Yeast ORFan Gene Project: Module 2 Guide

Structure-Based Evidence

The following tools will help you to obtain additional information about the probable function of your gene's product in the cell based on its predicted structure and similarity to structures of known proteins in several different databases.

Conserved Domain Database Search (CDD)

Proteins often contain several modules or domains, sometimes with distinct evolutionary origins and functions. The Conserved Domain Database of the NCBI is a collection of well-annotated multiple sequence alignment models for conserved domains and full-length proteins. In this module, CDD will be used to find Clusters of Orthologous Groups (COGs) that have significant similarity to the query sequence. A very close COG match to the query can be interpreted as strong likelihood that the query gene belongs in the set of orthologs that were aligned to build the COG model. While COG hits are highly specific, they are not as reliable as those obtained by comparison of a query with more sophisticated conserved-domain models such as TIGRFAMs or Pfams.*

This search is automatically run in parallel with any NCBI BLAST search. After performing a BLAST search, [*go to SGD, enter your gene name, go to the Proteins Tab, scroll to the bottom and under Resources and the Homologs section click on* ***BLASTP at NCBI,*** *click BLAST and wait for your search to complete*] at the top of the page in the *Graphic Summary*, *Show Conserved Domains* section you will see a graphical representation of putative conserved domains (superfamilies, COGS, Pfams, etc.), if any have been identified. Click on this graphic to view the CDD search results page.

In the Module 2 Worksheet, record the Accession number, Name, E-value, Description and Interval of the top two hits that begin with the prefix "COG" (or at least do not begin with "Pfam" or "TIGRFAM") from the *List of Domain Hits* section. Enter "N/A" if no hits other than Pfams or TIGRFAMs are obtained.

Note: You will learn how to search the Pfam and TIGRFAM databases manually next in this module, which is why we are omitting them here. CDD queries many large databases for protein models (database sources listed below and indicated by the beginning of the accession number).

Accession starts with:	Source Database
cd	Curated at NCBI
pfam	Pfam
smart	SMART
COG	COGs
KOG	KOGs (available as a separate search set via CD-Search (RPS-BLAST); not searchable by text term in Entrez)
PRK	PRotein K(c)lusters (Entrez database)
CHL	Chloroplast and organelle proteins; subset of the PRK database.
MTH	Mitochondrial proteins; subset of the PRK database.
PHA	Phage proteins; subset of the PRK database.
PLN	Plant-specific (non-chloroplast) proteins; subset of the PRK database.
PTZ	Protozoan proteins; subset of the PRK database.
TIGR	TIGRFAM
LOAD_	Library of Ancient Domains (LOAD) data set. (available as a separate data set via FTP; not searchable by text term in Entrez)

Accessions that start with "**cl**" are for superfamily cluster records, which can contain domain models from one or more source databases.

* 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 2 Guide

* 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 2 Guide

TIGRFAM

TIGRFAM[2] is a database of protein families that features curated multiple sequence alignments, hidden Markov models (HMMs), and associated information designed to support the automated functional identification of proteins by sequence homology. In contrast to Pfams (which you will use in the next section), TIGRFAMs are often constructed from full-length protein sequences or well-conserved and functionally understood domains. Therefore, TIGRFAM results are very useful for predicting the name and/or function of the gene product.*

Go to TIGRFAM at http://tigrblast.tigr.org/web-hmm.

Paste the protein sequence in FASTA format into the search box and click "Start HMM Search". *Protein sequence in FASTA format can be downloaded from the Summary Tab on SGD from the Sequence dropdown box by selecting Protein.*

After searching the TIGRFAM database, raw text results will show which TIGRFAMs match. The name of the TIGRFAM hit ('**Description'**' column) may be cut off. If this is the case, identify the TIGRFAM number (e.g. **TIGR#####**) by the code found in the '**Model**' column and then go to the following page:

http://jcvi.org/cgi-bin/tigrfams/Listing.cgi

Search the database with the full number to find the entire TIGRFAM name.

