Yeast ORFan Gene Project: Module 5 Guide

Cellular Localization Data (Part 1)

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 the options of membrane-bound organelles and the plasma membrane.

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)[1]. 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 your FASTA-formatted protein sequence in the search box. [*Recall this can be downloaded from the Protein tab on your gene page of SGD*] Leave the default **Output Format** as **Extensive, with graphics** and click the "Submit" button to begin your search.

You will be presented with a graphic 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 curve indicate portions of the sequence that match a Transmembrane Helix 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 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 in your **Module 5 Worksheet**. You may also want to record the amino acid sequence ranges covered by these TMHMMs.

Save the graphic generated by TMHMM and upload it to the designated space in your Module 5 Worksheet.

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

[1]A Hidden Markov Model is a mathematical transformation that can reveal the existence of patterns in a data set.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.

Yeast ORFan Gene Project: Module 5 Guide

SignalP

The signal peptide (SP) is an amino acid sequence found at the N terminus of newly-translated proteins that need to be secreted or become integrated into a membrane. The signal peptide directs the ribosome synthesizing the protein to a transport complex on the cytoplasmic face of the 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 (euk networks) comprises three different scores: *C*, *S* and *Y*. Two additional scores, the *S-mean* and the *D-score*, are reported in the SignalP-4.1 output but these are only stated as numerical values.
For each class of organism, two different 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. 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. *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.*

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 SignalP-4.1 output should be very low, at least in theory. *

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 >0.75.

Upload the HMM graph to the Module 5 Worksheet.

Report the Signal Peptide Probability in the Module 5 Worksheet.

[1]A Hidden Markov Model is a mathematical transformation that can reveal the existence of patterns in a data set.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.

Yeast ORFan Gene Project: Module 5 Guide

PSORT II

PSORT II is another useful tool for predicting the subcellular location of a protein. PSORT II predicts the subcellular localization sites of proteins based on their amino acid sequences. The method makes predictions based on both known sorting signal motifs and some correlative sequence features such as amino acid content. **In this version of PSORT, parameters for analyzing yeast (or plant) sequences are the same with parameters for animal sequences. Yeast (and plant) have a candidate site, vacuole, instead of lysosome in animal. In yeast, the consensus sequence for ER-lumen retention is HDEL rather than KDEL in others.***

Navigate to PSORT II at http://psort.hgc.jp/form2.html. Enter the amino acid sequence in FASTA-format in the space provided (NOTE: this program requires the REMOVAL of the FASTA header which is the >genename). Select "Yeast/Animal" for source of input sequence, the click Submit to run the search.

Scroll down the Results page to view the localization predictions. The scores will vary depending on the particular algorithm that has been run. Interpreting this data will take attention to detail, reading through the explanations below and determining how to deduce meaning from each component. As you go through each section paste the information in your **Module 5 Worksheet** ***and comment on how you have interpreted the data***.

Recognition of Signal Sequence

In eukaryotes, proteins sorted through the so-called vesicular pathway (bulk flow) usually have a signal sequence (also called a leader peptide, and different in character from the signal peptide described above) in the N- terminus, which is cleaved off after the translocation through the ER membrane. Some N-terminal signal sequences are not cleaved off, remaining as transmembrane segments but it does not mean these proteins are retained in the ER; they can be further sorted and included in vesicles. *

PSORT first predicts the presence of signal sequences, it considers the N-terminal positively-charged region (N-region) and the central hydrophobic region (H-region) of signal sequences. A discriminant score is calculated from the three values: length of H-region, peak value of H-region, and net charge of N-region. These results are summarized in **"PSG"**. A large positive discriminant score means a high possibility to possess a signal sequence but it is unrelated to the possibility of its cleavage. *

Next, PSORT applies a weight-matrix method and incorporates the information of consensus pattern around the cleavage sites (the (-3,-1)-rule) as well as the feature of the H-region. Thus it can be used to detect signal-anchor sequences. The output score of this **"GvH"** is the original weight-matrix score (for eukaryotes) subtracted by 3.5. A large positive output means a high possibility that it has a cleavable signal sequence. The position of possible cleavage site, *i.e.,* the most C-terminal position of a signal sequence, is also reported. *

