Networks_hw

February 25, 2021

1 Networks HW

The goal of this homework is to gain hands on experience working with real biological networks, and to start thinking about the possible ways that choice of network might affect your inference about the biology that you are studying.

In this assignment we will focus on comparing Mendelian disease genes, genes that are frequently somatically mutated in cancer, and genes that are neither. We would like to understand whether cancer genes are similar to Mendelian disease genes as has been previously suggested in the literature: > Torkamani, Ali, and Nicholas J. Schork. "Prediction of cancer driver mutations in protein kinases." Cancer research 68.6 (2008): 1675-1682.

2 Instructions

- Copy all necessary files for this homework to your home directory with the following commands:

```
cd ~
cp -r /datasets/cm262-wi21-A00-public/hw2 .
```

- Please write your code directly into this notebook. Add your written answers also directly into this notebook, in comment form.
- Save your Jupyter notebook as **PDF** (`File -> Download as -> PDF via LaTeX`) and submit to GradeScope.

Hint: All commands needed to complete this homework can be found in the exercise notebooks completed in class!

[ ]:
```
# Load libraries for network analysis
library(igraph)
```

[ ]:
```
# Read in two networks
# First a binary protein interaction network constructed from an unbiased␣
 ↪yeast2hybrid experimental screen
Y2H <- read.table(file="~/hw2/data/Networks/HI-II-14.tsv",header=T,sep="\t")
head(Y2H)
# Second an literature curated network of high confidence protein interactions
lit <- read.table(file="~/hw2/data/Networks/Lit-BM-13.tsv",header=T,sep="\t")
head(lit)
```

```
# These networks are hosted here: http://interactome.dfci.harvard.edu/H_sapiens/
↪index.php
```

[ ]:
```
# Load networks
edgelist <- cbind(as.character(Y2H$Symbol.A),as.character(Y2H$Symbol.B))
g <- graph.data.frame(edgelist, directed=F)
edgelist2 <- cbind(as.character(lit$symbol_a),as.character(lit$symbol_b))
g2 <- graph.data.frame(edgelist2, directed=F)
```

3 Question 1 (2 points)

In this homework, we will investigate similarities and differences in the networks generated by systematic screen versus literature curation.

1a) How many nodes?

[ ]:

1b) How many edges?

[ ]:

1c) Get a list of unique node names - hint: use `names()` with the solution to 1a.

[ ]:

1d) Compare the node degree distributions of the 2 graphs. Do they both follow a power law distribution?

[ ]:

1e) Finally, compare the diameters of the graph. Based on the results of 1a-1e, briefly explain similarities and differences between properties of the two graphs.

[ ]:

4 Question 2 (4 points)

Evaluate coverage of different gene sets in the graph.

> **Hint**: use this syntax to get the nodes in the graph that are also in a list of interesting genes.
>
> `nodesinlist <- nodenames[which(nodenames %in% genelist)]`
>
> The - symbol gives the names **not** in the list.
>
> `nodesinlist <- nodenames[-which(nodenames %in% genelist)]`

[ ]:
```
# Load in disease gene lists
mend <- scan("~/hw2/data/OMIM/Mendelian_HGNC.txt",what=as.character())
cancer <- scan("~/hw2/data/Cancer/cancer_genes.2_sources.txt",what=as.
 ↪character())
length(mend)
length(cancer)
```

2a) Determine how many mendelian disease genes are in the Y2H graph.

[ ]:

2b) Determine how many cancer genes are in the Y2H graph.

[ ]:

2c) Make a list of the nodes in the Y2H graph that are neither cancer nor mendelian disease genes.

[ ]:

Repeat the same statistics for the **literature-based** graph.

2d) Determine how many mendelian disease genes are in the literature-based graph.

[ ]:

2e) Determine how many cancer genes are in the literature-based graph.

[ ]:

2f) Make a list of the nodes in the literature-based graph that are neither cancer nor mendelian disease genes.

[ ]:

5 Question 3 (4 points)

Compare graph measures between disease genes, cancer genes, and non-disease genes in the Y2H network

Hint: You want a number for each gene in the group. Boxplots are a good way to compare distributions. If you get warnings for a method you run here that's ok.

3a) Plot degree distributions for each class of gene

[ ]:

3b) Plot clustering coeffcient distribution for each class of gene

[ ]:

3c) Plot closeness centrality for each class of gene

[ ]:

3d) Plot betweenness centrality for each class of gene

[ ]:

Now repeat for the literature curated network.

3e) Plot degree distributions for each class of gene

[ ]:

3f) Plot clustering coeffcient distribution for each class of gene

[ ]:

3g) Plot closeness centrality for each class of gene

[ ]:

3h) Plot betweenness centrality for each class of gene

[ ]:

3i) Do your conclusions about the properties of these different classes of genes change when you use different networks?

[ ]:

6 Question 4 (2 points)

Next, compare enrichment for 4 node motifs in the Y2H network versus the literature based network.

Hint: There are 6 unique motifs where edges connect all 4 nodes in an undirected graph; there are 11 total undirected 4 node motifs when the subgraph doesn't have to be connected.

4a) Visualize the possible 4 node motifs.

[ ]:

4b) Count the number of motifs in each graph.

[ ]:

4c) Do the graphs differ in terms of the number motifs? Which motifs are more common in the Y2H network? Which in the literature derived network?

6.1 Bonus

Are these motifs over-represented relative to similar random networks?

Hint: you can perform degree preserving permutation using `rewire(g, with = keeping_degseq())` - see `igraph` documentation - The `niter` parameter is the number of edges that will be randomly reassigned - This might be computationally intensive!

[ ]: