Final Research Project
BMI 565/665 Python Scripting

Eric Leung

December 8th, 2014

Part I: Identify Pathway Focus

The first part of the project was completed with the `odds.py` file. We were also given three files to analyze.

- `H5N1_VN1203_DE_Probes.txt`
- `H5N1_VN1203_UNIVERSE_Probes.txt`
- `KEGG_Pathway_Genes.txt`

Upon analysis, I found that there were six pathways that have a larger number of differentially expressed genes than would be expected by chance. This was determined with an odds ratio greater than 1.5.

	DE Genes	Non-DE Genes
Target Pathway Genes	A	B
Non-Pathway Genes	C	D

Table 1: A 2 × 2 contingency table of differential expression among target and non-target pathways used to calculate an odds ratio.

The odds ratio for each pathway was calculated with the following equation:

$$OR = \frac{\frac{A}{A+B} \Big/ \frac{B}{A+B}}{\frac{C}{C+D} \Big/ \frac{D}{C+D}} = \frac{A \times D}{B \times C} \tag{1}$$

The variables from the equation above reference those in Table 1. A summary of the significant pathways are in Table 2 below. The variables (A, B, C, D) represent the number of genes based on each pathway that was being focused on. Thus, these variables will change depending on which pathway is focused on at the time.

Pathway	Odds Ratio
ECM-receptor interaction	2.2812
RIG-I-like receptor signaling pathway	**2.2828**
Focal adhesion	1.5999
Cytokine-cytokine receptor interaction	1.5165
Small cell lung cancer	2.3641
Toll-like receptor signaling pathway	2.1792

Table 2: Significant pathways with odds ratios greater than 1.5

The pathway I chose is “RIG-I-like receptor pathway.” It has an odds ratio of 2.2828, which is greater than 1.5, which we deem significant enough to study. An unhighlighted Figure 1 can be found at `http://www.kegg.jp/kegg-bin/show_pathway?hsa04622`.

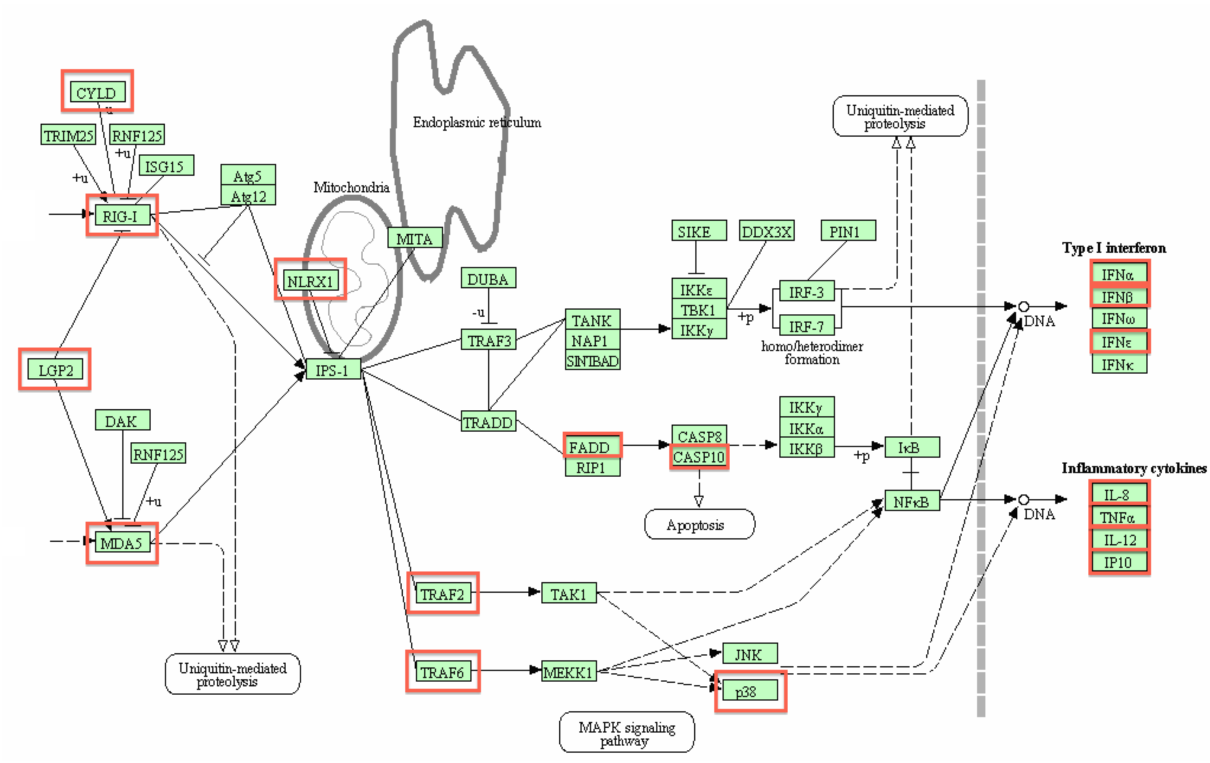

Figure 1: RIG-I-like receptor signaling pathway for Homo sapiens with differentially expressed genes highlighted.

