BMI 565/665 Bioinformatics Programming and Scripting
Research Project
Submit final project through Sakai submissions tool.

Send a short progress report by Monday of Week 9 to: mooneymi@ohsu.edu

There is a significant need to understand and treat respiratory virus infections such as H1N1 influenza, H5N1 influenza and SARS coronavirus. In this research project you will be studying a cell line host response to the H5N1 VN1203 strain of influenza, or the avian flu. You will be provided with a list of Agilent microarray probes that are differentially expressed up to 12 hours after infection. From this data you will identify a KEGG pathway that has been affected as a result of infection. Using Python, you will write a program to estimate if there is greater evolutionary conservation of the differentially expressed genes versus the non-differentially expressed genes in your pathway.

You will be provided with the following files:

1. H5N1_VN1203_DE_Probes.txt: contains a list of probes that are statistically differentially expressed after H5N1 infection
2. H5N1_VN1203_UNIVERSE_Probes.txt: contains all probes from the gene expression experiment.
3. KEGG_Pathway_Genes.txt: contains a list of KEGG pathways and gene members

Part I: Identify Pathway Focus

- Use Python to perform an odds ratio calculation to identify pathways containing a larger number of differentially expressed genes than would be expected by chance (this will be discussed in class)
- First identify a pathway with an odds ratio greater than 1.5 that is of interest to you, then confirm your choice with the instructor (mooneymi@ohsu.edu). Pathways can be viewed here: http://www.genome.jp/kegg/pathway.html

Part II: Examine Cross-Species Conservation of Pathway Genes

- Use Entrez eUtils efetch to download sequences pertaining to genes within your pathway for human, mouse, and one other species (your choice). Assume cross-species genes (orthologs) can be identified using the same gene symbol.
- Use clustalw to align these sequences and assess the similarity between all 3 sequences. One option is to use the edit distance between pairs of sequences. Remember to adjust for the length of the sequences (e.g. long genes are likely to have more differences than short genes).
- Create a boxplot comparing the similarity measures for genes that are differentially expressed versus genes that are not. Use a **Mann-Whitney U** test (**scipy.stats.mannwhitneyu**) to determine if the differences are significant.

Deliverables:

1. Write Part I and Part II as separate python programs that are both called from a linux bash script. This script should check for the existence of relevant files and check that Part I has successfully completed before running Part II. Turn in both python programs and the bash script. (60 points)
2. Screen shot of pathway with highlighted differentially expressed genes (10 points)
3. Boxplot labeled with p-value of **Mann-Whitney** statistic (10 points)

4. 1-2 page write-up summarizing your findings. This should be a word document that includes at least 2 figures corresponding to your pathway and boxplot. Discuss the limitations of the study and any obstacles/problems you experienced. Discuss the relevance of the pathway to H5N1 infection. Discuss why the conservation of the affected pathway might be important for the study of H5N1 (Hint: think about our use of model organisms to study infectious diseases). (20 points)