

Laboratory Worksheet, Monday, Nov. 28.

I. PHX Data. This week we will scrap the previously gathered data for some more systematically cared for data. Dr. de Wet found an alternative pathway with little hand curation required: be sure to thank him for this, since it is your hand-curation he spared you of! Basically, we will go to The Institute for Genomic Research (TIGR) where they have well curated genomic data for reference species. The gens will be readily found, but there will be some minimal curation required: you will search for, e.g., “chaperones”, and will receive a managable number of genes with chaperone in the annotation. There are relatively few misleading ones, unlike Genbank, but there are some, and each of you will have to go over one family to remove the related (upstream or downstream in a pathway, e.g.) genes which come up, i.e., you will have to remove several not-quite-chaperones from the list provided. We will save this to a multi-FastA file, entitled *E.coli chaperones*, e.g. Here are Jeff’s instructions for using the TIGR resources:

Here are four numbered steps that guide you through the process of selecting sets of genes by a keyword search of the annotations.

Step 1. Enter a search term (chaperone, ribosomal protein, or transcriptional) in the text field.

Step 2. Select a genome from the list of available genome and add to the search list by clicking the add button. We will use the *E. coli* K12-MG1655 genome

Step 3. Select which annotation sets to search. Use Primary and TIGR annotation (Default)

Step 4. Click the Submit button.

An Annotation Search Report page will be returned.

On the search report click the Select All button at the upper left of the report.

Next, go through the list of genes and unselect any genes that dont look like they belong with your group of genes.

To get the selected genes, click the options button at the upper right of the report. This will bring up a new window. First select the sequence type (nucleotide) and then click the download button.

You will get a new page with the sequences in it. These sequences are the open reading frames for the genes, and they include the stop codons. This page also has a download button that will give a new window with just the sequences in it. Right click on this page and you will get a page with the sequences as plain text that you can copy, paste into an editor (e.g. Nedit), and save as a file.

Family curation: Kim Davis - chaperones, Malini Sridharan – Ribosomal proteins, Xiaoxia Wang – transcription factors. Be sure to go to the Karlin-Mrazek paper to see what transcription factors means for them.

II. PHYLIP. PHYLIP is the name of the suite of phylogeny programs developed by

Joseph Felsenstein of the University of Washington. In lab we saw some of these programs mounted at the Institut Pasteur in Paris. Please check the “mothership” at: <http://evolution.genetics.washington.edu/phylip.html>. This is also now linked to the Web Resources page, as well as to the Lab Worksheets page where you found this Worksheet. The “exercise” for now is simply to read the documentation for the bootstrap and for DNAML before next Monday. We will find the relevant pages in the lab.

III. Perl for PHX. Time permitting we can get through the hash part of the exercise. Now that we have our data arranged, we can try to check whether the program sorts data properly.