This program may also pull Pfam hits in which case the Model designator will be PF##### instead of TIGR#####.

For each significant TIGRFAM hit, report the Model, Description, Score, and E-value in the **Module 2 Worksheet**. If a GO (Gene Ontology) number or an EC (Enzyme Commission) number is shown in the full description of the TIGRFAM entry, you should record these as well since they may prove very useful when attempting to predict the function of the gene product. Only record TIGRFAM hits (those starting with the prefix "TIGR") at this point. In the next section, you will search for Pfam hits.

Pfam

Pfam is a database of protein families, each represented by hidden Markov models generated from manually-curated multiple-sequence alignments of common protein families and domains. [Recall that domains are "modules" in proteins that usually have conserved tertiary structure and function]. Pfam can be used to identify likely protein domains within an amino acid query sequence. For each domain identified, Pfam can provide a great deal of information (if it is available) pertaining to the domain function, sequence conservation, and critical residues. A protein sequence may contain multiple Pfam domains, but these will never overlap.*

Navigate to the Pfam Search page at http://pfam.xfam.org/ Click the SEQUENCE SEARCH link, enter your ORFan's amino acid sequence and click "Go".

* 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 2 Guide

When your Sequence search results load you will be presented with a graphical overview at the top and below that a table of Significant Pfam-A Matches. The graphic at the top has a horizontal gray cylinder representing the entire polypeptide for your ORFan with any Pfam hits as colored bubbled regions along that gray cylinder. [*You can right click and open this graphic then copy it if you would like to add this to your **Module 2 Worksheet**.*] The table below then lists out the Pfam Family of the domain hit, its Description, the alignment of start and end amino acid locations for that domain, whether or not the entire domain is found within your protein (HMM From and To regions), and an E-value. Use this results page and its data and hyperlinks to work through the sections below.

Predicted Domains

Identify Pfam domains that are found within the protein sequence.

Note the E-value for the hit (match between the query and a database sequence). Why might it be useful to examine a hit even though it has a relatively high E-value?

Record the Pfam name (Description), Score, and E-value in the **Module 2 Worksheet**. If there are any amino acids listed under "Predicted active sites", record these in the **Module 2 Worksheet** as well.

In some cases, there might be more than one Pfam hit for the query sequence. In this case, be sure to record the relevant information as above, and consider why the different domains might exist in the same protein. This could help in determining the identity or function of the gene product.

Is the whole Pfam hit (listed as an HMM) covered by the alignment between your protein and the HMM? If not, the text in either the 'From' or 'To' field under "HMM" will be highlighted in **burgundy**. If a large portion of the Pfam domain is missing due to truncation, it is possible that the domain in your protein may not fold in the same way or perform the same function as it does in other family members.

Pairwise Alignment

To examine the alignment between the query sequence and the Pfam HMM to determine how similar what is present in your protein is to the consensus domain, click the "Show" button under "Show/hide alignment". Record this alignment in the **Module 2 Worksheet**. [*Note: You may need to screenshot or snip this image (SnippingTool or Grab) and insert it separately if you are having trouble copying and pasting it into the worksheet.*]

The first row in the alignment shows the consensus sequence for the HMM. In this sequence, capital letters correspond to highly conserved amino acids in the alignment used to generate the HMM. The second row lists residues in your query sequence that are identical to those in the consensus. You should record the identities and positions of these residues in the **Module 2 Worksheet**.

Be sure to comment on details such as the alignment and E-value and how this information contributes to your annotation.

* 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 2 Guide

Domain Summary

Now look at the **Domain Summary** page for the domain that your query sequence matched by clicking on the hyperlink under the Pfam Family.

On the resulting page, copy the Pfam Family descriptor and number (formatted as PF#####) from the top of the page and the Clan name and number (if available)(CL####) from the Clan tab into your **Module 2 Worksheet**.

