Yeast ORFan Gene Project: Module 6 Guide

Cellular Localization Data – Part II

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.

Philius

"Hidden Markov models (HMM) have been successfully applied to the tasks of transmembrane protein topology prediction and signal peptide prediction. Our model, *Philius*, is inspired by a previously published HMM, Phobius, and combines a signal peptide sub-model with a transmembrane sub-model. We introduce a two-stage (dynamic Bayesian networks) DBN decoder which combines the power of posterior decoding with the grammar constraints of Viterbi-style decoding. Philius also provides protein type, segment, and topology confidence metrics to aid in the interpretation of the predictions. We report a relative improvement of 13% over Phobius in full-topology prediction accuracy on transmembrane proteins, and a sensitivity and specificity of 0.96 in detecting signal peptides. We also show that our confidence metrics correlate well with the observed precision. In addition, we have made predictions on all 6.3 million proteins in the Yeast Resource Center (YRC) database.*"

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 depending on the segments that are found you may be get a Predication Image Map and/or a Prediction Sequence Map; look over these three pieces of data and enter the results in your **Module 6 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. Copy this image and include it in your **Module 6 Worksheet.**

The Prediction Sequence Map provides colorimetric data about the localization of these regions and the confidence values of these predictions. Copy this image with the color detail and include it in your **Module 6 Worksheet**. Scroll your mouse over the colored sections of the sequence and record the confidence values for these predictions.

* 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 6 Guide

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 two relevant numbers: 1) mTP 2) SP. These numbers are reported as probabilities out of 1 that your protein contains a mitochondrial targeting peptide or a secretory pathway signal peptide. Then an RC – Reliability Class – score is reported. *"The RC is from 1 to 5, 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 these values on your **Module 6 Worksheet** and comment about whether you believe your protein contains either of these peptide, both or neither.

NucPred

"NucPred (pronounced *newk*-pred) analyses 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 https://nucpred.bioinfo.se/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 given amino acid being part of a nuclear localization signal.

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

* 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 6 Guide

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

Copy the sequence of your protein from this page (in color) and copy it into the space provided in your Module 6 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. (See the next page for interpretation of your data)

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 estimated using the sequences it was trained on (by cross-validation). *

NucPred score threshold	Specificity	Sensitivity
	fraction of proteins predicted to be nuclear that actually are nuclear	fraction of true nuclear proteins that are predicted (coverage)
0.10	0.45	0.88
0.20	0.52	0.83

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

0.30	0.57	0.77
0.40	0.63	0.69
0.50	0.70	0.62
0.60	0.71	0.53
0.70	0.81	0.44
0.80	0.84	0.32
0.90	0.88	0.21
1.00	1.00	0.02

Yeast Protein Localization Database (YPL)

"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 and Confocal Laser Scanning Microscopy (CLSM). YPL.db is an ongoing project with the aim to provide localization data for all yeast proteins. Labs using the same technique are welcome to submit their data.*"

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 6 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 6 Worksheet** and record the site of localization.

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