Part II: Examine Cross-Species Conservation of Pathway Genes

I studied the conservation of genes among Humans (*Homo sapiens*), Mice (*Mus musculus*), and Chimpanzees (*Pan troglodytes*). Instead of taking genes from the "KEGG Pathway Genes" text file, I referenced the KEGG website [1]. I relied on the KEGG API to search for and parse the right page to extract the list of genes.

I disregarded the file given because I found some "genes" listed for my pathway did not correspond to any human gene. Two were cDNA clones (AW511634 and AI423557), a partial CDS for a transcription factor (U88316), and another was from HeLa cells (U56433). Thus, I went straight to the source (KEGG) for the list of genes to study. Additionally, I came across a duplicate gene, but under difference names. Gene CR936771 in the KEGG pathways file is actually SIKE1, which is already in my list, but it is listed as SIKE, but SIKE1 is the official gene name, according to HGNC.

The decision to go to the KEGG database for the list of genes to study resulted in eighteen genes from the "KEGG Pathway Genes" file not being found for my pathway. As it was mentioned earlier, I found that I wouldn't have been able to find legitimate genes for five "genes". Thus, I did not search for thirteen genes. This may have had an effect on the final analysis. However, as you will see in the end, these thirteen missing genes must have had radical differences in them to change the overall statistical test.

I wanted to only compare genes that existed among all three species. This makes

[1] http://www.kegg.jp/dbget-bin/www_bget?hsa04622

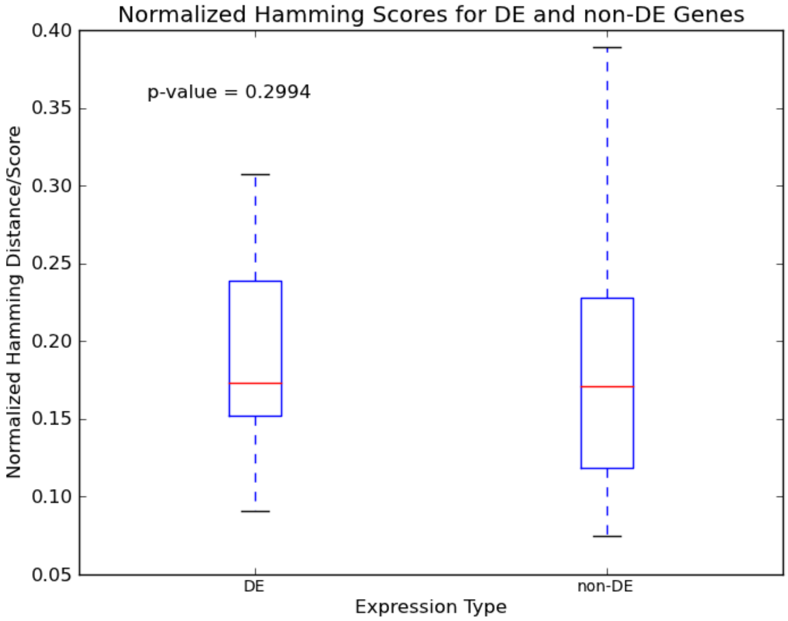

Figure 2: Boxplot of normalized Hamming distances for differentially expressed and non-differentially expressed genes in the RIG-I-like receptor pathway for Humans (*Homo sapiens*), Mice (*Mus musculus*), and Chimpanzees (*Pan troglodytes*).

analysis easier instead of having to make a separate analysis for a pair of genes and all three genes. I found there was no significant difference between differentially expressed genes and non-differentially expressed genes in terms of Hamming distance among three different species (p-value $= 0.2994 > 0.05$).

One obstacle I found was how to search for the correct gene sequence to retrieve. If I used the Entrez E-Utils and searched by gene name, I could only get close to the correct gene I wanted (a RefSeq mRNA sequence), but it would vary from each gene. I was not satisfied with retrieving any sequence (e.g. predicted or anything not RefSeq if possible). So I tried to find other, more precise ways of searching. This is another reason why I decided to reference the KEGG database because before each gene in a specified pathway is an ID that can be search precisely in Entrez to get the RefSeq database.

H5N1 Infection Reflection

The H5N1 infection is a type of bird flu. The pathway I chose (RIG-I-like receptor pathway) is "responsible for detecting viral pathogens and generating innate immune responses"[2]. This can be very important in responding to the onset of the flu virus, which supports the reason why this pathway's genes are differentially expressed.

Conservation of the affected pathway in model organisms is important when studying H5N1 because successful treatments on affected model organisms will have more confidence in translating the successful treatments to humans compared to distantly related model organisms.

[2] http://www.kegg.jp/dbget-bin/www_bget?hsa04622