Yeast ORFan Gene Project: Module 3 Guide

Structure-Based Evidence (Part 2)

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.

In Part 1 of the Structure-Based Evidence module we checked 4 (CDD, TIGRFAM, PFAM and PDB) of the largest structural information databases. There are many more databases in existence and they all use different algorithms to predict and detect structure regions of proteins. Therefore, in this module we will check additional structure detection programs for potential domains, you may get many more results, duplicate results, or no hits, all of which are informative.

SUPERFAMILY

"***SUPERFAMILY is a database of structural and functional annotation for all proteins and genomes. The SUPERFAMILY annotation is based on a collection of hidden Markov models, which represent structural protein domains at the SCOP (Structural Classification of Proteins) superfamily level. A superfamily groups together domains, which have an evolutionary relationship. The annotation is produced by scanning protein sequences from over 2,414 completely sequenced genomes against the hidden Markov models." "Nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification***.*"

For each protein sequence/gene name you run through the Superfamily database the algorithm will output any SCOP classifications it can find. You can also look in greater detail at the domain organization and any sequence alignments between the consensus sequences of identified domains.

Copy your predicted protein sequence, in FASTA format, from SGD or a previous worksheet and paste it into the box that appears when you click ***Sequence search*** at http://supfam.org/ , then click the Submit button. Once your search has completed you will need to click the "**View the domain assignment results**" hyperlink on the page that has opened. You will open a new page that will contain the results of your search. *[Note we have created a blank in the worksheet for 1 domain, if you have more than 1 domain please simply copy and paste and make additional sections as needed]*

The new page should show a data table with information about any identified domains. Within this table is information about the name of the Superfamily, the Family name, and their respective E-values. Copy this information into your **Module 3 Worksheet**.

Click on the hyperlinked Superfamily/Family domain name to learn more your particular domain. A new window will open with a section titled **SCOP Classification**. Subsections of this field will differ from domain to domain, so please copy the information here into your **Module 3 Worksheet** (there will not be subsections labeled since we cannot predict what they might be).

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

Scroll down the rest of the page and make note of additional information that might be of interest, if you are not sure what something is feel free to ask your professor or take a screen shot and send an email. We are unable to predict what information might exist in order to give a comprehensive list of what it all might mean. Copy and make any notes about this data in your Worksheet.

When you have finished with this page, go back to the other tab/window that you were originally directed to (it says "YourInputSequence" at the top). Under YourInputSequence is a cartoon drawing of the domain(s) the program has identified (usually a colored rectangle drawn on a line). If you click on this picture you can see a phylogenetic distribution of genomes that have proteins with the same domain architecture as the one found in your protein. Scroll through the list and comment in the **Module 3 Worksheet** as to the abundance of your domain, that is, in what organisms is it very abundant, not very abundant, is it only found in eukaryotes, etc.

SMART

"SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. More than 500 domain families found in signaling, extracellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. Each domain found in a non-redundant protein database as well as search parameters and taxonomic information are stored in a relational database system. User interfaces to this database allow searches for proteins containing specific combinations of domains in defined taxa.*"

[*Note: If the database allows you to choose between Normal SMART and Genomic SMART modes, select Normal*.] Copy your predicted protein sequence, in FASTA format, from SGD or a previous worksheet and paste it into the Protein Sequence box at http://smart.embl-heidelberg.de/ click the **Normal Mode SMART** button.

This may direct you to a page to select your sequence based on your ORFan name (go ahead and chose yours) or may directly link you to a new page, which loads a simplistic representation of your protein at the top as a gray rectangle. Any domains of interest are represented as colored geometric shapes along your gray protein sequence. A table towards the bottom of this page will give you more information about any domain regions. This information starts with the sentence "The SMART diagram above represents a summary….." Please copy the information about any domains found, including name, beginning and ending sites and information gained from clicking on the name of the domain into your **Module 3 Worksheet.**

GENE3D

* Information in quotations 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 3 Guide

"Gene3D takes CATH domains (from PDB structures) and assigns them to the millions of protein sequences (using Hidden markov models) with no PDB structures. Assigning a CATH superfamily to a region of a protein sequence gives information on the gross 3D structure of that region of the protein. CATH superfamilies have a limited set of functions and so the domain assignment provides some functional insights. Furthermore, most proteins have several different domains in a specific order, and so looking for proteins with a similar domain organization provides further functional insights.

More recently we have subdivided (the sometimes large and functionally diverse) CATH superfamilies into functionally coherent functional families (FunFams) and the majority of CATH superfamily assignments to protein sequences have a FunFam assigned. There are many other uses of domain assignments for example helping to interpret protein interaction networks, or in genome comparison.*"

Go to the CATH Search page at http://www.cathdb.info/search/by_sequence. Enter your protein FASTA sequence ***with header*** in the open box on the **Search by Sequence** tab and hit **Search**.

The main page should change to show a Progress section at the bottom with a green line indicating if the current search is Waiting, Queued, Running or Done. Once complete a green button with show up indicating how many matches were found. You may get multiple green buttons representing searches in different sets, for instance the CATH Structural Domain set and the CATH Functional Families set. Click on the green button for the CATH Structural Domains, when the new page loads a data table should be present with the Summary information about the domains found in your protein. For matches with significant e-values <0.001, record information on the match name, region and e-value in **Worksheet 3**, then click on the match name to learn more about this region and take notes in your worksheet.

Next click to the Matching FunFams Tab. Again, for matches with significant e-values <0.001, record information on the match name, region and e-value in **Worksheet 3**, then click on the match name to learn more about this region and take notes in your worksheet.

PANTHER

* Information in quotations 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 3 Guide

“***The PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System is a unique resource that classifies genes by their functions, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence. Proteins are classified by expert biologists according to: 1) Gene families and subfamilies, including annotated phylogenetic trees 2) Gene Ontology classes: molecular function, biological process, cellular component 3) PANTHER Protein Classes 4) Pathways, including diagrams. PANTHER is part of the Gene Ontology Reference Genome Project. PANTHER is supported by a research grant from the National Institute of General Medical Sciences [grant GM081084] and maintained by the Thomas lab at the University of Southern California.****”

Navigate to http://www.pantherdb.org/ and click the Sequence Search tab near the top. Copy your predicted protein sequence, in FASTA format, from SGD or a previous worksheet, and paste it into the ‘Enter a protein sequence box’ and click the **Submit** button.

If a domain is found a page that reads: PANTHER HMM SEQUENCE SCORING RESULTS will open and the PANTHER Hit, E-value and Alignment will be displayed. Copy this information into your **Module 3 Worksheet.** [Formatting can be disrupted by simply copying the alignment and pasting it in, to remedy this problem after copying the sequence, right click, select Paste Special and click Unformatted Text. If the alignment is not fixed you can take a screenshot.]

Next click on the hyperlink of the PANTHER Hit Named Domain and read through the information on this page. Record the Family Name, The PANTHER protein class(es) and make note of useful information in the Gene Ontology (GO) sections. [Gene Ontologies, or groupings, categorize genes based on molecular function, biological process or cellular component, these groups are then applicable to all members of the group].

* Information in quotations regarding the functionality of each software/algorithm and operational information was taken directly from the websites and included here for instructional purposes only. [rev 6.6.2022 update superfamily link]