Recognition of Transmembrane Segments

PSORT II assumes that all integral membrane proteins have hydrophobic transmembrane segment(s) which are thought to be alpha- helices in membranes. PSORT detects potential transmembrane segments finding the most probable transmembrane segment from the average hydrophobicity value of 17-residue segments, if any. It predicts whether that segment is a transmembrane segment (INTEGRAL) or not (PERIPHERAL) comparing the discriminant score (reported as 'ALOM score') with a threshold parameter. For an integral membrane protein, position(s) of transmembrane segment(s) are also reported. Their length is fixed to 17 but their extension, *i.e.,* the maximal range that satisfies the discriminant

[1]A Hidden Markov Model is a mathematical transformation that can reveal the existence of patterns in a data set.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.

criterion, is also given in parentheses. The item 'number of TMSs' is the number of predicted transmembrane segments. Specifically, PSORT first tentatively evaluates the number of TMSs using less stringent value (0.5). Then, it re-evaluates the number by using a more stringent threshold (-2.0). If it is still predicted to have at least one TMS, the former threshold value is used. Record the ALOM Score and number of TMs.*

Recognition of Mitochondrial Proteins

Although many proteins which engage in mitochondrial protein targeting have been characterized, their exact pathways have not been fully understood. Many proteins transported to mitochondria have a mitochondrial targeting signal on their N-terminus. Some seem to have internal signals but they may be recognized by a common cytosolic factor (MSF).
PSORT employs a very simple method to recognize mitochondrial targeting signals: the discriminant analysis (called **"MITDISC"**) whose variables are the amino acid composition of the N-terminal 20 residues. In this version, further discrimination of signals directing to the substructures of mitochondria, *e.g.,* intermembrane space, is not attempted although proteins which are predicted to have both the mitochondrial targeting signal and transmembrane segment(s) are likely to be localized at the inner membrane. PSORT also reports some consensus patterns around the cleavage sites ("Gavel").*

Note the score and data for this section but wait until completion of TargetP in the next module before deciding upon mitochondrial localization.

Recognition of Nuclear Proteins

Although it seems possible that a protein without its own nuclear localization signal (NLS) enters the nucleus via cotransport with a protein that has one, many nuclear proteins have their own NLSs. Presently, NLSs are classified into three categories. The classical type of NLSs is that of SV40 large T antigen. PSORT uses the following two rules to detect it: 4 residue pattern (called 'pat4') composed of 4 basic amino acids (K or R), or composed of three basic amino acids (K or R) and either H or P; the other (called 'pat7') is a pattern starting with P and followed within 3 residues by a basic segment containing 3 K/R residues out of 4. *

Another type of NLS is the bipartite NLS, first found in *Xenopus* nucleoplasmin. The pattern (called 'bipartite') is: 2 basic residues, 10 residue spacer, and another basic region consisting of at least 3 basic residues out of 5 residues.
The last category of NLS is the type of an N-terminal signal found in yeast protein, Mat alpha2. However, PSORT doesn't try to find it because this type of signal has not been well studied. Nor the knowledge of Nuclear Export Signals (NESs) has not been incorporated yet. *

In the yeast genome, nuclear proteins occupy the majority. Since the precise discrimination of NLSs is presently difficult, the prediction of nuclear proteins affects much to the total prediction accuracy. Then, PSORT uses a heuristic that nuclear proteins are generally rich in basic residues: If the sum of K and R compositions are higher than 20%, then the protein is considered to have higher possibility of being nuclear than cytoplasmic. Moreover, a score (called **"NNCN"**) which discriminates the tendency to be at either the nucleus or the cytoplasm is calculated based on the amino acid composition. Most scores mentioned above are combined by a discriminant function to give the 'NLS score'. In addition, PSORT examines the presence of RNP (ribonucleoprotein) consensus motif (called 'RNA-binding motif') because some RNPs are transported to the nucleus by signals existing in the bound RNAs. However, it is apparently insufficient for actual prediction. In this version, we classify ribosomal proteins as cytoplasmic proteins although some of them have NLSs and are once transported into the nucleus. *

[1]A Hidden Markov Model is a mathematical transformation that can reveal the existence of patterns in a data set.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.

Yeast ORFan Gene Project: Module 5 Guide

Note the score and data for this section but wait until completion of NucPred in the next module before deciding upon nuclear localization.

