Guide for Bioinformatics Project Module 1

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 sequencing 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 a "putative protein of unknown function." For this project, each of you will be assigned one of these putative protein ORFs and will be tasked with performing both hands-on "wet-bench" experiments as well as bioinformatics analysis of the ORF to attempt to gain more knowledge about the function of this gene product. These guides are meant to lead you through the steps of gathering biological information about your assigned gene. While it is possible to complete them wherever there is an internet connection, **most students find it valuable to complete during class**, when time is available, so that they can consult with classmates and/or their instructor. You will be responsible for keeping track of all of this information in the worksheets and submitting them via Blackboard at two time points during the semester.

Saccharomyces Genome Database

The sequence information from the 1996 project has been automatically annotated at a basic level and then 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.

Basic Information

Description

On the **SGD Homepage** in the **search bar** section in the upper right corner, type in the **name of your gene**. Any information known about your ORF will be summarized on the **Summary Tab** also known as your **Gene Homepage** that is now displayed. Please copy down the information in the Feature Type and Description sections in your Module#1 Worksheet. Note if there is a citation (a superscript # included in the sentence) you can link to that reference and read about how this information was discovered.

Record the Feature Type and Description information. (We recommend working within the Worksheet for each given module, but that is not feasible write this information out and/or store it in a Word or Google document that you save in your SkyDrive, email to yourself, or keep on a thumb drive.)

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].

From your **Gene Homepage**, **record the DNA Coordinates** (shown in the format **####### to #######**) in the Chromosomal Location portion and **what chromosome the gene is found on**. **Calculate the size of your gene.**

[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.

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.

On the **SGD Summary** page scroll down to the **SEQUENCE INFORMATION** section, next to Retrieve sequences make sure S.c. Reference Strain S288C is selected in the 1st dropdown box and ORF Genomic DNA in the 2nd and hit **Get Sequence.**

You will be taken to a new page where the sequence information is given. **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 on this page 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.

On the **SGD Summary** page, at the Top click on the **Protein Tab**:

From the **PREDICTED SEQUENCE** 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).***

Next to the **Formatted Sequence** Hyperlink in the **Predicted Sequence** section is a button to download the FASTA formatted version of your sequence, this is the unnumbered version you will need to enter in future runs. **Click the Download button and copy this into your Module 1 Worksheet in the space provided** (leave off the asterisk if there is one)**.**

Sequence-Based Similarity

A starting point for predicting function of your gene encoded protein of interest is to see if there is anything known about what a similar protein might be doing in another organism. Do perform this comparison and test significance we will start by utilizing a program called BLAST (Basic Local Alignment Search Tool) that will find attempt to find entire proteins with similar sequences in other organisms. We will complement this search by then using the CDD or Conserved Domain Database to look and see if there are regions or parts of your protein, known as domains because they fold and sometimes function independently of the rest of the protein, that are conserved in other proteins and/or organisms.

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**).

Guide for Bioinformatics Project Module 1

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 50% identical in amino acid sequence across their entire length.

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 software finds regions of local similarity between sequences. The program compares nucleotide or protein sequences you provide 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.

While on the Protein Tab for your gene, scroll down to the **EXTERNAL LINKS FOR** section at the bottom of the page and click on **BLASTP (NCBI)**. Your query sequence identifier should be automatically entered into the Query Sequence box, check that there is a NP_###### text in this location.

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. **SwissProt** 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.

Click the "BLAST" button to search for the best protein sequence match. After a small wait while the search is completed you will land on the BLAST Results page. Scroll down to the section labeled **Sequences producing significant alignments** (under Descriptions). **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 the Description column to view the alignment.

Note: when you run the SwissProt BLAST, the format of the results is different and the organism information is not included in the Description column. To access this data you must click the Accession Column Hyperlink, this will redirect you to the NCBI entry for this piece of data and you can find SOURCE ORGANISM on this page that tells you where this particular protein is from.

Determine if this hit meets our E-value (Expect) 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-

[2]BLAST scores are based on the number of matches between two sequences, the length of the aligned region, and the number of gaps that have to be added to maintain the alignment. The E-value is the probability that an equivalent score could be obtained by comparing the query to a set of randomly-generated sequences.

value indicates a lower probability that the match observed is due to random chance rather than an evolutionary relationship.

For the top-scoring match, 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**.

While you are reviewing BLAST results, check to see if there are any Conserved Domain Database (CDD) hits (the graphical view near the top of the Results page). See more information regarding CDDs below.

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 (which we will investigate in module 3).

This search is automatically run in parallel with any NCBI BLAST search. After performing a BLAST search you will see a graphical representation of putative conserved domains (superfamilies, COGS, Pfams, etc.), if any have been found, at the top of the Results page (under Graphic Summary). Click on this graphic to view the CDD search results page.

In the Module 1 Worksheet, record the number, name, E-value and Bit Score of the top two hits that begin with the prefix "COG" (or at least do not begin with "Pfam" or "TIGRFAM"). 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 in the Structure-based Evidence 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).

Abbreviation	Database Name	Description
SMART	**S**imple **M**odular **A**rchitecture **R**esearch **T**ool	SMART is a web tool for the identification and annotation of protein domains, and provides a platform for the comparative study of complex domain architectures in genes and proteins. SMART is maintained by Chris Ponting, Peer Bork and colleagues, mainly at the EMBL Heidelberg. CDD contains a large fraction of the SMART collection.
Pfam	**P**rotein **fam**ilies	Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. Pfam is maintained by Alex Bateman and colleagues, mainly at the Wellcome Trust Sanger Institute. CDD contains a large fraction of the Pfam collection.
COGs	**C**lusters of **O**rthologous **G**roups of proteins	COGs is an NCBI-curated protein classification resource. Sequence alignments corresponding to COGs are created automatically from constituent sequences and have not been validated manually when imported into CDD.
TIGRFAM	**T**he **I**nstitute for **G**enomic **R**esearch's database of **p**rotein **fam**ilies	TIGRFAM, a research project of the J. Craig Venter Institute, is a collection of manually curated protein families from The Institute for Genomic Research and consists of hidden Markov models (HMMs), multiple sequence alignments, Gene Ontology (GO) terminology, cross-references to related models in TIGRFAM and other databases, and pointers to literature.
PRK	**PR**otein **K**(c)lusters	Protein Clusters is an NCBI collection of related protein sequences (clusters) consisting of Reference Sequence proteins encoded by complete prokaryotic and chloroplast plasmids and genomes. It includes both curated and non-curated (automatically generated) clusters.

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)

[3]Proteins consist of one or more polypeptides which are composed of several amino acids. When the amino acids form a peptide bond, the elements of water are removed, and what remains of each amino acid is called an amino acid residue.