

## BIOL 362 – Cellular and Molecular Biology – Bioinformatics Project

To accompany your molecular biology laboratory experience, you will be performing several tasks to gain some initial experience in the field of bioinformatics. This is designed to introduce you to the extensive applications that are available in this emerging field of Biology, Chemistry, and Computer Science.

As you have been informed during your laboratory project, you will be creating random mutations in the genome of *Candida albicans* in order to associate a particular gene with a specific phenotype. Since this project will take the entire semester to complete, I will be providing you a DNA sequence as if you had acquired at the end of the project. Separately, you will select an output sequence and perform the tasks listed below. The information below is designed to help you navigate through the San Diego Supercomputer Center workbench for bioinformatics, however you will need to troubleshoot at some points throughout the process. You will be working in your groups, so I expect you to work through these challenges together before coming to me for assistance. If you continue to struggle, I will be glad to point you in the right direction. As a group, you will need to work together. Remember that there will be an anonymous peer grading at the completion of this project.

1. Log on to the San Diego Supercomputer at <http://workbench.sdsc.edu>. Click on “Click to enter the biology workbench 3.2”.
2. You will need a user ID and password that your instructor will provide to each group.
3. Select Nucleic Acid tools.
4. Add your sequence by going to “add new sequence”, enter a label for your sequence, such as unknown Tn7 mutagenesis, enter your sequence, click “update”, then selecting “save”.
5. Select your sequence and run BLASTN.
6. Select the refseq\_fungi database and click submit...this will focus your search to fungal genomes since you are attempting to identify the unknown sequence of DNA from *Candida albicans*.
7. Look for the *C. albicans* sequence with the best e value, select, and click on “show record”.
8. Identify various key features of this piece of DNA, such as the name of the gene; the abbreviation for the gene; whether it is from a mRNA, CDS, or partial sequence; length of sequence in database (if complete sequence, identify length of coding region); and miscellaneous features. Be sure to also note the accession number.
9. To go back, select import sequence and return.
10. Select the imported sequence from the database and run BLASTX.
11. Select SwissProt.
12. Choose 10 protein sequences that matched the search...in addition, be sure to select the \_CANAL sequence, which should have an e value of 0.0). Avoid selecting more than one sequence from the same organism. You do not have to pick 10 with the best e value...you may select whichever 10 you wish. The higher the e value, the more

obscure the sequence will be, so you will want to use some sense when choosing the sequences.

13. Import these 11 sequences.
14. Select the *Candida albicans* protein sequence and view database records of imported sequence. Click run and show record.
15. Identify key characteristics of this protein sequence: number of amino acids; function; pathway in metabolic reaction; subunits; location; features table; amino acids involved in function, etc.
16. Return and select the *Candida* protein sequence and run RPSBLAST for domain search. Click run. Leave the defaults and select submit.
17. Identify conserved domains with functions. What is the percent identity and positives? What do you think is meant by a + rather than a match?
18. Return and select all the protein sequences (all 11), ClustalW, and run. Keep the defaults.
19. Copy and paste the Phylip unrooted guide tree into a word document.
20. Import the alignment.
21. Select the alignment and run Texshade with the defaults.
22. Copy and paste each page into your word document.
23. Pay particular note to the consensus sequence for your alignment. What guidelines are used in the designations of this consensus sequence? Identify any regions that appear highly conserved. What connection can be made to the features table from your database records of the *Candida* protein sequence?
24. Return and select the alignment. Run DrawGram. Keep default settings.
25. Copy and paste in a word document.
26. Return and select your alignment. Run ClustalDist and keep defaults. Copy and paste the distance matrix with your phylogenetic tree. Analyze the distances that were used to create the tree.

From here, you must begin to search information outside the workbench website.

1. Research the function of the gene/protein based on these 11 sequences you have extracted and analyzed. You can search websites, as long as you are aware of the origin of the information. Be very suspicious of information you find on random webpages.
2. Research the organisms that you have used in your phylogeny. If you are unsure of the organism, you can use the workbench site to identify the organism by selecting that particular sequence and looking at the database record. If you have already aligned your sequences, you can split the alignment into its component sequences.
3. You will be responsible for preparing a presentation to the group at the end of the semester. Separate instructions will come later in the semester as to the format of these presentations. For now, begin your work on the SDSC site and research of the function and organisms. This information will be used in your presentations.