Recognition of ER (endoplasmic reticulum) Proteins

PSORT postulates that the proteins with N-terminal signal sequence will be transported to the cell surface by default unless they have any other signals for specific retrieval, retention, or commitment; a luminal protein will be secreted constitutively to the extracellular space and a membrane protein will reside at the plasma membrane. The retrieval signal of ER luminal proteins from the bulk flow is the consensus motif, KDEL (HDEL in yeast), in the C-terminus. In addition, these proteins should have a cleavable signal sequence in their N-terminus but the existence of KDEL is often practically sufficient. Although PSORT only recognizes the (K/H)DEL pattern, it is known that some variations of this motif are allowed in some organisms and/or cell types. *

The retrieval signals for ER membrane proteins appear more complex. Two kinds of signals are known; one is the di-lysine motif (the KKXX motif) which exist near the C-terminus of type Ia proteins and the other is the di-arginine motif (the XXRR motif) which exist near the C-terminus of type II proteins. Note that both of these motifs exist close to the terminus of the cytoplasmic tail. However, for the practical prediction, the existence of these motifs itself is neither necessary nor sufficient for the localization at the ER membrane. *

Recognition of Peroxisomal Proteins

Peroxisomes, sometimes called glyoxisomes, glycosomes, or microbodies, are organelles found in almost every eukaryotic cell. Several sorting signals into peroxisomes have been. As for peroxisomal-matrix targeting sequences (PTSs), two kinds of them are known. One is the tripeptide, (S/A/C)(K/R/H)L, at the C-terminus, known as PTS1 or the SKL-motif. PSORT calculates score of a given sequence empirically. The other is the N-terminal segment known as PTS2. The consensus patterns of known PTSs, (R/K)(L/I)xxxxx(H/Q)L is searched in the present version, although its importance has not been fully verified. The sorting signal of peroxisomal membrane proteins (mPTSs) is not well understood although the importance of a hydrophilic loop of 20 residues facing the matrix in Pmp47 is known. *

Analysis of Proteins in Vesicular Pathway

The signals that govern the protein sorting through the vesicle transport are complex because they are inter-related and they seem to be recognized in various situation. Moreover, there are redistribution processes via endocytosis and a pathway to lysosomes. As already exemplified above, many sorting signals of membrane proteins transported through these pathways exist in the cytoplasmic tail, a short terminal segment exposed to the cytosol in type Ia, Ib, and II proteins. For membrane proteins with (usually) a single transmembrane segment, PSORT tries to find the following motifs in the cytoplasmic tail: the YQRL motif which directs the transport from cell surface to Golgi; the tyrosine-containing motif and the dileucine motif for selective inclusion in clathrin-coated vesicles (endocytosis) and lysosomal targeting. *

Lysosomal and Vacuolar Proteins

Lysosomes are acidic organelles that contain numerous hydrolytic enzymes. In yeast and plant cells, similar activities are seen in vacuoles (lysosome-like vacuoles) and the protein sorting mechanisms for these organelles are conserved to some degree.*

In mammalian lysosomes, one sorting signal of soluble (lumenal) proteins is a post-translational modification, addition of mannose-6-phosphate, but there is also a pathway which is independent of mannose-6-phosphate. Two kinds of

[1]A Hidden Markov Model is a mathematical transformation that can reveal the existence of patterns in a data set.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.

mannose-6-phosphate receptor are known but the substrate specificity of the enzyme which adds mannose-6-phosphate is not well understood although the importance of a specific conformation, a beta-hairpin motif, has been implicated. *

Although yeast vacuoles have been studied as a model system of mammalian lysosomes, soluble proteins of yeast vacuole do not use the mannose-6-phosphate dependent pathway. Analyses of several yeast proteins suggest that the pro-peptides, which are exposed to the N-terminus after the cleavage of 'pre' signal sequence, work as a sorting signal. However, there are not apparently conserved motifs except for a weak motif, (T/I/K)LP(L/K/I), which PSORT searches for. The sorting mechanism of lysosomal membrane proteins seems different from that of lysosomal luminal proteins. The existence of the GY motif within 17 residues from the membrane boundary in the cytoplasmic tail seems to be important for some of them. *

Lipid Anchors

The protein modification reactions which bind lipid molecules to proteins are important because a linked lipid moiety can be integrated into various membranes and can anchor the bound protein. *

