Yeast ORFan Gene Project: Module 1 Guide

Introduction to *Saccharomyces cerevisiae*

Saccharomyces cerevisiae are common budding yeast used in food and drink production as well as scientific laboratories around the world. This yeast was the first eukaryotic organism to have its genome fully sequenced in 1996, a factor that plays an important role for this organism as a powerhouse in genetics research. As with any sequence project of any organism, while this data tells us much about potential gene locations and features of the genome, it cannot accurately identify all protein producing open reading frames (ORFs) nor can it tell us anything about the function of the protein that might be produced. Approximately 6400 ORFs have been identified within the >12 million base pairs of the *S. cerevisiae* genome, however many hundreds of these are still identified only as "putative protein of unknown function". For this project, each student is assigned one of these putative protein ORFs and will be tasked with performing bioinformatics analysis of the ORFan/Gene of Unknown Function (GUF) to attempt to gain more knowledge about the function of its gene product. These modules are meant to guide you through the steps of gathering biological information about your assigned gene. You will be responsible for keeping track of all of this information and reporting back your findings in a neat and organized way via your lab notebook.

Saccharomyces Genome Database

The sequence information for the 1996 project has been annotated and stored in a publicly available database called the Saccharomyces Genome Database or SGD and can be found at www.yeastgenome.org. This website should serve as a jumping off point for much of your data collection throughout this project. On the **SGD Homepage** in the **search bar** section in the upper right, type in the **name of your gene** and hit **enter.**

Basic Information

Description

You should now be looking at your GUF's **SGD Summary** page on SGD. You will start in the LOCUS OVERVIEW section and be able to scroll down through all of the tabs for more information or click between tabs at the top of the page.

Any basic information known about your GUF will be summarized in this first Locus Overview section. Please copy down and make note of this data. Note if there is a citation (a superscript # included in the sentence) you can link to that reference and read about how they found out this information.

Record this information in your **Module 1 Worksheet** in the space provided.

DNA Coordinates

DNA coordinates define the location of a particular sequence in the genome. Numbering of coordinates is based on the order of nucleotides within a sequencing scaffold[1].

[1]Here, the term *scaffold* refers to a set of partial genomic sequences in which the individual sequences are known to be in the correct order but not necessarily connected directly to one another.

* 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 1 Guide

As you scroll down your gene page into the SEQUENCE section, note the DNA Coordinates (shown in the format **####### to #######**) and what chromosome the gene is found on. Calculate the size of your gene.

DNA Sequence

You will be using the DNA nucleotide sequence as a query for some bioinformatic tools, so it is helpful to record this for future reference.

In the **SEQUENCE** section, using the grey Download (.fsa) box, select Genomic DNA from the drop down and a new window should pop-up prompting you to open or save your sequence file. **[***This file can be opened with a Text Editor (Notebook or TextEdit), Word or a Sequence Analysis program (such as SnapGene Viewer for which there is a free downloadable version.)]*

Please copy this DNA Sequence and paste it into your **Module 1 Worksheet** in the space provided.

Note the DNA coordinates and chromosome information are listed in the sequence file as well, double check you have everything entered correctly in your worksheet.

Protein Sequence

The protein (amino acid) sequence predicted from a gene is the most common query used for searching bioinformatic databases.

In the **SEQUENCE** section, using the grey Download (.fsa) box, select Protein from the drop down and a new window should pop-up prompting you to open or save your sequence file. **[***This file can be opened with a Text Editor (Notebook or TextEdit) or Word or a Sequence Analysis program (such as SnapGene Viewer for which there is a free downloadable version.)]*

Please copy this Protein Sequence, without the header line, and paste it into your **Module 1 Worksheet** in the space provided.

On the **SGD Summary** page, now scroll down to the **PROTEIN** section.

Record the Length (a.a.)

Does this length make sense with your calculated DNA sequence?

Record the Molecular Weight (Da)

Record the Isoelectric Point (pI): ***The isoelectric point, or pI, is the pH at which a protein carries no net charge. While the pI of an individual protein might not provide much useful information, the variation in abundance of acidic and basic proteins can be correlated with taxonomy, cellular localization, ecological niche of an organism, and proteome size (the proteome of a cell, tissue, or organism is the complete set of proteins made at a given time under defined conditions).***

[1]Here, the term *scaffold* refers to a set of partial genomic sequences in which the individual sequences are known to be in the correct order but not necessarily connected directly to one another.

