Akavia Assingment 2

Since this was posted on Friday, it is due midnight of Friday, November 9, 2018.

This assignment will include the usage of structures (structs) to save space. The GO Annotations are loaded in an array of structs, which have the following fields
`Terms` - GO Annotations using the format GO:0097084
`Genes` - genes that include or not this annotation
`Matrix` - logical matrix, where 0 means the row/column combination is not annotated, 1 means it is annotated
`Filename` - a string which will describe what part of the ontology it includes (BP, CC, MF) and whether or not it includes parents

You can load everything you should need for this exercise with the files `Assignment3.mat`
The functions detailing how things were created are attached for references. The gene names were modified for everything to work, so don't load the gene names nor expression via previous code.

There is no assignment dealing with GSEA.

Your major tools are going to be the Hypergeometric distribution functions, including `hygepdf` and `hygecdf`.

These two functions accept x, M, K, N, which are equivalent to k, N, K, n in the presentation, and represent number annotated in the group, universe, number annotated in the universe, size of group.
Presentation

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Matlab

$$\frac{\binom{K}{x}\binom{M-K}{N-x}}{\binom{M}{N}}$$

1. First, figure out how to use hygepdf/hygecdf to get the pvalue we are looking for. If you got X annotated in the group (overlap between group and annotated genes)
   Hygepdf calculates the probability of getting X.

$$\frac{\binom{K}{x}\binom{M-K}{N-x}}{\binom{M}{N}}$$

   hygecdf calculates the probability of getting AT MOST X.

$$CDF = \sum_{i=0}^{X} \Pr(i|M,K,N) = \sum_{i=0}^{X} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}$$

   We need to get the probability of getting i AT LEAST as big as X. That's the p-value you are looking for.

$$\sum_{i=X}^{Inf} \Pr(i|M,K,N) = \sum_{i=X}^{M} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}} = your\ Function$$

   You can calculate this with some actions on hygepdf or some actions on hygecdf. The answer to section 1 is NOT just hygepdf nor just hygecdf. You need to do something to hygecdf.

   Your function should give lower p-value the closer X is to K and p-value closer to 1 the closer X is to 0.

   Hint: hygepdf accepts vectors. Hint2: Probabilty of all events sums up to 1.

In sections 2-4 you use your result from section 1 with hygepdf/hygecdf. If you haven't figured it out, note that you haven't figured it out and calculate the p-value with what you've got - mistakes won't carry forward.

2. Please calculate the functional enrichment of genes that are differentially expressed between breast and prostate. Calculate once without parents
3. Calculate once with parents. Note the enrichments that appear only when using parents.
4. Remove GO terms that have less than 5 genes or more than 500 genes and repeat section 2.
5. **Optional** - Take one of your results, and convert it to GO terms, rather than IDs using the provided file `GOIDsToTerms.txt` . What seem to be the main biological differences between breast and prostate cell lines (10 bonus points).