CSC 315, Group Project
Cancer Biology

Analysis Requirements

Background: The *C. elegans* gene *cep-1* is related to *TP53*, a transcription factor and tumor suppressor gene in humans that regulates the cell cycle. Mutations in *TP53* can result in uncontrolled cell growth, which can result in cancer.

Objective: Identify candidate targets of cep-1, which are genes that have a TP53 binding consensus sequence in their promoters (up to 1000 bp upstream of the gene). Identify all *C. elegans* genes with at least 2 binding sites, and provide a table with the following information: the worm gene name, the number and locations of the binding sites, the human orthologs (if they exist), a summary of the human gene. (You may be asked to identify Gene Ontology (GO) terms at a later date.

The TP53 binding site has the consensus sequence:
G/A, G/A, C, A/T, A/T, G, T/C, T/C

Requirements:

1. Download the promoter regions of all *C. elegans* genes using the UCSC Table Viewer, and save to a file.
2. Write a python script that does the following:
    a. Reads in each sequence, and identifies all sequences containing at least 2 occurrences of the consensus sequence.
    b. For each sequence containing at least 2 occurences, outputs the following to a file:
        i. The transcript ID of the sequence
        ii. The number of candidate binding sites
        iii. The positions of the candidate binding sites
3. Using BioMart, find the orthologs for all transcripts that you have identified
4. Add the ortholog information to your output from (2), creating a new file.
5. Write a python script that reads in the file from (4), and creates a new file that adds the gene summary information for each ortholog. The summary information should be obtained from the Entrez Gene database.

Due Date: For each group, scripts must be submitted by Monday, April 17th. Additional submission requirements will be provided at a later date.