* 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 1 Guide

Sequence-Based Similarity

Some terms that you need to know before starting this module*:

Homolog – (1) a sequence that shares a common ancestor with another (2) a gene related to a second gene by descent from a common ancestral DNA sequence. The term "homolog" may apply to genes separated by speciation (see **ortholog**) or by genetic duplication within a species (see **paralog**).

Ortholog – (1) a sequence that shares a common ancestor with another but evolved independently because of a speciation event (2) a gene in one of two or more different species that evolved from a common ancestral gene. Normally, the products of orthologs retain the same functions in the course of evolution.

Paralog – (1) a sequence that is similar to another because both are descendants of a duplicated ancestral gene (2) a gene related to another by duplication within a genome. Paralogs often evolve new functions, even if these are related to the original one.

Domain – distinct modular region of a protein that serves a particular function, such as DNA-binding.

Family – a group consisting of proteins that are more than 25% identical in amino acid sequence across their entire length, share structural features and often have related functions.

Superfamily/Clan – group of protein families that are related by detectable levels of sequence similarity reflective of an ancient evolutionary relationship (accession numbers begin with "cl").

Basic Local Alignment Search Tool (BLAST)

BLAST finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to those in specified databases and calculates the extent and statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences and to help identify members of gene families. An important aspect of this type of pairwise alignment (comparison of two sequences) is evaluation of the quality of matches or "hits". Pay attention to the length of any match, its numerical score and its E-value[2]. A BLAST hit may have a good E-value, for example, but if it is based on alignment with only a small part of the query sequence, the results should be interpreted with caution.*

[1]Here, the term *scaffold* refers to a set of partial genomic sequences in which the individual sequences are known to be in the correct order but not necessarily connected directly to one another.

* 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 1 Guide

Switch from the **SGD Summary** Tab to the **SGD Protein** Tab by clicking on the **PROTEIN** tab at the **TOP** of the page. Scroll down to the **RESOURCES** section at the bottom of the page and click on **BLASTP at NCBI**. Your query sequence identifier should be automatically entered into the Enter Query Sequence box, check that there is a NP_###### text in this location. [If this does not work, copy the protein sequence from SGD and paste it into the query sequence box.]

Select a database to search against from the **Database** dropdown menu. **NR** (nonredundant) is a massive repository of protein sequences that are largely predicted from sequenced genomes. The vast majority of the sequences in NR have never been manually annotated, so while it is more likely your query sequence will closely match a sequence in NR than in another database, the predicted gene product identity may or may not be reliable. **UniProtKB/Swiss-Prot** is a much smaller sequence database that contains only curated (manually annotated) sequences. It is less likely that your sequence will match an entry in SwissProt than in NR, but if it does you can have greater confidence in the predicted identity. It is good practice to run a BLAST search against each database and note any differences in the results.

With **NR** selected, click the "BLAST" button to search for protein sequence matches. You will be directed to the BLAST Results page. [*Note: this page can take a little while to load based on website traffic and you may predict slower times if working on this all at once in a large class.*]

Scroll down to the Descriptions section to the portion labeled **Sequences producing significant alignments**. **Look for the best hit that is NOT from *Saccharomyces cerevisiae*.** Note that the organism may still belong to the genus *Saccharomyces*, as long as it is a different species. Click the hyperlink for in the "Description" column to view the alignment.

Determine if this hit meets our E-value cutoff: it should be less than E-03 ($1x10^{-3}$ or 0.001). If the E-value is satisfactory, note the identity of the organism and the gene product name. Remember that in bioinformatics, a lower E-value indicates a lower probability that the match observed is due to random chance rather than an evolutionary relationship.

For the top-scoring match (in an organism other than *S. cerevisiae*), record the gene product name, organism name, alignment length (equals the number of the last query sequence **residue**[3] aligned minus the number of the first residue aligned, plus 1), score, and E-value in the **Module 1 Worksheet.** Copy and paste the alignment of the query and top-scoring BLAST hit into the **Module 1 Worksheet.** Comment on the E-value and compare the lengths of the query and subject (matching) sequences.

Go back to your starting BLASTp page and change your database to **UniProtKB/Swiss-Prot(swissprot)**, click the "BLAST" button to search for protein sequence matches in this alternate database. You will be directed to the BLAST Results page.

Perform all actions and data copying as described above for the NR database results and put them into the **UniProtKB/Swiss-Prot** section of your **Module 1 Worksheet.**

[1]Here, the term *scaffold* refers to a set of partial genomic sequences in which the individual sequences are known to be in the correct order but not necessarily connected directly to one another.

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