For example, myristoylations occur at the consensus sequence in the N-terminal 9 residues. PSORT predicts its presence by the 'NMYR' program but note that the N-myristoylated proteins are not always anchored to the membrane.
In contrast, all proteins linked to the glycosyl-phosphatidylinositol (GPI) molecules are thought to be anchored at the extracellular surface of the plasma membrane. In addition, GPI anchor plays some roles on the protein sorting in polarized cells. Although much is known about the biosynthesis of GPI-anchor, PSORT predicts GPI-anchored proteins by empirical knowledge that most of them are the type Ia membrane proteins with very short cytoplasmic tail (within 10 residues). *

Lastly, there is a lipid modification known as (iso)prenylation (*i.e.,* farnesylation or geranylgeranylation.) This modification requires a CaaX motif in the C-terminus, where 'a' denotes an aliphatic amino acid. Prenylated proteins have been found in the plasma membrane and the nuclear envelope. *

Miscellaneous Motifs

Experimentally the knowledge of various protein functional motifs in the PROSITE database has been included. To discriminate cytoskeletal proteins, two motifs of actin-binding proteins were examined as well, these PROSITE motifs are experimentally introduced: 63 DNA binding motifs, which may be useful to distinguish nuclear proteins; 71 ribosomal protein motifs, which may be necessary because the sorting processes of ribosomal proteins are complex; 33 prokaryotic DNA binding motifs, which might be useful for the prediction of bacterial sequences. *

Coiled-coil Structure

Coiled-coil structures are found in some structural proteins, *e.g.,* myosins, and in some DNA-binding proteins as the so-called leucine-zipper. In this structure, two alpha-helices bind each other making a coil, where these two alpha-helices show a 3.5-residue periodicity which is slightly different from the typical value, 3.6. Thus, the detection of coiled-coil structure by searching for 7-residue periodicity is relatively more accurate than usual secondary structure prediction. Although the presence of coiled-coil structure in a protein itself does not indicate its subcellular localization, such information might be useful for the people who try to characterize ORFs of unknown function. Currently, a classical detection algorithm is used and the sequences which are likely to be in a coiled-coil conformation are reported. *

[1]A Hidden Markov Model is a mathematical transformation that can reveal the existence of patterns in a data set.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.

Yeast ORFan Gene Project: Module 5 Guide

Ignore the Results of the *k*-NN Prediction for now.

Phobius

***Phobius combines the methods used by TMHMM and SignalP and generates a single graphical output. This tool is especially useful when TMHMM predicts that the N-terminal portion of the sequence forms a transmembrane helix, while SignalP predicts it is a signal peptide. Phobius will help you decide which of these predictions is most likely to be accurate.* ***

Navigate to http://phobius.sbc.su.se/. Enter the amino acid sequence in FASTA-format. Select "Long with Graphics" as your output format and submit the query. The prediction at the top of the output lists the most probable locations of transmembrane helices and a signal peptide (if any) in the sequence. The graphic plot shows the probability that a given amino acid residue is cytoplasmic, non-cytoplasmic, within a TM helix, or part of a signal peptide. Here one can see weak potential TM helices that fell below the prediction threshold and get an idea of the confidence level for each segment in the prediction. At the very bottom of the plot, in the range between 0 and -0.04 on the Y axis, the overall predictions are represented in graphic form. * **If the entire sequence is labeled as being cytoplasmic or non-cytoplasmic, the prediction is that it contains no transmembrane helices. AS WITH TMHMM, YOU SHOULD NOT INTERPRET THE STATED LOCATION AS ACCURATE IN THIS CASE SINCE THE LOCALIZATION MODEL IS ONLY VALID IF THE PROTEIN SPANS A MEMBRANE AT LEAST ONCE**.

Record the graph and probability data in the Module 5 Worksheet.

One might assume that a protein predicted to contain transmembrane helices would include a signal peptide to direct it to the membrane. If this is not the case with your gene product, what might that say about the protein sequence? Note, that examples of "internal signal sequences" that are not cleaved and remain as transmembrane domains of mature proteins have been reported in both bacterial and eukaryotic cells.

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 the final prediction in the **Module 5 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. ***Note – we will do another module on Cellular Localization so this preliminary hypothesis may be changed after further experimentation and addition of data.***

[1]A Hidden Markov Model is a mathematical transformation that can reveal the existence of patterns in a data set.

* Information regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only.