting the results.

Navigate to NucPred at http://www.sbc.su.se/~maccallr/nucpred/, click the Single Protein hyperlink and enter the FASTA-formatted protein sequence in the search box. [*Recall this can be downloaded from your protein data on your gene page of SGD*] Click Submit Query to begin your search.

Your sequence will be returned to you in color, with the colors representing the probability of any giving amino acid being part of a nuclear localization signal.

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

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

There also may be regions of the sequence that are underlined, if so this is the main part of the sequence that is contributing to your NucPred score. That is, these are the amino acids, usually lysines and arginines (Ks and Rs) that compose the nuclear localization signal or NLS.

Copy the sequence of your protein from this page (in color) and copy it into the space provided in your Module 4 Worksheet. Comment in the space provided on if any regions of your protein likely contain a NLS signal and where they are located if they are present.

What does the NucPred score mean?

You have to decide on a **NucPred score threshold**. Sequences which score greater than or equal to this threshold are predicted to spend some time in the nucleus. Higher thresholds yield fewer predicted nuclear proteins, but these predictions are more accurate (you can have higher confidence in them). The table below gives more details of the performance of NucPred estimates using the sequences it was trained on (by cross-validation).

NucPred score threshold	Specificity
	fraction of proteins predicted to be nuclear that actually are
0.1	0.45
0.2	0.52
0.3	0.57
0.4	0.63
0.5	0.7
0.6	0.71
0.7	0.81
0.8	0.84
0.9	0.88
1	1

Yeast Protein Localization Database (YPL)

Welcome to the Yeast Protein Localization Database, YPL.db. The intention of this site is to provide information about the subcellular localization of proteins in the yeast *Saccharomyces cerevisiae*. For the localization studies YPL.db has used the GFP fusion technique (a green fluorescent protein tag is added to the end of each protein) and Confocal Laser Scanning Microscopy (CLSM).

Navigate to http://yeastgfp.yeastgenome.org/. Enter your gene name in the Quick Search box and hit GO. If your gene-fused to GFP has been tested for localization then a box will appear that indicates the estimated number of molecules/cell and also says, “please click cartoons at right to view cell image”. **Record the molecules/cell information in your Module 4 Worksheet and then Click “please click cartoons at right to view cell image” to display the actual data image files of where your GFP-fused protein appears in the cell. Scroll through some of the pictures and select a good representative to include in your Module 4 Worksheet and record the site of localization.**

Hypothesis - *Where do you expect to find your protein?*

Take the results of all of the above analyses into consideration and make a final localization prediction. Did all the tools yield the same result? If one disagreed with the others, what might that tell you about the protein's function, based on the prediction method used in that tool? **Record your final prediction, with justification, in the Module 4 Worksheet**. If the combined results of the analyses are inconclusive, enter “Unknown”. Feel free to consult one of the instructors if you’re unsure about this.