Read any text on the Summary and Clan pages carefully. Since each Pfam domain has been manually curated, this information can be extremely useful in predicting the function of a query gene product containing a match to the domain. Make notes within your worksheet on information from these summaries.

If a GO or EC number is shown, record that number in the **Module 2 Worksheet** as it will aid in predicting the function of the gene product.

HMM Logo

On the left-hand menu, click "HMM logo". This tool is for visualizing amino acid conservation among the sequences used to build the Pfam domain. HMM Logos provide the researcher with a quick overview of the features of a Pfam HMM while conserving as much information as possible. The larger a letter is in an HMM Logo, the more conserved this residue is in the protein family. Colors correspond to different amino acid types (e.g. neutral, acidic, etc.). Letters are sorted in descending order depending on their probability of occurring at a given position in a sequence that contains the domain.

Save the HMM logo as a .PNG file. Go back to the **Module 2 Worksheet** and upload the file. [*Note: You may need to screenshot or snip this image (SnippingTool or Grab) and insert it separately if you are having trouble copying and pasting it into the worksheet.*]

Curated Alignment

To find the active site residues, we can look at the Curated Alignment (HMM logos do not identify active site residues).

On the left-hand menu, click "Alignments". Under View options, on the HTML line, click on the check mark under Seed to land on the Curated Alignment page. [*Seed represents a limited set of key sequences used for the consensus of this domain, you can view all alignments by clicking the check mark under Full.*]

A new window will open with the See Sequence Alignment for your Domain. The conserved residues are highlighted in various colors. The key for the color indication is to the bottom of the page. Active site residues are highlighted in black. By comparing information from the pairwise alignment, the HMM logo, and the curated alignment, you should be able to identify any key functional residues for your protein.

Report the key functional or structural amino acid residues in the **Module 2 Worksheet**.

* 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 2 Guide

Protein Data Bank (PDB)

The Protein Data Bank is the single worldwide depository of information about the three-dimensional structures of large biological molecules, including proteins and nucleic acids. A variety of information associated with each structure is available through the PDB including sequence details, atomic coordinates, crystallization conditions, 3-D structure neighbors computed using various methods, derived geometric data, structure factors, 3-D images and a variety of links to other resources.*

Copy your FASTA format protein sequence and go to:
http://www.rcsb.org/pdb/search/advSearch.do?st=SequenceQuery

You will land on a Sequence Search page. Copy in the amino acid sequence of your query gene. The default cut-off E-value is 10; change this to 0.01.

Click the "Submit Query" button and review the results. This runs a BLAST search just like you did in NCBI BLAST in the Sequence-based Similarity module. In this case, however, the query sequence is searched against all of the protein sequences that have solved structures in PDB.

Each hit will show up with a picture of the solved matching structure, Domain name information which will link to the full PDB page for that structure, an alignment depiction, and Length, E-value and Identity information. Examine the quality of the alignments between the query and the BLAST hits in the Protein Data Bank. If the E-value meets the cutoff of 0.01 and a significant length of the protein is aligned, the match is considered a good one. When two proteins are very similar in amino acid sequence and have approximately the same length, it is highly probable that they fold in a similar manner. Therefore, the structure corresponding to the PDB BLAST hit predicts how the query gene product is likely to fold.

If your query results in one or more good matches, record the PDB Code, Name, Length, Identities, and E-value in the **Module 2 Worksheet**.

Copy and paste the Alignment into the **Module 2 Worksheet**. [*Note: You may need to screenshot or snip this image (SnippingTool or Grab) and insert it separately if you are having trouble copying and pasting it into the worksheet.*]

If there is a literature reference associated with the protein structure, it may be beneficial to read this (or at least attempt to). When a structure is published, the authors will frequently characterize the function of the protein and identify important residues within its amino acid sequence. If these residues are conserved in your gene product, this helps to confirm its identity and function.

Don't be concerned if there are no significant hits with your query sequence since not all proteins are included in the PDB database.

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